This is to certify that the

dissertation entitled

A 'REARRANGEMENT PROCEDURE' FOR ADMINISTERING ADAPTIVE TESTS WHEN REVIEW OPTIONS ARE PERMITTED

presented by

Elena C. Papanastasiou

has been accepted towards fulfillment
of the requirements for

Ph. D. degree in Measurement and Quantitative Methods

Mark D. Reckase

Mark D. Reckase
Major professor

Date 6/25/01

A 'REARRANGEMENT PROCEDURE' FOR ADMINISTERING ADAPTIVE TESTS WHEN REVIEW OPTIONS ARE PERMITTED

By

Elena C. Papanastasiou

AN ABSTRACT OF A DISSERTATION

Submitted to
Michigan State University
in partial fulfillment of the requirements
for the degree of

DOCTOR OF PHILOSOPHY

Department of Counseling, Educational Psychology and Special Education

2001

Major Professor: Mark D. Reckase

ABSTRACT

A 'REARRANGEMENT PROCEDURE' FOR ADMINISTERING ADAPTIVE TESTS WHEN REVIEW OPTIONS ARE PERMITTED

By

Elena C. Papanastasiou

Computerized adaptive testing (CAT) has gained increased popularity during the last decades. Consequently, admissions tests such as the GRE, the TOEFL, as well as other certification and licensure exams have been transformed from their paper-and-pencil versions to computerized adaptive versions. However, a major difference between the two test formats, from an examinee's point of view, is that examinees are usually not allowed to revise their answers on CATs. Some researchers feel that the validity of a CAT can increase with item review, since it allows them to rethink their answers and make corrections to items that have been misread or miskeyed. Other researchers believe that item review can decrease the efficiency and validity of a CAT since item review allows examinees to cheat on the test.

The purpose of this study was to test the efficiency of a '*rearrangement procedure*' that rearranges and skips certain items in order to better estimate the examinees' abilities. This was examined through a simulation study. This rearrangement procedure permits examinees to change their answers on a CAT without allowing them to artificially inflate their test scores. If this procedure were adopted, it could help reduce the stress of examinees who feel that they have more control over the testing situation when they can revise their answers. This procedure could also help

improve the reliability and validity of the CAT since errors due to misread and miskeyed answers would be corrected.

The results of this simulation study have shown that when the Maximum Likelihood estimation was used, the rearrangement procedure was effective in reducing the bias of the ability estimates. With the Bayesian estimation, the rearrangement procedure increased the reliability, and slightly decreased the standard error of the estimates. However, the Bayesian method was not effective in reducing the bias of the estimates after the rearrangement procedure.

There were not many differences in the accuracy of the estimates after the rearrangement procedure, when three or five items were reviewed. There were also very small differences in the effects that the item pool size had on the accuracy of the ability estimates after the rearrangement procedure.

For my family

ACKNOWLEDGMENTS

Now that I have finished my dissertation, it is great to be able look back and think about at how the various aspects of my education at Michigan State University have contributed to this dissertation. I suppose it all started when my father had suggested that I should apply to Michigan State University for my doctoral degree, even though I had been looking at other schools back then. And his suggestion was a brilliant one. I cannot imagine getting a better education at any other school.

Ever since my admission to the Measurement and Quantitative Methods program, I have obtained invaluable support from all of the professors in the department. Betsy Becker, Richard Houang, Ken Frank, William Mehrens, Mark Reckase, William Schmidt, Teresa Tatto, and Edward Wolfe, have all been great teachers as well as excellent role models during my training as an educator and a researcher. I especially owe great thanks to Mark Reckase, Richard Houang, William Mehrens, Maria Teresa Tatto, and Edward Wolfe who were all members of my dissertation committee.

During my first year in the program, my current advisor, Mark Reckase came to interview for a position at Michigan State University. At that point, it would was hard for me to imagine all the ways in which he would influence me during the rest of my studies. As an advisor, he has always provided excellent guidance throughout all of the stages of my doctoral work. Through his sincere interest in my studies, he always provided me with invaluable encouragement and support. As a teacher, his excitement about Item Response Theory, as well as his brilliant teaching skills had also influenced me to pursue my dissertation in the area of Item Response Theory. In addition, as the director of my dissertation committee, he has always provided me with outstanding suggestions and guidance. I will always be indebted to him.

The assistantships that I have had during my studies here have also contributed to a large part of my education. During the first two years of my studies, I was fortunate to have the opportunity to work with William Schmidt and Richard Houang on the Third International Mathematics and Science Study. Richard Houang was the one who initially patiently introduced me to programming in SAS, which I would later use for developing the program code for my dissertation. Richard Houang should also be especially thanked for the numerous hours that he spent discussing my dissertation with me. He has managed to push my thinking further, and has help me refine by dissertation ideas. I would also like to thank Edward Wolfe for introducing me to the development and planning of simulation studies. Without his support, it would have been very difficult for me to conceptualize the model for the simulation that I used in my dissertation.

William Mehrens, with his thoughtful comments and suggestions has been a very inspiring role model throughout my studies, as well as throughout this dissertation process. I would also like to thank Teresa Tatto for encouraging me throughout this process, and for providing me with invaluable input on the practical aspects and implications of my research study.

I would also like to thank the various sources that have provided me with financial support throughout my doctoral studies. These are the Robert Ebel scholarship, the Leventis Foundation, and the Society for Multivariate Experimental Psychology (SMEP). Frederic Robin should also be thanked for providing me with the Computer-Based Testing Simulation and Analyses Computer Program, which simulated a large portion of the data that I used for my dissertation study.

I am also very grateful to my parents Constantinos and Georgia Papanastasiou for their teachings, and their support. They are the ones who have instilled in me the value of education, and who have encouraged me to excel in my studies, throughout my student life. Σας αγαπώ πολύ. Finally, I would also like to thank all of my friends that

have provided me with encouragement and stress relief during the whole doctoral process. Thank you all!

TABLE OF CONTENTS

LIST OF TABLES. x

LIST OF FIGURES. xii

CHAPTER 1. INTRODUCTION. 1
Significance of the study 3

CHAPTER 2. LITERATURE REVIEW. 5
Item review 6
Problems with item review in CAT 8
Cheating strategies 9
Effects of item review 11

CHAPTER 3. METHODS. 14
Simulation specifications 15
Test specifications 16
Item pool development and characteristics 17
Ability estimation methods 19
Examinee characteristics 21
Simulation of the examinee's test taking behaviors 22
Item revision algorithm- The rearrangement procedure 25
Item skipping in the rearrangement procedure 26
Types of answer changing and the rearrangement procedure 26
Type 1 change. Incorrect to incorrect changes 26
Type 2 change. Incorrect to correct changes 27
Type 3 change. Correct to incorrect changes 28
Making two or more answer changes. Rearranging items in the rearrangement procedure 30
Convergence plots 33
Exceptions to the rule 34
Stopping rules 35
Dependent variables 36
Independent variables 37

CHAPTER 4. RESULTS. 38
Condition 1. 250 items with 3 reviews maximum 39
Results based on bias (Condition 1) 42
Results based on the standard error of the θ estimate (Condition 1) 49
Reliability of test scores (Condition 1) 54
Examinee anxiety effects (Condition 1) 55
Condition 2. 250 items with 5 reviews maximum 58
Results based on bias (Condition 2) 60
Results based on the standard error of the θ estimate (Condition 2) 66
Reliability of test scores (Condition 2) 71
Examinee anxiety effects (Condition 2) 72
Comparison of three and five answer changes with a 250 size item pool 74
Condition 3. 500 items with 3 reviews maximum 78

Results based on bias (Condition 3) 80
Results based on the standard error of the θ estimate (Condition 3) 86
Reliability of test scores (Condition 3) 92
Examinee anxiety effects (Condition 3) 92
Condition 4. 500 items with 5 reviews maximum 95
Results based on bias (Condition 4) 97
Results based on the standard error of the θ estimate (Condition 4) 103
Reliability of test scores (Condition 4) 107
Examinee anxiety effects (Condition 4) 109
Comparison of three and five answer changes with a 500 size item pool 111
Item pool size differences 114
Overall final comparison 116

CHAPTER 5.CONCLUSIONS. 118
How does the rearrangement procedure affect the bias of the ability estimates? ... 119
What are the effects of the rearrangement procedure on the reliability of the examinee's final ability estimate? 121
How does the rearrangement procedure affect the ability estimates of the examinees that have computerized-test-anxiety? 121
How does the choice of the ability estimation procedure affect the results from the rearrangement procedure? 122
How does the maximum number of item changes affect the examinee's final ability estimates? 123
Implications for practice 123
Limitations 125

APPENDIX A. SAS PROGRAM CODE. 129

BIBLIOGRAPHY. 165

LIST OF TABLES

Table 1. Target distributional characteristics of the item parameters. 19

Table 2. Examinee frequency distribution. 22

Table 3. Replication conditions of the simulation 3

Table 4. Percentage of actual answer changing patterns in condition 1 41

Table 5. Overall bias of the 250 pool with 3 reviews estimates (Condition 1) 43

Table 6. Conditional Maximum Likelihood bias when 3 reviews are permitted with a 250 sized item pool (Condition 1) 44

Table 7. Conditional Bayesian bias when 3 reviews are permitted with a 250 sized item pool (Condition 1) 46

Table 8. Overall standard deviation of the θ estimates obtained from the pool of 250 items, when 3 reviews were permitted (Condition 1). 49

Table 9. Conditional Maximum Likelihood standard error when 3 reviews are permitted with a 250 sized item pool (Condition 1). 50

Table 10. Conditional Bayesian standard error when 3 reviews are permitted with a 250 sized item pool (Condition 1). 51

Table 11. Reliability of the ability estimates with a pool of 250 items, and 3 permitted reviews (Condition 1). 55

Table 12. Percentage of actual answer changing patterns in condition 2 59

Table 13. Overall bias of the 250 pool with 5 reviews estimates (Condition 2) 60

Table 14. Conditional Maximum Likelihood bias when 5 reviews are permitted with a 250 sized item pool (Condition 2) 62

Table 15. Conditional Bayesian bias when 5 reviews are permitted with a 250 sized item pool (Condition 2) 63

Table 16. Overall standard deviation of the θ estimates obtained from the pool of 250 items, when 5 reviews were permitted (Condition 2). 66

Table 17. Conditional Maximum Likelihood standard error when 5 reviews are permitted with a 250 sized item pool (Condition 2) 67

Table 18. Conditional Bayesian standard error when 5 reviews are permitted with a 250 sized item pool (Condition 2). 68

Table
Table
Table
Table
Table
Table
Table
Table
Table
Table
Table
Table
Table
Table
Table
Table
Table
Table
Table
Table

Table 19. Reliability of the ability estimates with a pool of 250 items, and 5 permitted reviews (Condition 2). 72

Table 20. Percentage of actual answer changing patterns in condition 3..... 79

Table 21. Overall bias of the 500 pool with 3 reviews estimates (Condition 3)..... 81

Table 22. Conditional Maximum Likelihood bias when 3 reviews are permitted with a 500 sized item pool (Condition 3) 82

Table 23. Conditional Bayesian bias when 3 reviews are permitted with a 500 sized item pool (Condition 3) 83

Table 24. Overall standard deviation of the θ estimates obtained from the pool of 500 items, when 3 reviews were permitted (Condition 3). 87

Table 25. Conditional Maximum Likelihood standard error when 3 reviews are permitted with a 500 sized item pool (Condition 3)..... 88

Table 26. Conditional Bayesian standard error when 3 reviews are permitted with a 500 sized item pool (Condition 3). 89

Table 27. Reliability of the ability estimates with a pool of 500 items, and 3 permitted reviews (Condition 3). 90

Table 28. Percentage of actual answer changing patterns in condition 4..... 96

Table 29. Overall bias of the 500 pool with 5 reviews estimates (Condition 4)..... 97

Table 30. Conditional Maximum Likelihood bias when 5 reviews are permitted with a 500 sized item pool (Condition 4) 98

Table 31. Conditional Bayesian bias when 5 reviews are permitted with a 500 sized item pool (Condition 4) 100

Table 32. Overall standard deviation of the θ estimates obtained from the pool of 500 items, when 5 reviews were permitted (Condition 4). 103

Table 33. Conditional Maximum Likelihood standard error when 5 reviews are permitted with a 500 sized item pool (Condition 4)..... 104

Table 34. Conditional Bayesian standard error when 5 reviews are permitted with a 500 sized item pool (Condition 4). 105

Table 35. Reliability of the ability estimates with a pool of 500 items, and 5 permitted reviews (Condition 4). 108

Table 36. Overall results comparison 117

Fig
Fig
Fig.
Fig.
Fig.
Fig.
Fig.
Fig.
Figu
Figu
Figu
Figu
Figu
Figu
Figure
Figure
Figure

LIST OF FIGURES

Figure 1. Example of an incorrect-to-correct answer change on a CAT (type 2 change) 29

Figure 2. Example of a correct-to-incorrect answer change on a CAT (type 3 change) 30

Figure 3. Rearrangement procedure without a rearrangement of the item order 32

Figure 4. Rearrangement procedure with a rearrangement of the item order 33

Figure 5. Convergence plot with correct-to-incorrect change 34

Figure 6. Condition 1 ML and Bayesian bias after the rearrangement procedure (250 pool and 3 reviews) 48

Figure 7. Percentage of ARP bias improvement with 3 changes and a pool of 250 items (condition1) 48

Figure 8. Condition 1 ML and Bayesian standard error after the rearrangement procedure (250 pool and 3 reviews) 53

Figure 9. Percentage of ARP standard error improvement with 3 changes and a pool of 250 items (condition 1). 54

Figure 10. Anxiety effects and bias of the ability estimates (condition 1). 56

Figure 11. Standard error estimates of examinees with anxiety (condition 1). 57

Figure 12. Condition 2 ML and Bayesian bias after the rearrangement procedure (250 pool and 5 reviews). 64

Figure 13. Percentage of ARP bias improvement with 5 changes and a pool of 250 items (condition 2) 65

Figure 14. Condition 2 ML and Bayesian standard error after the rearrangement procedure (250 pool and 5 reviews) 69

Figure 15. Percentage of ARP standard error improvement with 5 changes and a pool of 250 items (condition 2) 71

Figure 16. Anxiety effects and bias of the ability estimates (condition 2) 73

Figure 17. Standard error estimates of examinees with anxiety (condition 2) 74

Figure 18. The effects of the number of items reviewed on the ML bias improvement from a pool of 250 items 75

Figure
Figure
Figure
Figure
Figure
Figure
Figure
Figure
Figure
Figure
Figure
Figure
Figure
Figure
Figure
Figure
Figure
Figure
Figure
Figure

Figure 19. The effects of the number of items reviewed on the Bayesian bias improvement from a pool of 250 items 75

Figure 20. The effects of the number of items reviewed on the ML standard error improvement from a pool of 250 items 76

Figure 21. The effects of the number of items reviewed on the Bayesian standard error improvement from a pool of 250 items 77

Figure 22. Condition 3 ML and Bayesian bias after the rearrangement procedure (500 pool and 3 reviews) 85

Figure 23. Percentage of ARP bias improvement with 3 changes and a pool of 500 items (condition 3) 86

Figure 24. Condition 3 ML and Bayesian standard error after the rearrangement procedure (500 pool and 3 reviews) 90

Figure 25. Percentage of ARP standard error improvement with 3 changes and a pool of 500 items (condition 3). 91

Figure 26. Anxiety effects and bias of the ability estimates (condition 3) 93

Figure 27. Standard error estimates of examinees with anxiety (condition 3) 94

Figure 28. Condition 4 ML and Bayesian bias after the rearrangement procedure (500 pool and 5 reviews). 101

Figure 29. Percentage of ARP bias improvement with 5 changes and a pool of 500 items (condition 4) 102

Figure 30. Condition 4 ML and Bayesian standard error after the rearrangement procedure (500 pool and 5 reviews) 106

Figure 31. Percentage of ARP standard error improvement with 5 changes and a pool of 500 items (condition 4). 108

Figure 32. Anxiety effects and bias of the ability estimates (condition 4) 109

Figure 33. Standard error estimates of examinees with anxiety (condition 4). 110

Figure 34. The effects of the number of items reviewed on the ML bias improvement from a pool of 500 items 111

Figure 35. The effects of the number of items reviewed on the Bayesian bias improvement from a pool of 500 items. 112

Figure 36. The effects of the number of items reviewed on the ML standard error improvement from a pool of 500 items 113

Figure

Figure

Figure

Figure 37. The effects of the number of items reviewed on the Bayesian standard error improvement from a pool of 500 items. 113

Figure 38. Item pool size effects on the estimation bias when 5 reviews are permitted 114

Figure 39. Item pool size effects on the standard error when 5 reviews are permitted. 115

last t

like t

penc

used

psyc

regu

exar

exar

ther

Gui

conc

prob

estr

exar

make

By al

more

such

and va

CHAPTER 1

INTRODUCTION

Computerized adaptive testing (CAT) has gained increased popularity during the last two decades (Reckase, 2000). Consequently, many tests such as admissions tests like the GRE, the SAT, and the TOEFL, have been transformed from their paper-and-pencil versions to computerized adaptive versions. Adaptive tests are now also being used for certification and licensure purposes (Stone & Lunz, 1994).

The popularity of CAT has prompted researchers, measurement specialists and psychometricians to reconceptualize many of the processes that were established for regular paper-and-pencil tests (Pommerich & Burden, 2000; Reckase, 2000). For example, a major difference between paper-and-pencil tests and adaptive tests, from an examinee's point of view, is that in many cases, examinees are not allowed to revise their answers on CATs (Vispoel, Rocklin & Wang, 1994; Wise, Roos, Plake & Nebelsick-Gullett, 1994; Wise, 1997a; Wise, 1997b). Studies have shown that this is a major concern for students who feel anxious while taking tests. This anxiety is especially problematic since anxiety can be an additional source of error in the examinee's ability estimates. So some researchers feel that the validity of a CAT can increase when examinees can revise their answers, since it allows them to rethink their answers and make corrections to items that might have been misread or miskeyed (Vispoel, 1998a). By allowing revisions, the final ability estimate could represent an examinee's ability more accurately because it will be closer to his/her actual ability when small mistakes such as miscodings are corrected (Wise, 1996; Vispoel, Henderickson & Bleiler, 2000).

Other researchers, however, believe that item review can decrease the efficiency and validity of a CAT since item review allows examinees to cheat on the test. An

example of a cheating strategy is the Wainer strategy (Wainer, 1993), in which examinees might purposely answer all the items incorrectly when they are first administered so that they can have the easiest items administered to them. The second step of the Wainer strategy involves going back to the test items, and answering all of the items on the test correctly. Answering all the items correctly should not be very difficult for these examinees since the test would consist of very easy items that have low difficulty levels. This would result in an artificial inflation of the examinee's ability estimates.

Other studies on item review have also shown that the efficiency of a test decreases when item review is permitted (Stocking, 1997; Vispoel, Rocklin, Wang & Bleiler, 1999). For this reason, item review is not permitted in most adaptive tests at this time (Vispoel, Henderickson & Bleiler, 2000).

The purpose of this study is to test the effectiveness of a rearrangement procedure that permits examinees to review previously presented items without allowing them to artificially inflate their test scores by using test-wiseness strategies. More specifically, the research questions that will be answered in this study are the following:

1. What are the effects of the rearrangement procedure on the reliability of the estimates?
2. How much statistical bias and error does the rearrangement procedure create?
3. How does the rearrangement procedure affect the ability estimates for the examinees who have anxiety because of the computerized format of adaptive tests?
4. How does the choice of the ability estimation procedure affect the estimates after the rearrangement procedure?

Significance Of The Study

The issue of item review is of great importance to examinees who are administered tests, as well as to testing organizations that administer tests. From the perspective of the examinee, tests are stressful situations overall, and even more so when they are high stakes tests. Therefore, examinees would like to have as much control of the testing situation as possible when they are taking tests, so that they can perform to the maximum extent of their capabilities. Such control is achieved by allowing the examinees to use the test taking strategies that they have been accustomed to. So when examinees are permitted to review answers on a test, the majority of them choose to do so (Bowles & Pommerich, 2001).

When examinees are administered paper-and-pencil tests, each individual uses various strategies while completing the test (Vispoel, Hendrickson & Bleiler, 2000). For example, some examinees choose to go through the test once and answer all the questions immediately no matter how confident they are of their responses. After they answer all the questions, they go over the whole test again, they rethink all of their answers, and they might make any changes that are necessary to their original answers. Other examinees choose to omit questions that they are unsure of and go back to those items after they reached the end of the test (Stocking, 1997).

When taking computer adaptive tests, however, in most cases examinees are not allowed to go back and revise their answers. So the examinees who have been using the strategies mentioned above cannot use those anymore on computer adaptive tests. This may cause stress and anxiety to many examinees, especially if they are taking high stakes tests (Wise, Roos, Plake & Nebelsick-Gullett, 1994; Wise, 1997a; Wise, 1997b). This might cause even bigger problems and stress to international students who have to take high stakes admissions tests such as the TOEFL and the SAT or GRE in a foreign

language to get admitted to universities in the USA. Therefore, a large number of examinees may actually be at a disadvantage when taking computerized adaptive tests, because the no-revision policy might prevent them from performing to the maximum extent of their abilities (Vispoel, 1998a).

The stress that is caused by computer adaptive tests might even cause examinees to make mistakes on questions to which they know the answers. For example, due to stress, examinees might choose an incorrect option accidentally even though they knew the correct answer to a question (Lunz, Bergstrom & Wright, 1992). It is also possible that the stress can cause examinees to make foolish arithmetic errors although they have the ability and skills to answer them correctly. So if examinees lose points on tests due to such reasons, and if they are not allowed to go back and revise their answers, their test scores will not be valid indicators of their true abilities (Vispoel, Henderickson & Bleiler, 2000).

However, item review can also be costly to testing organizations that believe that the efficiency of their tests, as well as the validity of the test scores would be compromised if item review were permitted (Gershon & Bergstrom, 1995). For example, examinees might try to "trick" the computer to artificially inflate their test scores with item review (Wainer, 1993). In addition, when item review is permitted, examinees might have more time to memorize test items, which would jeopardize the security of these tests (Patsula & McLeod, 2000). Therefore, this study will attempt to provide a compromised solution to this problem. This would involve a solution that would allow examinees to revise their answers, without jeopardizing the quality and efficiency of the test.

by ye

2000)

testing

One c

Greer

estima

length

tailor

(More

tests a

pencil

times.

provide

control

adminis

paper-a

conclus

(1997b)

CHAPTER 2

LITERATURE REVIEW

The popularity of computerized adaptive testing (CAT) has been increasing year by year, especially after the increase in the power of desktop computers (Reckase, 2000). The reason for this popularity is because the numerous benefits of adaptive testing have become evident to testing companies, test developers, and administrators. One of the main benefits of adaptive tests is that they are very efficient (Wainer, Dorans, Green, Mislevy, Steinberg & Thissen, 2000). Due to this efficiency, adaptive tests can estimate an examinee's ability more accurately with fewer questions than with fixed-length paper-and-pencil tests (McBride, Wetzel & Hetter,1997). This is achieved by tailoring each CAT to each examinee's individual estimated ability.

Another advantage of CAT is that the tests can be offered more frequently (Moreno, 1997). With adaptive tests, examinees can schedule the administration of their tests at a time and date of their choice. This is in contrast to the traditional paper-and-pencil tests where all the examinees are administered the tests at specific dates and times.

Some other advantages of the CAT, as listed by Wainer (1993), are that a) it provides the final scores to examinees immediately after they finish taking the test, b) it controls for cheating by preventing stolen booklets from being circulated, c) it permits the administration of types of questions that could not have been asked in the traditional paper-and-pencil format, and d) it maintains more control of the item pool.

However, there are some practical issues in adaptive testing that have not been conclusively resolved. Mills and Stocking (1995), have listed 18 such issues. Wise (1997b) divided these issues further into four clusters of issues that were labeled as: a)

item p

issues

the ma

integri

exam

reduce

On a m

compu

adapt

main re

reason

categor

Legitim

answers

conside

accurate

item pool development and maintenance issues, b) administering and scoring issues, c) issues in the protection and integrity of the item pool, and d) examinee issues. One of the major issues that has to do with the examinees, as well as with the protection and integrity of the item pool, is that of *'item review'*. Item review is the process of allowing examinees to go back and revise or change their answers on an adaptive test.

On a theoretical level, allowing examinees to revise their answers on such tests reduces the efficiency of these tests (Lunz, Bergstrom & Wright, 1992; Wainer, 1993). On a more practical level, wise test-takers could use the review options to "trick" the computer and obtain scores that do not accurately reflect their ability. Therefore, adaptive tests typically do not allow examinees to revise their responses.

Item Review

Why do examinees want to review items? Harvill and Davis (1997) found ten main reasons why students might review and change their answers on exams. The reasons for making these changes were

> 1) Reread and better understood the test item, 2) Rethought and conceptualized a better answer, 3) Gained information from other test items, 4) Gained information from the instructor, 5) Remembered more information 6) Used a clue or cue within the test item, 7) Made a clerical (recording) correction, 8) Corrected an arithmetic /mathematic error 9) 'Gut feeling' that the new response was a better answer, 10) one wild guess replaced the other (p.97).

These ten reasons for choosing to review items can be divided in two major categories: into legitimate and illegitimate reasons for changing answers (Wise, 1996). Legitimate reasons are the ones in which examinees change incorrect to correct answers due to knowledge that was possessed at the beginning of the test. This can be considered good practice since the final score would reflect an examinee's ability more accurately. In turn, the validity of the test increases.

corre

othe

not b

infor

decre

stude

stude

quest

the q

this s

were

permi

entry

that ex

answe

and-pe

student

scores

> Allowing answer changes following review also could increase test score validity if the changes reflect corrections of typing errors, misreading of items, temporary lapses in memory, or reconceptualizations of answers to previously administered items. Under these conditions, item review would yield more valid scores because the scores would represent the examinee's skill level at the end of the test more accurately, and the scores would not be contaminated with clerical or other inadvertent errors (Vispoel, 1998b, p.338).

Illegitimate reasons for changing answers include the cases in which examinees correct an incorrect response due to test wiseness (e.g. by gaining information from other test items, or by the instructor), as described in points 3 and 4 above. This would not be considered as good practice since the final score would provide misleading information about an examinee's true ability. In addition, the validity of the test would decrease in such situations.

Schwartz, McMorris and DeMers (1991) have found that the majority of the students would change items because of legitimate reasons. Forty five percent of the students would change their answers because they reread and better understood the question; 31% would change their answers because they rethought and conceptualized the question better, and 20% because they remembered more information. In addition, this study also found that the students that gained the most out of their answer changing were the students in the middle or highest third of their class. It is an issue of fairness to permit the examinees to demonstrate their true knowledge by checking for calculation or entry errors or for uncertain responses (Lunz, Bergstrom & Wright, 1992).

In addition, prior research that was based on paper-and-pencil tests has shown that examinees tend to increase their test scores when they are allowed to revise their answers. In a study conducted by Geiger (1991), it was found that on regular paper-and-pencil tests where students have the opportunity to review items, 97% of the students had changed at least one item. In addition, 70% of the students increased their scores by changing their answers on the test (Geiger, 1991). Wagner, Cook, and

Friedm

show t

of thos

points

conduc

were n

to inco

67% o

and Lu

were fr

options

Proble

consec

anxiety

environ

situatio

stress (

consist

inability

exposed

decreas

depende

Friedman (1998) found similar results with a sample of fifth grade students. Their results show that 85% of the students changed their answers during the test, and that only 23% of those students lost points by the changes. Fifty-seven percent of the students gained points by their changes, while 20% had no change in their final scores. A meta-analysis conducted with 75 studies on answer changing, found that 57% of the answer changes were made from incorrect to correct options, and 21% of the changes were from correct to incorrect options (Waddell & Blankenship, 1994).

Vispoel (1998) found similar results for a computer adaptive test. He found that 67% of the examinees had made changes to their answers on the adaptive test. Stone and Lunz (1994) also found that 47% of the total answers changed on an adaptive test were from incorrect to correct options, and 27% were changed from correct to incorrect options.

Problems With Item Review In CAT

The inability of examinees to revise their answers can have negative consequences on the examinees. First of all, the no-revision policy causes increased anxiety to the examinees that perceive that they have little control over the testing environment (Wise, 1997). Studies have shown that examinees can cope with stressful situations and test anxiety much better when they have some control over their source of stress (Wise, Roos, Plake & Nebelsick-Gullett, 1994). In addition, examinees have consistently reported that one of the main disadvantages of adaptive testing is their inability to go back and revise their answers to questions that had been previously exposed to them (Vispoel, Rocklin & Wang, 1994). Consequently, this anxiety can decrease the examinee's performance on adaptive tests (Wise, 1997a).

Item review and revisions can also have negative consequences. Due to item dependence, examinees might obtain clues to a correct answer based on the stems

and or

create

overall

depend

respon

have t

adapti

Theref

of revie

which

adapti

the exa

item, is

a temp

X, by c

change

that wo

less inf

1993).

to revie

the test

and/or response options that were provided on other items (Wise, 1996). This can create dependencies in the errors, which would further decrease the efficiency and the overall psychometric properties of the test and the test scores (Wise, 1996). These dependencies are problematic because they violate one of the main assumptions of item response theory, that the items are all independent of each other for examinees that have the same ability level θ (Lord, 1980). However, this problem exists for non-adaptive tests as well, and it cannot be fully eliminated either with or without item review. Therefore, this reason alone is not a sufficient reason to refuse the examinees the option of reviewing and possibly revising their items.

Another major problem with item review is that it does not follow the logic on which adaptive testing is based, which can greatly compromise the efficiency of an adaptive test (Wise, 1996). More specifically, when an examinee takes an adaptive test, the examinee's ability estimate is calculated after the response to each item. The next item, is then selected to provide the maximum information (or smallest standard error) at a temporary ability estimate for the examinee. However, if an examinee revises an item X, by changing it after reaching the end of the test, the examinee's ability estimate might change. Consequently, the item Y that was administered after item X will not be the item that would provide the maximum information at that ability estimate. So by providing less information, the standard error for the final ability estimate will be larger (Wainer, 1993). This is the reason why the test will have less efficiency if examinees are allowed to review items. A possible solution for this problem would be to increase the length of the test to reduce the standard error of the estimate (Wainer, 1992).

Cheating strategies

Wainer (1993) described a situation called the Wainer strategy, in which examinees would intentionally answer all the items wrong in order to obtain the easiest

items
could
Wainer
and pr
simulat
profice
they ar
exam
(Wise,
Wainer
adopte

(1984),
item diff
their res
For exa
the follo
examine
rationale
that their
the exam
answered

such diff
Enders, a
more diff

items on their test. If this were the case, and item review were permitted, the examinees could then go back and answer those easy items correctly. "The logic underlying the Wainer strategy is basically that if the test is easy enough, then invariance will not hold, and proficiency estimates will consequently be higher " (Wise, 1996, p.10). However, a simulation study conducted by Gershon and Bergstrom (1995) found that only highly proficient examinees would be able to profit from this strategy, with the assumption that they answer all the items correctly when revising their answers. However, these examinees are the ones that are least likely to need to perform such cheating strategies (Wise, 1996). In addition, even if examinees were able to successfully use them, the Wainer strategy would be very easy to spot. In turn, the tests of the examinees that adopted such cheating strategies could be invalidated (Wise, 1996).

Another strategy first noted by Green, Bock, Humphreys, Linn, and Reckase (1984), that was further elaborated by Kingsbury (1996), has to do with detecting the item difficulty of the items. In the Kingsbury strategy, examinees could try to determine if their response to a previous question was correct based on the subsequent question. For example, after responding to a specific question 10, an examinee might notice that the following question 11 is a more difficult than question 10. This would lead the examinee to assume that their answer to question 10 was correct. With the same rationale, if the following question 12 appeared to be easier, the examinee could assume that their answer to the previous question 11 was wrong. So if item review is permitted, the examinees could just go back and revise their answers to those items that they answered incorrectly.

However, it is not exactly clear to what extent examinees would be able to detect such differences in item difficulties. In a study conducted by Wise, Freeman, Finney, Enders, and Severance (1999), it was found that examinees were not able to identify the more difficult item at a better than chance level when comparing pairs of items without

having

proble

results

most s

examir

Finney

actual

on an a

examir

taught

which w

examin

the sec

efficien

is that t

Effects

the effec

compare

response

the items

efficiency

by Lunz,

revision o

having to solve them. However, when the examinees had to actually solve the problems, which were administered only one at a time, the examinees had slightly better results in identifying the more difficult items. In addition, the examinees that were the most successful in this process were the high ability examinees. Overall, however, the examinees were not very proficient in discriminating item difficulties (Wise, Freeman, Finney, Enders, & Severance, 1999). Similar results were found by a study that had actually taught examinees how to use the Kingsbury strategy, and advised them to use it on an adaptive test (Vispoel, Clough, Bleiler, Henderickson & Ihrig, 2001). Those examinees were also not effective in inflating their test scores even after they were taught how to use the strategy.

Another disadvantage of item review is that it could increase the testing time, which would have two negative consequences. First, by increasing the testing time, examinees would have more time to memorize the test items, which would jeopardize the security of the item pool. In addition, increase testing time would decrease the efficiency of adaptive tests since one of the original arguments in favor of adaptive tests is that they decreased testing time (Wise, 1996).

Effects Of Item Review

What are the effects of item review on adaptive testing? In order to understand the effects of review on a computer adaptive certification exam, Stone and Lunz (1994) compared the examinee responses before and after reviewing and possibly altering their responses on the test. Their results show that the error of measurement after reviewing the items increased by approximately 0.0025. This means that the loss of precision and efficiency on the test, caused by the item review was minimal. Another study performed by Lunz, Bergstrom and Wright (1992) found that the loss of information due to the revision of items was less than the amount of information that would be added if one

addic

same

signifi

same

might

allowe

might

no exa

inflate

studies

into ac

their m

or three

actuall

the exa

possible

The rea

(includi

test wou

in the st

likelihoo

standard

the chan

specifica

additional item targeted to an examinee's ability were added to the test. Moreover, the same study found that the examinees who were able to review their answers performed significantly better than an equivalent group of examinees that were administered the same test, but were not allowed to revise their answers. This increase in test scores might also be due to the comfort that the examinees feel when they know that they are allowed to go back and revise their answers, and correct possible careless errors they might have made (Lunz, Bergstrom, & Wright, 1992). It should be noted, however that no examinee in the previous studies purposely used the Wainer strategy to artificially inflate their test scores (Stocking, 1997; Stone & Lunz, 1994). In addition, both of these studies were based on the one parameter, Rasch model. So these studies did not take into account the pseudo-guessing parameter (c) or the discrimination parameter (a) in their models. Therefore, it is not clear if these results would also be replicated with a two or three parameter logistic model.

Gershon and Bergstrom (1995) also examined whether cheating strategies could actually help examinees inflate their test scores. Their results show that even though the examinees might be able to get an easier test by using cheating strategies, it is very possible that their final ability estimate will be much lower than their true ability estimate. The reason is because the final ability estimate is based on all items and all responses (including the original incorrect responses). Therefore, trying to cheat on an adaptive test would clearly be an unwise procedure (Gershon & Bergstrom, 1995).

Intentional use of the Wainer strategy on a CAT can lead to an artificial increase in the standard error (SE). The SE can increase up to six times when the maximum likelihood (ML) ability estimation procedure is used, when compared to the mean standard error of a legitimate CAT (Vispoel, Rocklin, Wang, & Bleiler, 1999). However, the change in SE may differ when other estimation procedures are used. More specifically, Vispoel, Rocklin, Wang, and Bleiler (1999) found that the expected a

poste

Wain

stude

resu

revis

that w

large

is to t

revie

it is p

adapt

that w

would

posteriori (EAP) estimates were the ones that were least likely to increase due to the Wainer strategy.

Stocking (1997) also examined the effects of revising items on a CAT where students were purposely told to use the Wainer strategy while answering the test. Her results show that the conditional bias of a test, when up to two items (out of 28) were revisited and changed, was minimal. However, when there were seven or more items that were revisited, there was a positive bias in the test scores. This bias was especially large for examinees with approximately average or high scores.

Overall, the research presented above shows that permitting item review is to the benefit of the examinees. However, testing companies do not prefer item review, since it does not follow the logic on which adaptive tests are based on, and since it is prone to cheating strategies. Consequently, item review is not permitted in many adaptive tests. The purpose of this study is to examine the efficiency of a CAT algorithm that would permit examinees to review previously presented items, to determine if it would allow them to artificially inflate their scores by using test-wiseness strategies.

CHAPTER 3

METHODS

Examinees have usually not been permitted to revise their answers in many adaptive tests because of the possibility that the tests would lose their quality and effectiveness. For example, if an examinee went back to revise item A on an adaptive test, and changed a response from a correct to an incorrect response, then item B (that was selected because of the previous response to item A) would not be the most informative item at this new ability estimate. Consequently, the accuracy of the ability estimates and the efficiency of the test would both decrease. This loss of efficiency and accuracy could become even worse when test wiseness and cheating strategies are used to artificially inflate test scores (Lunz, Bergstrom & Wright, 1992; Stocking, 1997; Vispoel, Rocklin, Wang & Bleiler, 1999).

One of most talked about cheating strategies in the CAT literature, is the Wainer strategy (Wainer, 1993). Based on the Wainer strategy, examinees originally go over the test and purposely answer all the items incorrectly when they are first administered. Consequently, very easy items are administered to these examinees. Because most of the items on such a test are very easy, the examinees can easily go back and correct all of their answers. By doing so, they can artificially inflate their ability estimates. As a result, the ability estimate becomes inaccurate, and the standard error of the test could increase (Gershon & Bergstrom, 1995). In order to remedy such situations, testing organizations could consider two options. One option would be to administer additional items to the examinees until the standard error of the test decreases to the desired level. However, this would create a less efficient and very costly test. Another option is for testing organizations to restrict examinees from revising their answers. Most testing

organizations have chosen the second option, so at this time most adaptive tests do not allow item review (Vispoel, Henderickson & Bleiler, 2000; Wise, 1996).

The purpose of this study is to assess the effects of a specific 'rearrangement procedure' that rearranges and skips certain items in order to obtain a better estimate of the examinee's ability. It is hypothesized that the rearrangement procedure will improve the ability estimates of the examinee's scores. It is also expected that the rearrangement procedure will have three additional advantages if it were used in real life. First, by using this rearrangement strategy, the estimated ability levels of the examinees may become more valid since the examinees would have a chance to correct any errors or miscodings that they might have made. In addition, the rearrangement procedure will also help reduce the stress of examinees because they will have more control over the testing situation when they can revise their answers. Finally, the third advantage of the revision strategy that is essential for testing organizations, is that it will not permit the Wainer strategy from taking place.

To examine the rearrangement procedure, a simulation study was conducted to determine the effect that the rearrangement procedure would have on the accuracy of the examinee's ability estimates. All the simulation procedures for the no-review adaptive testing process, were performed using the Computer-Based Testing Simulation and Analyses Computer Program (CBTS) (Robin, 1999). The simulation of the rearrangement and the item review process was conducted using SAS (SAS Institute Inc, 1999).

Simulation Specifications

In order to determine the specifications for this simulation study, the adaptive testing literature was reviewed to make the simulation as realistic as possible (Ban,

Wang

Pats

Test S

when

achie

for tes

admir

was a

estima

with th

have a

Wang

pools

With th

examin

examin

answe

each it

$I(\theta,$

Wang, Yi & Harris, 2000; Camilli & Penfield, 1997; Camilli, Wang & Fesq, 1995; Eignor, Patsula & McLeod, 2000; Stocking, Way & Steffen, 1993; Vispoel, 1998a).

Test Specifications

In theory, a person's true (exact) ability estimate can be obtained from a test when an infinite number of test items are administered. However, this cannot be achieved in reality, since this process would be too tiring for examinees and too costly for testing organizations. In practice, only a relatively small sample of items can be administered to each examinee. For this reason, a 30 item, fixed length adaptive test was administered to each examinee since a 30 item test would be sufficient to properly estimate the examinee's abilities (McBride, Wetzel & Hetter, 1997). This is in accord with the psychometric literature on adaptive testing, where many adaptive tests tend to have approximately 30 items (Eignor, Stocking, Way & Steffen, 1993; Stocking, 1997; Wang & Vispoel, 1998).

The items that were administered in the simulation were selected from the item pools based on the maximum information procedure (McBride, Wetzel & Hetter, 1997). With the maximum information procedure, the items that are administered to the examinees are the ones that provide the maximum information at each of the examinees' current ability estimates. These estimates are calculated after the examinee answers each of the items on the test. The formula for estimating the information of each item is given in equation 1 below;

$$I\{\theta, u_i\} = \frac{a_i^2(1-c_i)}{(c_i + e^{L_i})(1+e^{-L_i})^2} \quad (1)$$

Where θ is the person's true ability

a_i is the discrimination parameter for item i

c_i is the pseudo-guessing parameter for item i,

$L=a_i(\theta - b_i)$,

b_i is the difficulty parameter for each item i, and

U_{ij} is the response to item i from person j,

Item pool development and characteristics

Two item pools were created for this study. Item pool 1 included 250 items, and was the more realistic item pool in terms of its size, since most real item pools contain about 250 items (Vispoel, 1998a; Wang & Vispoel, 1998). Item pool 2 contained 500 items. The items in this pool were increased compared to the size of item pool 1, to determine if better ability estimates can be obtained when the item pool is larger in size. It is hypothesized that the ability estimates of the examinees will be more accurate in the larger item pool since there would be a larger variety of items to administer, that would provide more information, and be targeted closer to the ability estimates of the examinee's true ability.

The item pools from which the items were selected were assumed to fit the three-parameter logistic model (3PL) (Lord, 1980), which is only used for dichotomously scored items. The 3PL model is described in equation 2.

$$P\{U_{ij}=1|\theta_j\}=c_i+(1-c_i)\frac{e^{a_i(\theta_i-b_i)}}{1+e^{a_i(\theta_i-b_i)}} \qquad (2)$$

Where θ_j is the person j's true ability

a_i is the discrimination parameter for item i

b_i is the difficulty parameter for item i

c_i is the pseudo-guessing parameter for item i,

U_{ij} is the response to item i from person j,

and $P\{U_{ij}=1|\theta_j\}$ is the probability of a correct response to item i from person j.

The scaling factor of 1.7 was not used anywhere in the simulation procedure. Since I did not have to compare the results to those from the normal ogive model, there was no need to include the scaling factor in the simulation process.

The psychometric literature was also reviewed to determine the item pool characteristics, so that the simulated item pools would be typical of the items used in real adaptive tests. Based on published item parameters and the means and standard deviations from various item pools, a sample item pool was defined for this study.

The real test items that were used as the basic reference for the creation of this item pool, were the following; From the Iowa Test of Educational Development, 300 items were used (Wang & Vispoel, 1998); 480 items from the ACT Assessment Program Math Usage test (Luecht & Hirsch, 1992); 30 items from the Basic Skills Test (Camilli & Penfield, 1997); 200 items from the Iowa Test of Educational Development (Vispoel, 1998a). In addition, 3600 items were used from the ACT mathematics test (Ban, Wang, Yi & Harris, 2000). From these 4610 items, the means and standard deviations were obtained for each of the three parameters, to serve as a model for the item parameters of the item pool. What is missing from the information provided by these items, is if the scaling factor of 1.7 was used in the scaling of these items.

Table 1 describes the targeted distributional characteristics of the two item pools that were recreated for this simulation. In terms of the distributions of the item parameters, the *a*-parameter, which is the index of discrimination, usually has a log normal distribution. So, the distribution of the a-parameter that was created for this study was a log normal distribution with a mean of 1.10 and a standard deviation of 0.25. The values of the a-parameter were also restricted to range between 0.45 and 2.3. The *b*-parameter, which is the difficulty index, had a uniform distribution. The reason for the use of this distribution was to have an adequate amount of items to assess the ability

levels of all the examinees. Although most item pools do not have a uniform distribution of b-parameters, the ideal goal of the test developers is to achieve this distribution. The mean for the b-parameter was 0.00 with a standard deviation of 2.0. The values of the b-parameters for the uniform distribution ranged from -3.5 to 3.5. The *c*-parameter, the pseudo-guessing parameter, also had a uniform distribution in this simulation. This is consistent with many studies on adaptive testing (Harwell, Stone, Hsu & Kirisci, 1996; Luecht & Hirsch, 1992). The mean of the c-parameter was 0.17 with a standard deviation of 0.10. Finally, the values of the c-parameter distribution ranged from 0.0 to 0.35. The range of values for all of the distributions of the parameters were chosen to represent the values of the parameters that currently exist in the adaptive testing literature (Eignor, Stocking, Way & Steffen, 1993; Harwell, Stone, Hsu & Kirisci, 1996; Luecht & Hirsch, 1992; Wang & Vispoel, 1998). Items whose upper and lower bounds fell outside of the pre-specified range, were eliminated from the item pools.

Table 1. Target distributional characteristics of the item parameters

	Mean	SD	Type of distribution	Minimum	Maximum
a parameter	1.10	0.25	Log normal	0.45	2.30
b parameter	0.00	2.00	Uniform	-3.50	3.50
c parameter	0.17	0.10	Normal	0.00	0.35

Ability Estimation Methods

Two estimation procedures were used for the estimation of the examinee abilities in the simulation. The first estimation procedure was the Maximum Likelihood (ML) procedure (Lord, 1980). Equation 3 is the formula for the iterative process used to obtain the ML estimates of the examinee abilities.

$$[\hat{\theta}_j]_{t+1} = [\hat{\theta}_j]_t + \left[\frac{\sum_{i=1}^{n} \alpha_i W_{ij}[(u_{ij} - P_{ij})/P_{ij}Q_{ij}][P_{ij}^*/P_{ij}]}{\sum_{i=1}^{n} \alpha_i^2 W_{ij}[P_{ij}^*/P_{ij}]^2} \right]_t \quad (3)$$

where t is the number of iterations,

where $Q_{ij} = 1 - P_{ij}$

where is $W_{ij} = P_{ij}Q_{ij}$ and

where $P_{ij}^* = \frac{P_{ij} - c_i}{(1 - c_i)}$

The second estimation procedure used in this study was the Owen's Bayesian Estimation procedure (Owen, 1975). In the Bayesian estimation procedure, as well as in the Maximum Likelihood procedure, examinees obtain new ability estimates after each item is answered. The posterior mean and variance formulas used for the estimation of the Bayesian estimate, are expressed as follows:

$$E(\theta_j \mid 1)_j = \bar{x}_{prior_j} + \frac{(1-c_i)V_{prior_j}(a_i^{-2} + V_{prior_j})^{-\frac{1}{2}}\phi(D_j)}{c_i + (1-c_i)\Phi(-D_j)} \quad (4)$$

$$E(\theta_j \mid 0)_j = \bar{x}_{prior_j} - \frac{V_{prior_j}(a_i^{-2} + V_{prior_j})^{-\frac{1}{2}}\phi(D_j)}{\Phi(D_j)} \quad (5)$$

$$\mathrm{var}(\theta_j \mid 1)_j = V_{prior_j}\{1 - (1-c_i)(1 + a_i^{-2}V_{prior_j}^{-1})^{-1}\phi(D_j)\cdot \frac{[(1-c_i)\frac{\phi(D_j)}{A_j} - D_j]}{A_j}\} \quad (6)$$

$$\mathrm{var}(\theta_j \mid 0)_j = V_{prior_j}\{1-(1+a_i{}^{-2}V_{prior_j}^{-1})^{-1}\phi(D_j).\frac{[\frac{\phi(D_j)}{\Phi(D_j)}+D_j]}{\Phi(D_j)}\} \quad (7)$$

where $A_j = c_i + (1-c_i)\Phi(-D_j)$ (8)

where $D_j = \frac{(b_i - \bar{x}_{prior_j})}{\sqrt{(a_i{}^{-2} + V_{prior_j})}}$ (9)

Φ(.) is the cumulative normal distribution that ranges from 0 to 1, and

φ(.)is the density of the function, which is the ordinate at θ,

var(θ|1)is the variance of the posterior distribution when a question is answered correctly,

var(θ|0) is the variance of the posterior distribution when a question is answered incorrectly,

E(θ/1) is the posterior expected value of θ when a question is answered correctly, and

E(θ/0) is the posterior expected value of θ when a question is answered incorrectly.

Examinee Characteristics

A group of 26000 examinees was simulated for this study. These simulees were created on 13 equally spaced θ levels. The goal was to have an adequate amount of simulees at each ability level so that the distribution of the ability estimates would be approximately normal. The θ level groupings were -3.0, -2.5, -2.0, -1.5, -1.0, -0.5, 0.0, 0.5, 1.0, 1.5, 2.0, 2.5, and 3.0. This is approximately the average number of intervals that were referenced in the computer adaptive test literature (Ban, Wang, Yi & Harris, 2000; Eignor, Stocking, Way & Steffen, 1993; Robin, Xing, Scrams & Potenza, 2000;

Patsula & McLeod, 2000). Table 2 includes the frequencies of examinees that were created at each ability level.

Table 2. Examinee frequency distribution

Examinee true ability level (θ)	Frequency
-3.0	78
-2.5	242
-2.0	728
-1.5	1712
-1.0	3154
-0.5	4548
-0.0	5133
0.5	4536
1.0	3138
1.5	1699
2.0	721
2.5	239
3.0	72

Simulation Of The Examinee's Test-Taking Behavior

According to Powers (1999), examinees tend to have higher anxiety when they are administered computer adaptive tests than with paper and pencil tests. For this reason, three types of simulated examinees were created for this simulation. The one group of examinees had no anxiety while taking the test. The second group of examinees had anxiety throughout the test (overall anxiety), due to their unfamiliarity

with the computerized adaptive testing format. These examinees were not able to perform to the maximum extent of their abilities on the test, so their ability while answering the items on the test for the first time, was 0.5 θ points lower than their true ability. There were 7800 examinees, which corresponded to 30% of the total sample, displayed overall test anxiety in the simulation. The specific percentage (30%) of examinees with overall anxiety was a close approximation to the results of a study by Powers (1999). According to Powers (1999), 25% of the examinees that he sampled, strongly agreed that they 'froze up' while taking the CAT GRE test, while another 28% of the examinees also strongly agreed that they felt very panicky while taking the CAT test.

Powers also found that 60% of the same sample reported that they 'felt unsure and tense while taking the CAT GRE General Test'. For this reason, a third group of examinee was created. These are the examinees who have high anxiety at the beginning of the test, but who are able to overcome their anxiety after they start moving through the test. It was hypothesized that the 'start anxiety' of such examinees will prohibit them from performing to the maximum extent of their abilities on the first 7 items on the test. So another 30% of all the examinees answered the first 7 items on the test as if their ability estimates were 1 θ points lower than their true ability estimates (van der Linden & Krimpen-Stoop, 2001). This percentage of examinees was also based on the results found by Powers (1999).

If the examinees were allowed to go back and change their answers on the test, there would be three types of questions they would consider changing their answers to. The first type of questions that would be changed, are the questions in which the examinees made 'stupid mistakes' such as calculation errors, even though they had the ability to answer those items correctly. Therefore, if the examinees had the opportunity to go back and change their answers, those answers would be changed from incorrect to correct ones. In the simulation procedure those cases were the questions to which

examinees had an 0.80 or higher probability of answering correctly, but were answered incorrectly. So all of those answers would be changed by the examinees from incorrect to correct answers.

The second type of questions that the examinees would consider changing their answers to, are the ones that were very difficult for them. If these questions were too difficult for the examinees, it is very likely that they would go back to reread those questions and rethink about their answers. In some cases, the examinees might select the correct answer the first time that they went through the test, just by chance. In the simulation procedure, those cases would be identified by the questions to which the examinees only had a 0.33 or lower probability of answering them correctly, but were answered correctly. Therefore, it was hypothesized that if the examinees were able to go back and reread those questions, they would change their answers from correct to incorrect answers with a probability of 1.0.

The third type of questions that the examinees would consider changing, would be the questions that were well matched to their true abilities. In this case, the examinees would be unsure of their answers to such questions, and could wish to reread and rethink their answers. In the simulation procedure, these cases were identified by the questions to which the examinees had approximately a 0.50 probability (0.47- 0.53) of answering correctly. A meta-analysis study that examined the examinee's item changing behaviors has shown that 72% of the examinees that change their answers on tests, change them from incorrect to correct answers (Waddell & Blankenship, 1994). For this reason, a binomial random number was generated with a 0.72 probability of answering the item correctly. So the examinees that had originally answered such a question incorrectly, would change their answer to a correct one with a 0.72 probability. A second binomial random number would be generated, for those questions that were answered correctly, with a probability of 0.28. So there would be a

0.28 probability that the examinees would change their answer on that item from a correct to an incorrect answer, based on this random number.

In the cases where examinees had more items that needed to be reviewed than the number of items that were permitted, then the items that would eventually be reviewed were randomly selected by the simulation procedure.

Item Revision Algorithm- The Rearrangement Procedure

The rearrangement procedure will not be visible from the perspective of the examinees. All that they will know is that they will be allowed to change up to 5 of their answers on the test. No additional time will be provided for the examinees to change their answers. Only if they finish answering all 30 items on the test before the end of the allotted time will they be allowed to review their answers. The time limit for the completion of the test will be fixed. So if an examinee manages to finish answering all 30 items on the test before the time limit has expired, that examinee will also have the opportunity to revise their answers. If an examinee does not manage to finish answering all of the items on the test by the end of the time limit, no additional time would be provided for them to make any revisions.

So, if item review were permitted on a 30-item test, and the examinees finished answering all the items before the time limit expired, they would have the option to go back to review and possibly change any of their answers. The test would then officially terminate either at the end of the time limit, or after the examinees finished making up to five changes to their answers on the test, whichever came first. (It should be noted that another one of the conditions of the simulation permitted the examinees to change only up to three answers on the test). The rearrangement procedure will then take place after responding to all of the 30 items. However, the examinees will not be aware of the

rearrangement procedure since it will be part of the estimation procedure algorithm that would be used to obtain the examinee's final ability estimate.

Item Skipping In The Rearrangement Procedure

One of the strengths of adaptive testing, is that the items that are administered are selected to match the examinee's most recent ability estimate. This matching of items to the examinee's most recent estimate permits more efficient and accurate estimation of the examinee's ability. The ultimate goal is to obtain ability estimates that are as close to the examinee's true ability levels as possible.

However, with item review, after the answer to an item *i* is changed, the items that follow might no longer be as appropriate for estimating the examinee's posterior ability estimate. Therefore, instead of administering items that are not as appropriate for a new ability level, the rearrangement procedure will skip these items. The rearrangement procedure will then try to find an item *i* +k, that is more appropriate for the posterior ability estimate. It is hypothesized that by administering fewer items that are better targeted to an examinee's ability estimate, the final ability estimate will be less biased and closer to the examinee's true ability level than when less appropriate items are administered. This is consistent with Reckase (1975) who found that the bias of the ability estimates tended to increase by administering extreme items that were not properly targeted to the examinee's ability levels.

Types Of Answer Changing And The Rearrangement Procedure

There are three types of answer changes that could be made by the examinees; changing responses a) from an incorrect to an incorrect response, b) from an incorrect to a correct response, and c) from a correct response to an incorrect response.

Type 1 change. Incorrect to incorrect changes

If an examinee changes an answer from an incorrect option to another incorrect option, then no changes need to be made to the ability estimation of the examinee, and the examinee will obtain the same score as they did before the review. In addition, no change will take place in terms of the accuracy of the standard error of the test.

Type 2 change. Incorrect to correct changes

The second type of answer change that examinees can make, is the change from an incorrect to a correct answer. If this change were made to item *i*, the ability estimation $\hat{\theta}_i$ will be changed to $\hat{\theta}_i$'. However, if this occurs, question *i* +1 would probably not be the most informative item for the ability $\hat{\theta}_i$' since it would be easier and targeted at lower ability levels than $\hat{\theta}_i$'. This is a problem in adaptive testing, because it will cause the bias of the final ability estimate to increase (Reckase, 1975).

To solve this problem, the computer algorithm will skip question *i* +1 in the ability estimation procedure, since that would no longer be an appropriate item for that ability level. The algorithm of the rearrangement procedure will then jump to the first item X after question *i* +1 (e.g. item *i* +k, with 1<k<4) that was answered incorrectly since it was more difficult. It is hypothesized that this new item *i* +k would be more similar to the item that would have been administered after item *i*, if item *i* were answered correctly in the first place. So after the skipping of items *i* +1 through *i* +k-1, the rest of the test would remain the same, and no changes would be made to the test if no other answers were changed. So the next step would be to recalculate the ability estimate based on the rest of the items in the order that they were presented, until the end of the test. However, in

this s

leve

to ite

abil

diffic

until

these

next

it finds

answe

item th

answe

ignored

item / w

This wo

describe

where

correct

items /

this specific case, a total of 30-(k-1) items would be used to estimate the final ability level.

Figure 1 provides an example in which an incorrect-to correct change was made to item 2 of a test. In this case, question 3 was skipped since it was targeted at a lower ability level than $\hat{\theta}_2'$. So the algorithm jumped to item 4 since that was the first more difficult item that was answered incorrectly, that came after item 3.

However, it is also possible for the rearrangement procedure to jump 2 or 3 items until it finds the next incorrect item. In case 3 items have been skipped and none of these answers are incorrect, then the 4th item after the answer-changed-item will be the next item that will be used for the estimation of the examinee's ability estimate.

However, it is also possible for the rearrangement procedure to skip 3 items until it finds the next correct item. In case 3 items have been skipped and none of these answers are correct, then the 4th item after the answer-changed-item will be the next item that will be used for the estimation of the examinee's ability estimate.

Type 3 change. Correct to incorrect changes

If an examinee decides to change another item on the test (e.g. item *I*), and the answer is changed from a correct answer to an incorrect answer, item *I* +1 would be ignored in the ability estimation procedure. The reason for ignoring that item is because item *I* would be targeted at a higher ability level than $\hat{\theta}_I$ so it would be more difficult. This would result in a larger standard error of the final ability estimate. Figure 2 describes this situation. Therefore, the computer would select item Y (e.g. item *I* +K where 1<k<4) if that was the first item after item *I* that was easier since it was answered correctly. So it is hypothesized that the ability estimation would be more accurate if items *I* +1 through *I* +K-1 were ignored from the estimation procedure, and item *I* +K was

used after the item whose answer was changed. This is done because it is hypothesized that item *I* +K would be more similar to the item that could have been administered after item *I*, if item *I* was answered incorrectly in the first case. The next step would be to recalculate the ability estimate from the rest of the items in the order they were presented. In this case, a total of 30-(K-1) items would be used to estimate the final ability level.

Figure 1. Example of an incorrect-to-correct answer change on a CAT (type 2 change)

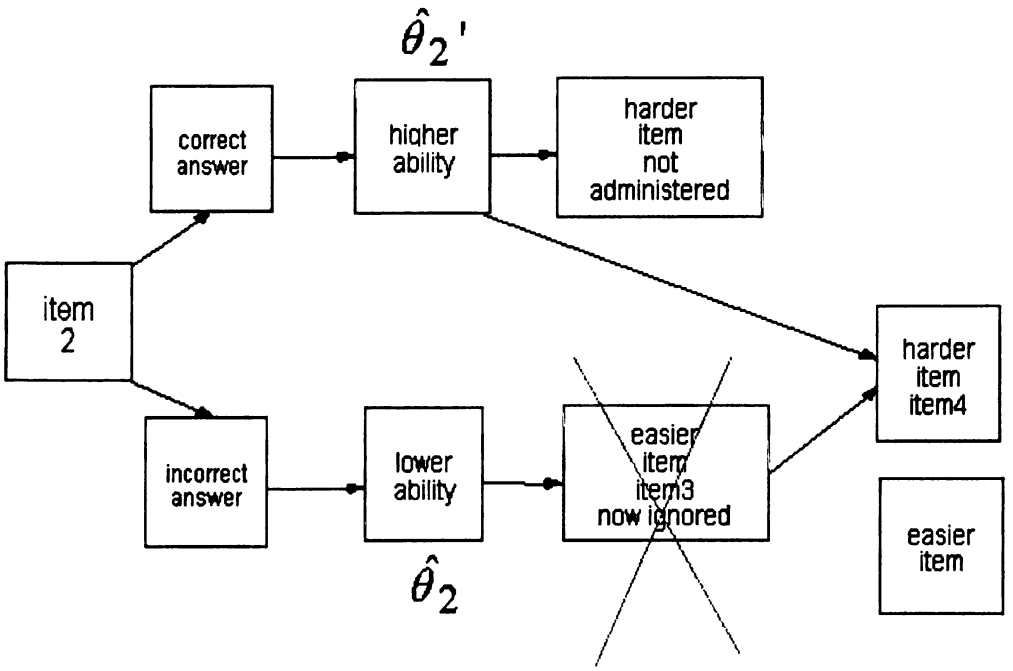
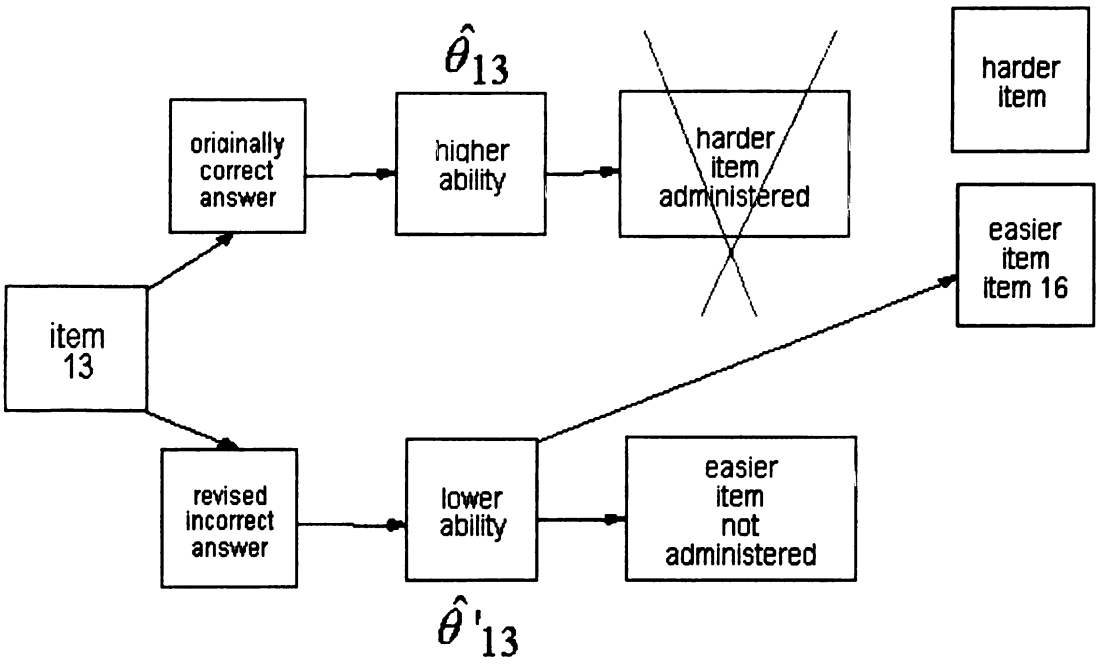


Figure 2 provides an example in which the answer to item 13 was changed from a correct to an incorrect answer. In this case, items 14 and 15 were too difficult for the examinee's new ability estimate $\hat{\theta}_{13}'$. Item 16 was the first easier item that was answered correctly that came after item 13. For this reason, the rearrangement procedure skipped items 14 and 15, and item 16 was used next in the estimation procedure.

Figure 2. Example of a correct-to-incorrect answer change on a CAT (type 3 change)



Making Two Or More Answer Changes: Rearranging Items In The Rearrangement Procedure

Consider the case in which an examinee makes two changes in his/her response patterns. This examinee might change the response to item 2 (from an incorrect to a correct response), and the response to item 13 (from a correct to an incorrect response). When the first change takes place, the algorithm will follow the same procedure as in the type 2 change. So item 3 would be ignored in the estimation procedure, and item 4 would be selected if that was the first item (after item 2) that was answered incorrectly. When the examinee continues through the test and changes the response to item 13 from a correct to an incorrect response, the algorithm would make a comparison to determine which items to use next in the estimation procedure. This determination would be made from the information that is provided by a) item 16, which is the first item after item 13 that was answered correctly, and b) any items that had been skipped in the estimation procedures at previous steps in the algorithm, such as item 3. The item that would provide the most information out of the two at $\hat{\theta}_{13}$ ',would be selected as the item that would replace item 14 that was skipped by the algorithm. The next step would be to recalculate the ability estimate from the rest of the items in the order that they were presented, until the end of the test. Figure 3 provides a hypothetical example of a convergence plot where no rearrangement took place since item 16 was more informative than item 3 at the posterior θ level of $\hat{\theta}_{13}$.

Estimated ability θ

be use

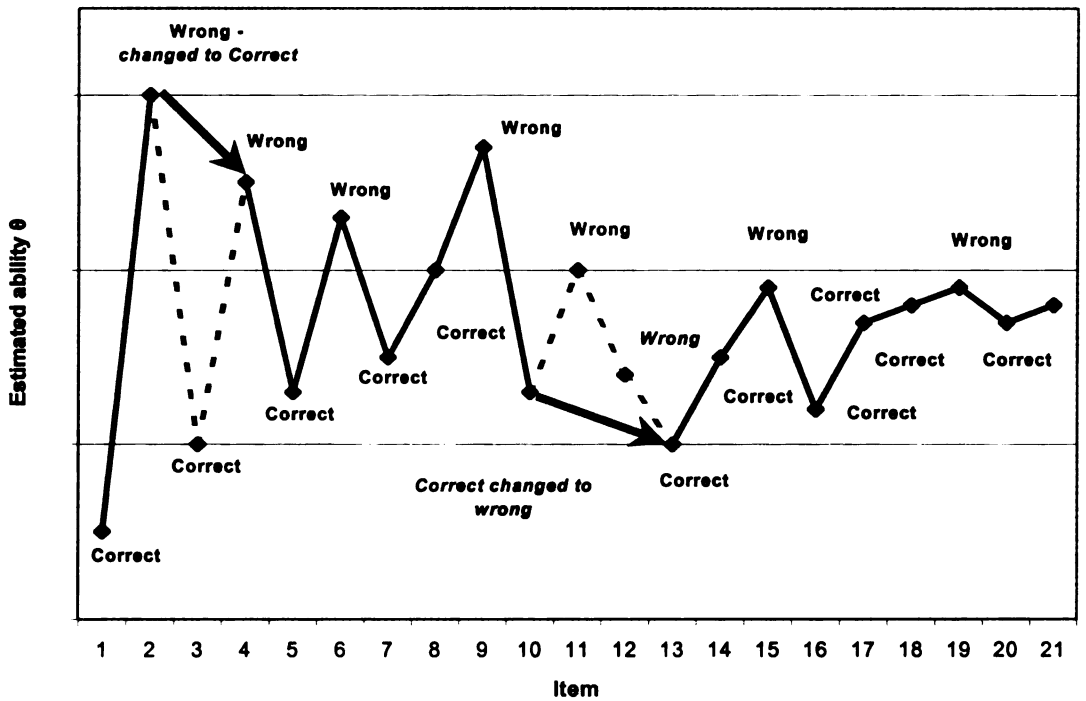
rearra

estima

23. 24.

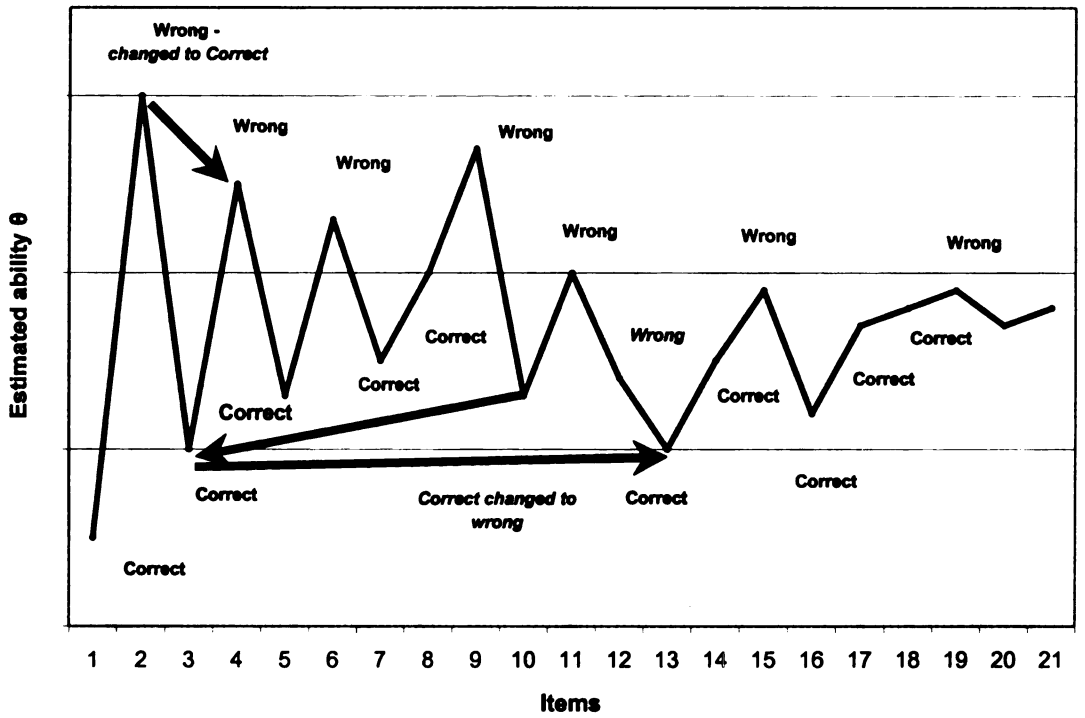
with a

Figure 3. Rearrangement procedure without a rearrangement of the item order



If item 3 was more informative than item 16 at the ability level $\hat{\theta}_{13}$', item 3 would be used after item 13 for the estimation of the examinee's ability estimate. So the rearranged order in which the items will be used for the estimation of the final ability estimate is the following: 1, 2, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, **3,** 16, 17, 18, 19, 20 21, 22, 23, 24, 25, 26, 27, 28, 29, 30. Figure 4 describes the above pattern of item responses, with a hypothetical set of data.

Figure 4. Rearrangement procedure with a rearrangement of the item order



Convergence Plots

Figure 5 describes the convergence plot of the ability estimates of a simulee who has a true ability of 0.00. This figure, which is based on the simulated data that are used in this study, reflects the way in which the Bayesian estimation procedure converges to the simulee's final ability estimate. The convergence is examined three times, which is once with each of the three points of the rearrangement process.

This examinee had originally answered item 2 correctly in the simulation. After item review, however, this examinee changed their answer to question 2 to an incorrect answer. Consequently, the rearrangement procedure skipped items 3 and 4, and continued with the use of item 5, which was answered correctly. So after the rearrangement procedure, the posterior θ after item 5 was -0.036. This θ estimate which

was

0.23

proc

(-0.1

Ability estimates (θ)

chan

items

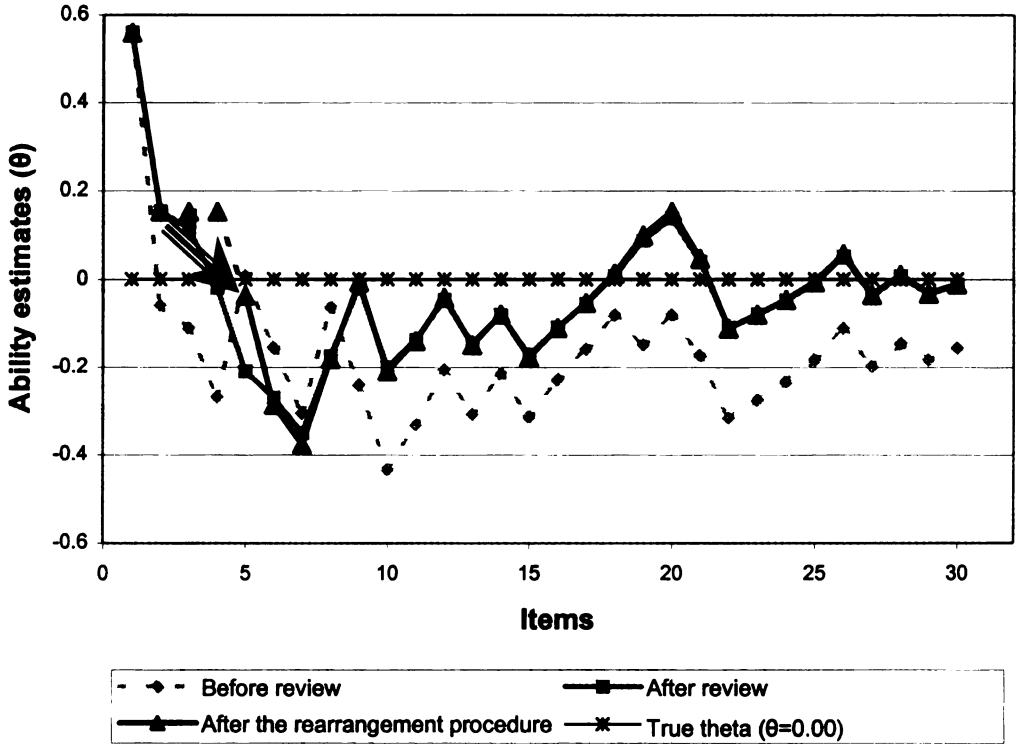
incor

estim

the fo

was closer to the examinee's true ability of 0.00 than the estimate after review that was -0.2094. Eventually, the examinee's final ability estimate after the rearrangement procedure was 0.0098. This was closer to the true score than the estimate before review (-0.1565) as well as the estimate after review (0.0125).

Figure 5. Convergence plot with correct-to-incorrect change



Exceptions to the Rule

A possible problem might exist in the cases where an examinee for example, changes an answer from an incorrect to a correct one, but there are no other appropriate items to replace them. More specifically, there might be no items that were answered incorrectly after the item whose answer was changed that could be used by the estimation procedure. In this case, the procedure would skip three items, and then use the fourth item that comes after the item to which the answer was changed. The same

situation could occur when an examinee changes an answer from a correct to an incorrect one, but there are no other items answered correctly that could be used by the rearrangement procedure. Again, like in the previous example, the procedure would skip three items, and then use the fourth item that comes after the item to which the answer was changed.

A second exception to the rule includes the case in which an examinee changes the last item on the test. In this case, no additional changes would have to be made to the estimation procedure, and the final (correct or incorrect) answer would be used to estimate the final ability estimate.

A third exception to the rule would be in the case where more than 3 items have already been skipped. This would cause a problem to the estimation procedure since the examinee's ability estimate would be much worse since there would be too few items that could be used for the estimation. For this reason, no items will be skipped if three items have already been skipped because of the rearrangement procedure.

Stopping rules

In order to avoid possible cheating strategies used by examinees, some restrictions would also have to be made on the revision policy. A meta-analysis conducted by Waddell and Blankenship (1994) found that the mean percentage of items changed in 75 studies was 5.1% when examinees have the option of revision. This means that on a 30 item test, an average of only 1.5 items are changed. So any large deviation beyond 15% might be an indicator that an examinee is trying to cheat. Therefore, a limit would have to be placed on the number of revisions that would be allowable for examinees to make. This would prohibit the Wainer strategy from taking place. So in the case of a 30 item test, a maximum of 5 items would be permitted to be changed for the rearrangement procedure. This should not appear as a major restriction

to the examinees since the typical examinee would only change about 2 items out of 30. It should also be noted that if an examinee changes their answer to the same question two times, that would count as one revision, not two.

Dependent Variables

The effects of the rearrangement procedure can be judged in many ways. Three dependent variables were used to help determine the effects that the rearrangement procedure had on the examinees' ability estimates; the bias, the conditional standard error, and the reliability estimate (Kim & Nicewander, 1993). The bias of the final ability estimate was calculated to determine how much the examinees' estimated scores deviated from their true scores. The formula for the bias is shown below:

$$\text{Bias}_i = \hat{\theta}_i - \theta_i \qquad (10)$$

Where $\hat{\theta}_i$ is an examinee's estimated ability and

θ_i is an examinee's true ability

Another way of judging the quality of the results was by estimating the reliability of the ability estimates before review, after review, and after the rearrangement procedure. Formula 11 was used to estimate the reliability of the examinees' ability estimates.

$$\rho_{\hat{\theta}\theta} = \frac{\sigma^2_{\theta}}{\sigma^2_{\theta} + \sigma^2_{\hat{\theta}/\theta}} \qquad (11)$$

where σ^2_{θ} is the variance of the examinee's true ability and

$\sigma^2_{\hat{\theta}/\theta}$ is the conditional variance of the ability estimates

A third and final way of judging the effects of the rearrangement procedure, is by comparing the conditional estimates of the standard error. The standard error discussed in this study is actually the standard deviation of the ability estimates, at each true ability level.

Independent Variables

The independent variables used in this study are the item pool size, the maximum amount of items that are allowed to be reviewed, the estimation procedures, as well as the examinee anxiety conditions. The item pools are of two sizes, 250 and 500 items. The maximum amount of items that are permitted to be reviewed is either 3 or 5 items. The estimation procedures used are the Bayesian and Maximum Likelihood procedure. Finally, there are three types of anxiety conditions; no anxiety, start anxiety, and overall anxiety.

CHAPTER 4

RESULTS

The purpose of this study is to test the efficiency of a rearrangement procedure that rearranges and skips certain items in order to better estimate the examinee's ability estimates. This procedure takes place only after the examinees have had a chance to change any of their answers on an adaptive test. More specifically, the research questions that will be answered in this section of the study are the following:

1. What are the effects of the rearrangement procedure on the reliability of the estimates?
2. How much statistical bias and error does the rearrangement procedure create?
3. How does the rearrangement procedure affect the ability estimates for the examinees who have anxiety because of the computerized format of adaptive tests?
4. How does the choice of the ability estimation procedure affect the estimates after the rearrangement procedure?

The effects of the rearrangement procedure have been examined under four conditions that are presented in Table 3. Two of the conditions included item pools that had 250 items. This size was selected since most real item pools include approximately 250 items. The other two conditions included larger item pools of 500 items. In addition, conditions 1, and 3 examined the magnitude of the results when the examinees were permitted to change up to 3 of their answers on the test. Conditions 2 and 4 were examined to determine the magnitude of the results when the examinees were permitted to change up to 5 of their answers on the test. For all 4 conditions of the simulation, the sample of examinees was the same.

Table 3. Replication conditions of the simulation

Condition	Item pool size	Maximum number of items changed
1	250	3
2	250	5
3	500	3
4	500	5

The results of this study will be discussed according to the four conditions that were examined in the study. In each condition, the effects of the rearrangement procedure will be discussed in terms of the standard error, the bias, and the reliability of the final simulated estimates. In addition, the effect of the rearrangement procedure on the examinees with anxiety will also be presented. Finally, a comparison will be made between the effects of the item pool size, of the maximum number of items that are allowed to be changed, and of the Maximum Likelihood and Bayesian estimation procedures.

Condition 1: 250 Items With 3 Reviews Maximum

The first condition with which the rearrangement procedure was examined, was where the examinees were allowed to make up to three revisions to their answers on the test, and where the item pool used to create the adaptive tests consisted of 250 items. To determine the accuracy and effectiveness of the rearrangement procedure (RP), the ability estimates were obtained three times, at the three points of the rearrangement process; before review, after review, and after the rearrangement procedure. The before review time point is the one before the examinees in the simulation had the opportunity to change their answers on the test. The point after review describes the ability estimate after the examinees in the simulation had the opportunity to revise and change their answers. Finally, the after the rearrangement procedure (ARP) time point describes the

ability estimates in the simulation after the rearrangement procedure was used. At each of the three points, the bias, the standard error and the reliability were estimated with two methods. They was obtained once when the ability estimates were obtained with the Maximum Likelihood (ML) estimation procedure and once again with the Bayesian method.

Overall, 40.4% of the examinees in the simulation made correct-to-incorrect, or incorrect-to-correct changes to their answers. These types of changes are the only ones that will be discussed since the rearrangement procedure takes place only when such changes have occurred. Table 4 describes the percentage of actual answer changes, that are divided in four categories. The first type of answer change is the one where examinees made an incorrect-to-correct change, to an item where they had approximately a 0.50 probability of answering correctly. The second type of answer change is the one where examinees made a correct-to-incorrect change, to an item where they had approximately a 0.50 probability of answering correctly. The third type of answer change is the one where examinees made an incorrect-to correct change to an item that they originally answered incorrectly by mistake, although they had the ability to answer it correctly in the first place. These are the items to which the examinees had a probability of 0.80 or higher of answering correctly. The fourth type of answer change is the one where examinees had a lower than 0.33 probability of answering correctly, but were originally answered correctly just by chance. After the item review, the answers to these items were changed from correct to incorrect answers.

As can be seen from Table 4, the majority of the changes that were made in the simulation (40.40%) were from incorrect to correct ones. In addition, the majority of those changes were from examinees that made one or two such changes throughout their test. There were also 14.52% of the simulated examinees that made correct-to-

incorrect changes to questions to which they had approximately a 0.50 probability of answering correctly.

Only 5.64% of the simulated examinees had made 'stupid mistakes' that were then changed to correct answers. Finally, there were also 3.04% of the same examinees that changed their answers to incorrect answers to an item that was originally answered correctly just by chance.

Table 4. Percentage of actual answer changing patterns in condition 1

	Number of changes	Number of examinees	Percentage of examinees (out of 26000)
Incorrect-to-correct changes (0.5 probability)	**1**	3806	14.64%
	2	3765	14.48%
	3	2933	11.28%
	4	0	0.00%
	5	0	0.00%
	Sum	**10504**	**40.40%**
Correct-to-incorrect changes (0.5 probability)	**1**	3064	11.78%
	2	625	2.40%
	3	85	0.34%
	4	0	0.00%
	5	0	0.00%
	Sum	**3774**	**14.52%**
Stupid' mistake corrections (incorrect-to-correct)	**1**	1167	4.49%
	2	298	1.15%
	3	0	0.00%
	4	0	0.00%
	5	0	0.00%
	Sum	**1465**	**5.64%**
Unlucky guess' changes (correct-to-incorrect)	**1**	639	2.46%
	2	150	0.58%
	3	0	0.00%

4	0	0.00%
5	0	0.00%
Sum	**789**	**3.04%**

After the examinees reviewed their items on the test, the rearrangement procedure was used. Because of the rearrangement procedure, there were 908 examinees (5.86%) in the simulation that used item review, to which 1 of their items on the test were ignored. There were also 767 examinees (4.95%) that used item review, to which 2 of their items on the test were ignored. Finally, there were also 8083 examinees (52.16%) of the examinees that used item review, to which 3 of their items on the test were ignored.

When items were rearranged because of the rearrangement procedure, the amount of information that was provided at each ability level was used as an indicator for which item should be selected to be used next. The average amount of information that was gained by rearranging the items in condition 1 was 0.0513 θ with a standard deviation of 0.0426. The minimum amount of information that was gained was 0.0001, while the maximum information that was gained was 0.3079.

Results Based On Bias (Condition 1)

Table 5 describes the average bias and standard deviation of both estimation procedures at each of the three time points of the rearrangement procedure. These results are averaged over all of the examinees that were simulated in the sample, including the ones with test anxiety. According to Table 5, the ML bias estimate before review was -0.1374. After review, the ML bias dropped in magnitude to 0.0673. After the rearrangement procedure, the bias decreased even further to 0.0567. So the ARP ML bias improved by 15.7% when compared to the bias that existed after review.

Table 5. Overall bias of the 250 pool with 3 reviews estimates (Condition 1)

Bias	Mean	Standard Deviation
Maximum Likelihood Bias		
Before review	-0.1374	0.3229
After review	0.0673	0.3308
After the rearrangement procedure	0.0567	0.3355
Bayesian Bias		
Before review	-0.2641	0.3099
After review	-0.0687	0.2964
After the rearrangement procedure	-0.1087	0.2935

The results that were based on the Bayesian estimation showed a different pattern of bias. The Bayesian bias estimate before review was -0.2641. After review, the Bayesian bias dropped in magnitude to -0.0687. After the rearrangement procedure, the bias increased in magnitude to -0.1087. This was a 58.2% increase in the bias when compared to the bias after review. By comparing the overall results in terms of bias, the ML estimates tend to be more accurate. The smallest bias exists with the ML estimate when it is estimated after the rearrangement procedure. This bias of -0.0567 is even lower than the smallest Bayesian bias (-0.0687) that is produced only after item review has taken place.

Table 6. Conditional Maximum Likelihood bias when 3 reviews are permitted with a 250 sized item pool (Condition 1)

Ability θ	Before Review	After Review	After Rearrangement procedure (APR)	Improvement from rearrangement procedure
-3.0	-0.1261	0.0463	-0.1273	
-2.5	-0.2234	-0.0450	-0.0661	
-2.0	-0.1765	-0.1072	-0.1152	
-1.5	-0.1284	0.0656	0.0514	Yes
-1.0	-0.1469	0.0388	0.0291	Yes
-0.5	-0.1595	0.0439	0.0331	Yes
0.0	-0.1398	0.1065	0.0949	Yes
0.5	-0.1422	0.0722	0.0605	Yes
1.0	-0.1034	0.0242	0.0203	Yes
1.5	-0.0883	0.1859	0.1686	Yes
2.0	-0.1442	0.2012	0.1918	Yes
2.5	-0.1087	-0.0233	-0.0114	Yes
3.0	-0.0501	-0.0504	0.1033	

The bias results are described more analytically in Table 6, which presents the conditional ML bias at each of the 13 ability levels from which the examinees were sampled. In some cases, such as at the θ levels of 1.5 and 2.0, the after review bias was larger than the before review bias. The reason for this increase is because in certain cases, the review process eliminated the randomness from the examinee's responses. This resulted in a mismatch between the examinee responses and the IRT model. For

example, examinees with an ability of θ=2.0 might have had a 90% probability of answering item i correctly. Consequently, it is expected that 90% of the examinees with a θ of 2.0 would answer item i correctly, and 10% would answer the item incorrectly. However, if 100% of these examinees answer the item correctly, then their response patterns will not match the IRT model, which consequently will increase the after review bias of the ability estimates.

Table 6 also shows that there were 9 out of the 13 ability levels where the ML bias decreased in magnitude from the after review time point to the ARP time point. These improvements existed at the θ levels of -1.5 to 2.5. So the effects of the rearrangement procedure were generally more effective at the positive rather than the negative end of the θ scale.

Table 7 presents the conditional bias when the Bayesian estimation procedure was used. The rearrangement procedure was not very effective in reducing the bias of the Bayesian ability estimates after review at most of the ability levels. The only exceptions were the θ levels of - 3.0 and -2.5 where the bias decreased in magnitude with the rearrangement procedure. However, even at the rest of the ability levels, the bias after the rearrangement procedure was still smaller in magnitude than the bias that existed before item review.

Table 7. Conditional Bayesian bias when 3 reviews are permitted with a 250 sized item pool (Condition 1)

Ability θ	Before Review	After Review	After Rearrangement procedure (APR)	Improvement from rearrangement procedure
-3.0	-0.0469	0.07274	0.0494	Yes
-2.5	-0.1237	0.02924	0.0030	Yes
-2.0	-0.1808	-0.07141	-0.0852	
-1.5	-0.1979	-0.01737	-0.0477	
-1.0	-0.2309	-0.04596	-0.0835	
-0.5	-0.2621	-0.07826	-0.1126	
0.0	-0.2680	-0.0388	-0.0895	
0.5	-0.2839	-0.06853	-0.1237	
1.0	-0.2874	-0.16386	-0.1721	
1.5	-0.2980	-0.04969	-0.1091	
2.0	-0.3577	-0.0599	-0.1261	
2.5	-0.3570	-0.22348	-0.2498	
3.0	-0.3656	-0.1576	-0.1956	

Figure 6 provides a comparison of the ML and Bayesian conditional bias after the rearrangement procedure has taken place. The overall pattern of the results shows that the ML bias tends to increase from a negative to a positive bias as the ability of the examinees increases. Consequently, examinees with lower ability estimates tend to have a negative ML bias, while examinees with abilities higher than -2.0 tend to have a positive ML bias. So examinees at the lower end of the distribution have lower estimated scores than true scores when the ML is used with the rearrangement procedure. The

situation is the opposite for examinees at the higher end of the ability distribution where a positive bias exists. This is expected since the ML estimator is more biased towards the extremes (Kim & Nicewander, 1993; Lord, 1986; Wang & Vispoel, 1998). So this pattern of bias is a function of the ML estimation procedure, rather than a function of the rearrangement procedure.

In contrast to the ML bias, the Bayesian estimator is biased towards the mean (Kim & Nicewander, 1993; Lord, 1986; Wang & Vispoel, 1998). So examinees at the lower end of the ability distribution tend to have higher score estimates than true scores when the Bayesian is used with review and with the rearrangement procedure. This is because the Bayesian bias is positive at the lower end of the distribution. Examinees at the higher end of the distribution who have negative bias have lower Bayesian ability estimates than their true scores after the rearrangement procedure is used. However, the bias towards the mean is a function of the Bayesian estimation procedure, rather than a function of the rearrangement procedure.

Figure 7 describes the percentage of bias reduction that has occurred from the after review estimates to the ARP estimates. The ML estimation procedure appears to work well at most ability levels since it has a positive percentage of bias improvement at 9 of the 13 ability levels. This means that the ML bias decreased in magnitude due to the rearrangement procedure. The ML estimation procedure was problematic, though for the examinees whose abilities were around -3.0 θ and 3.0 θ. The bias produced by the rearrangement procedure for the examinees at the -3.0 θ was actually 175.0% worse than the bias that existed after review. The ML bias became worse by 105.2% for the examinees whose true θ was 3.0. However, the Bayesian bias tended to increase at all but 2 θ levels. These were the levels of -3.0 and -2.5. So in terms of bias, the ML estimates tend to be more accurate than the Bayesian estimates when the rearrangement procedure is used.

Figure 6. Condition 1 ML and Bayesian Bias after the rearrangement procedure (250 pool and 3 reviews)

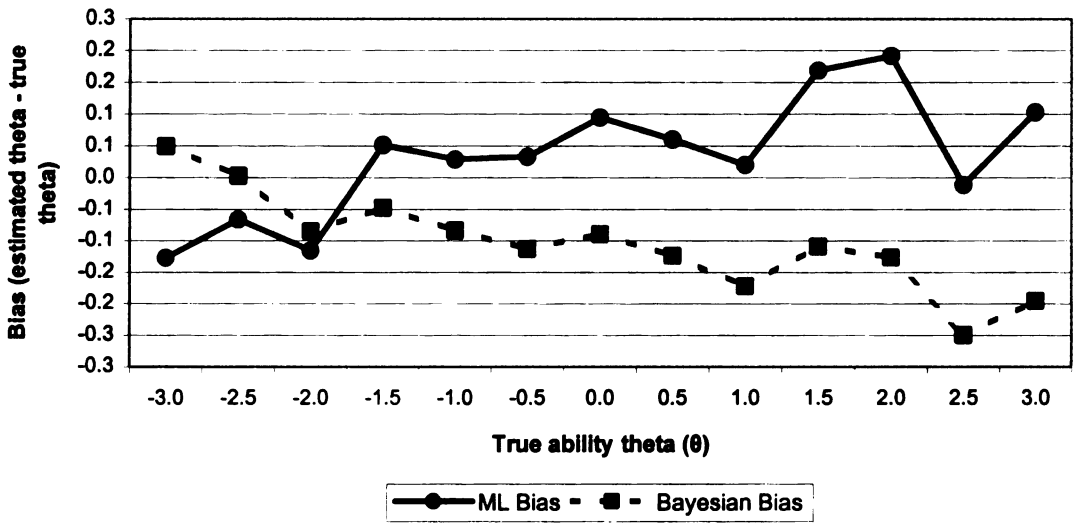
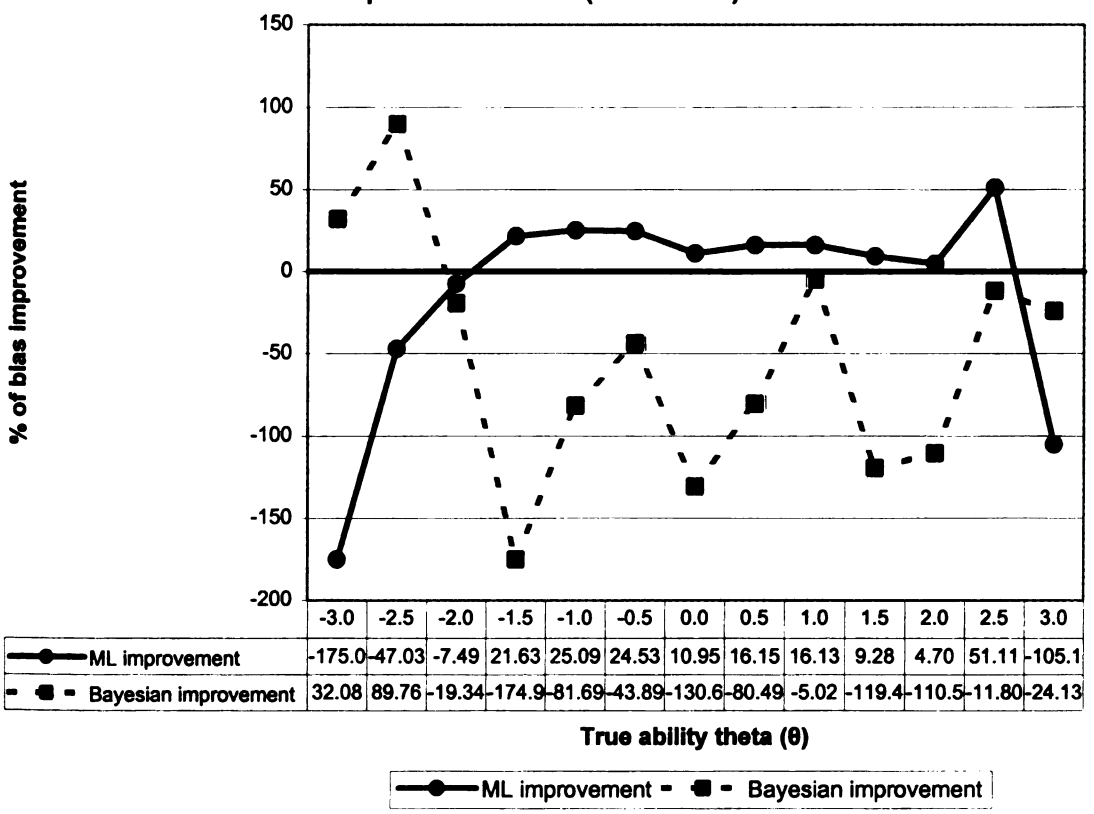


Figure 7. Percentage of ARP bias improvement with 3 changes and pool of 250 items (condition 1)



Results Based on the Standard Error of the θ Estimate (Condition 1)

Table 8 describes the average standard deviation (SD) of the θ estimates obtained from both estimation procedures at each of the three time points of the rearrangement procedure. These results are averaged over all of the simulated examinees in the sample, including the ones with test anxiety. According to Table 8, the ML SD before review was 1.0728. After review, the ML SD increased to 1.0878. After the rearrangement procedure, the standard deviation increased further to 1.0928. This was a 0.4% increase in the SD when compared to the SD after review.

Table 8. Overall standard deviation of the θ estimates obtained from the pool of 250 items, when 3 reviews were permitted (Condition 1)

	Standard Deviation of the θ estimates
Maximum Likelihood	
Before review	1.0728
After review	1.0878
After the rearrangement procedure	1.0928
Bayesian	
Before review	1.0187
After review	1.0273
After the rearrangement procedure	1.0223

The results that were based on the Bayesian estimation showed a different pattern of standard deviation. The Bayesian SD before review was 1.0187. After review, the Bayesian SD increased to 1.0273. After the rearrangement procedure, the SD

decreased to 1.0223. So the ARP Bayesian SD improved slightly by 0.4% when compared to the SD that existed after review.

Table 9. Conditional Maximum Likelihood standard error when 3 reviews are permitted with a 250 sized item pool (Condition 1)

Ability θ	Before Review	After Review	After Rearrangement procedure (APR)	Improvement from rearrangement procedure
-3.0	0.2331	0.1130	0.3366	
-2.5	0.3313	0.3139	0.3327	
-2.0	0.3570	0.3296	0.2947	Yes
-1.5	0.3204	0.2897	0.2953	
-1.0	0.3014	0.2905	0.3574	
-0.5	0.3334	0.3536	0.3214	Yes
0.0	0.3109	0.3172	0.3392	
0.5	0.3165	0.3371	0.3562	
1.0	0.3562	0.3519	0.2997	Yes
1.5	0.3026	0.2941	0.3552	
2.0	0.2966	0.3406	0.3774	
2.5	0.3702	0.3474	0.2191	Yes
3.0	0.2836	0.1058	0.2229	

The standard error (SE) results are described more analytically in Table 9, which presents the conditional ML SE at each of the 13 ability levels from which the examinees were sampled. Table 9 shows that there were only 4 out of the 13 ability levels where the ML SE decreased in magnitude from the after review time point to the ARP time

point. These improvements existed at the θ levels of -2.0, -0.5, 1.0, and 2.5. So the effects of the rearrangement procedure were generally less effective when the effects were determined based on the ML standard errors. This is not surprising since the standard error tends to increase when the length of a test is shortened, which is the case with the rearrangement procedure.

Table 10. Conditional Bayesian standard error when 3 reviews are permitted with a 250 sized item pool (Condition 1)

Ability θ	Before Review	After Review	After Rearrangement procedure (APR)	Improvement from rearrangement procedure
-3.0	0.3086	0.2637	0.2726	
-2.5	0.2766	0.2448	0.2525	
-2.0	0.3049	0.2644	0.2666	
-1.5	0.3024	0.2654	0.2634	Yes
-1.0	0.2984	0.2661	0.2621	Yes
-0.5	0.3124	0.3170	0.3115	Yes
0.0	0.3037	0.2966	0.2948	Yes
0.5	0.3012	0.3027	0.2933	Yes
1.0	0.3232	0.2954	0.3054	
1.5	0.3119	0.2747	0.2786	
2.0	0.3083	0.2985	0.3006	
2.5	0.3565	0.3185	0.3166	Yes
3.0	0.3154	0.2802	0.2837	

Table 10 presents the conditional standard error when the Bayesian estimation procedure was used. The rearrangement procedure was effective in reducing the SE of the Bayesian ability estimates after review at 6 of the 13 ability levels. These were at the θ levels of - 1.5, -1.0, -0.5, 0.0, 0.5, and 2.5. This decrease in the standard error tended to be quite small. However, even at the rest of the ability levels, the Bayesian SE after the rearrangement procedure was still smaller in magnitude than the SE that existed before item review.

Figure 8 provides a comparison of the ML and Bayesian conditional SE at each of the 13 ability levels after the rearrangement procedure has taken place. The overall pattern of the results shows that the ML SE tends to be larger than the Bayesian SE at most of the ability levels. This is consistent with Kim and Nicewander (1993) who concluded that the ML estimator produced the largest standard errors compared to other estimators such as the Bayesian modal estimation. The only exceptions to this pattern are at the extremes of the distribution, with the ML SE being smaller than the Bayesian SE at the θ levels of -3.0 and 3.0.

Figure 8. Condition 1 ML and Bayesian standard error after the rearrangement procedure (250 pool and 3 reviews)

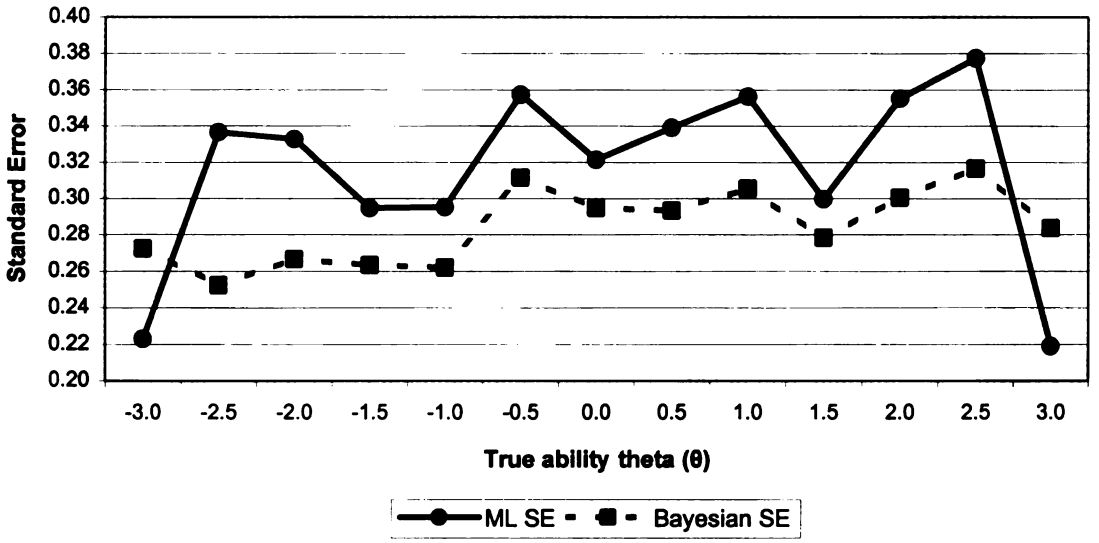
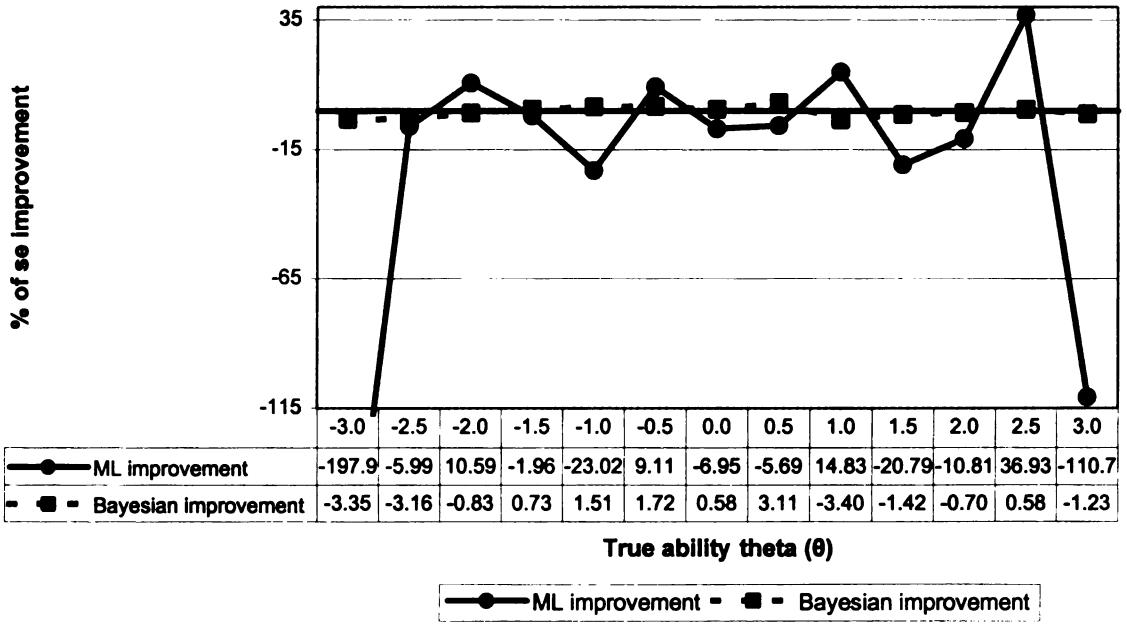


Figure 9 describes the percentage of standard error reduction that has occurred from the after review estimates to the ARP estimates. The Bayesian ARP standard errors tend to consistently show slight improvements in decreasing the se, when compared to the SE after review. In contrast, the ML estimation procedure appears to have a lot of extreme fluctuations in terms of its effect in decreasing the standard errors of the estimate. The worse fluctuation is at the θ of -3.0 where the ML SE increased by almost 200%, as well at the θ level of 3.0 where it became worse by approximately 110%. The increase in the standard errors at the extreme of the distribution is a function of the failure of the ML procedure to converge for examinees whose abilities are at the extremes of the distribution. This occurs when examinees get their answers on the test either all correct, or all wrong. So after review, and after the skipping of items in the rearrangement procedure, it is more likely that the examinees at the extremes of the distribution will get their answers either all wrong, or all correct. Consequently, the ML

estimation after the rearrangement procedure will have problems converging for these examinees, which in turn increases the ARP SE of the ability estimates.

Figure 9. Percentage of ARP standard error improvement with 3 changes and pool of 250 items (condition 1)



	-3.0	-2.5	-2.0	-1.5	-1.0	-0.5	0.0	0.5	1.0	1.5	2.0	2.5	3.0
ML improvement	-197.9	-5.99	10.59	-1.96	-23.02	9.11	-6.95	-5.69	14.83	-20.79	-10.81	36.93	-110.7
Bayesian improvement	-3.35	-3.16	-0.83	0.73	1.51	1.72	0.58	3.11	-3.40	-1.42	-0.70	0.58	-1.23

Reliability Of Test Scores (Condition 1)

The reliability of the ability estimates was also compared at the 3 time points of the rearrangement process, as shown in Table 11. The reliability of the ML estimates before review was 0.821. After the examinees changed their answers on the test, the ML reliability estimate dropped to 0.818. After the rearrangement procedure, the reliability of the scores dropped further by 0.008, to 0.810.

However, the Bayesian reliability estimates were higher than the ML reliability estimates. Before the review took place, the Bayesian reliability estimate was 0.834.

After review, the reliability jumped to 0.847. Finally, after the rearrangement procedure, the reliability increased further to 0.849.

Table 11. Reliability of ability estimates with a pool of 250 items, and 3 permitted reviews (Condition 1)

		Reliability
Maximum Likelihood	Before review	0.821
	After review	0.818
	After the rearrangement procedure	0.810
Bayesian	Before review	0.834
	After review	0.847
	After the rearrangement procedure	0.849

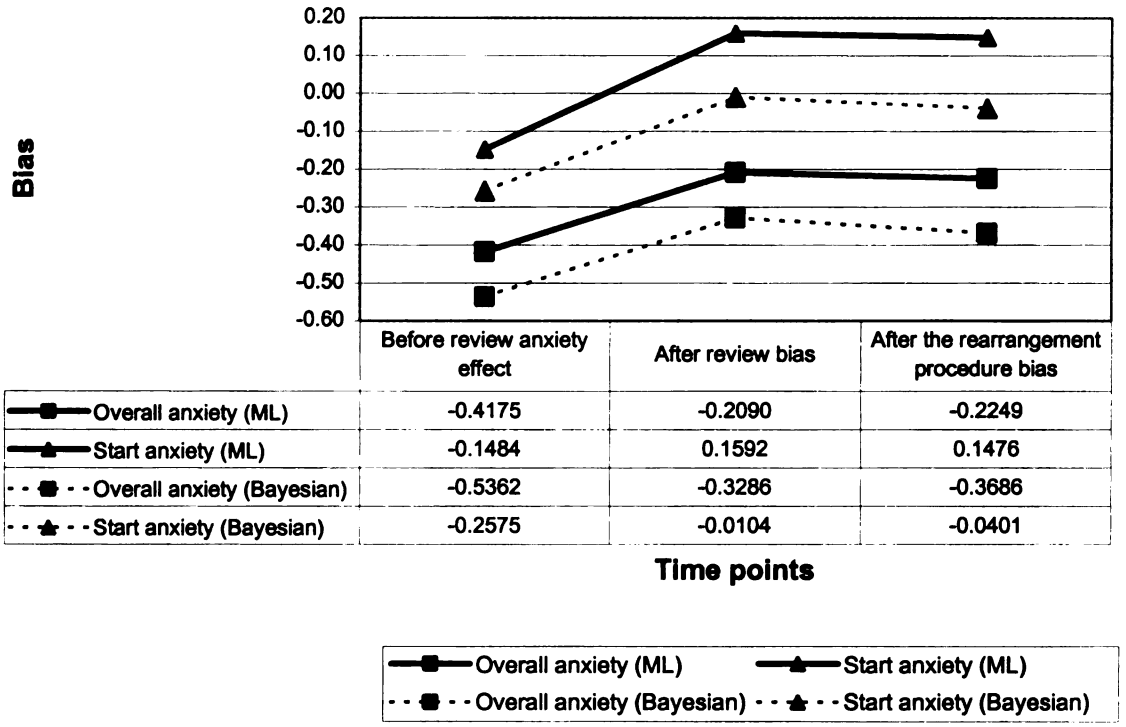
Examinee Anxiety Effects (Condition 1)

In terms of anxiety, there were two types of simulated examinees with aberrant responses. The first type of examinees, were the ones who had high anxiety at the beginning of the test, but who were able to overcome their anxiety after they started moving through the test. This anxiety was called 'start anxiety'. The second type of examinees, were the ones who had anxiety throughout the test, due to their unfamiliarity with the computerized adaptive testing format. This was called 'overall anxiety'. Both of these anxiety effects resulted in a decrease in the accuracy of the ability estimates of the examinees. Figure 10 describes how these examinees with anxiety were affected by the rearrangement procedure in terms of the bias of their score estimates.

Overall, all of the simulated examinees who had anxiety obtained more accurate ability estimates after review when compared to their before review estimates that contained the anxiety effects, that decreased the precision of their ability estimates. This was consistent with the ML and the Bayesian estimation procedures. However, the

rearrangement procedure was not very effective in reducing the bias of the ability estimates further. With the exception of the ML bias of the examinees with start anxiety, the rest of the ARP bias estimates increased when compared to the bias after review. However, even after the increase in the bias after the rearrangement procedure, that ARP estimates were still more accurate than the before review estimates that contained the anxiety effects.

Figure 10. Anxiety effects and bias of the ability estimates (condition 1)

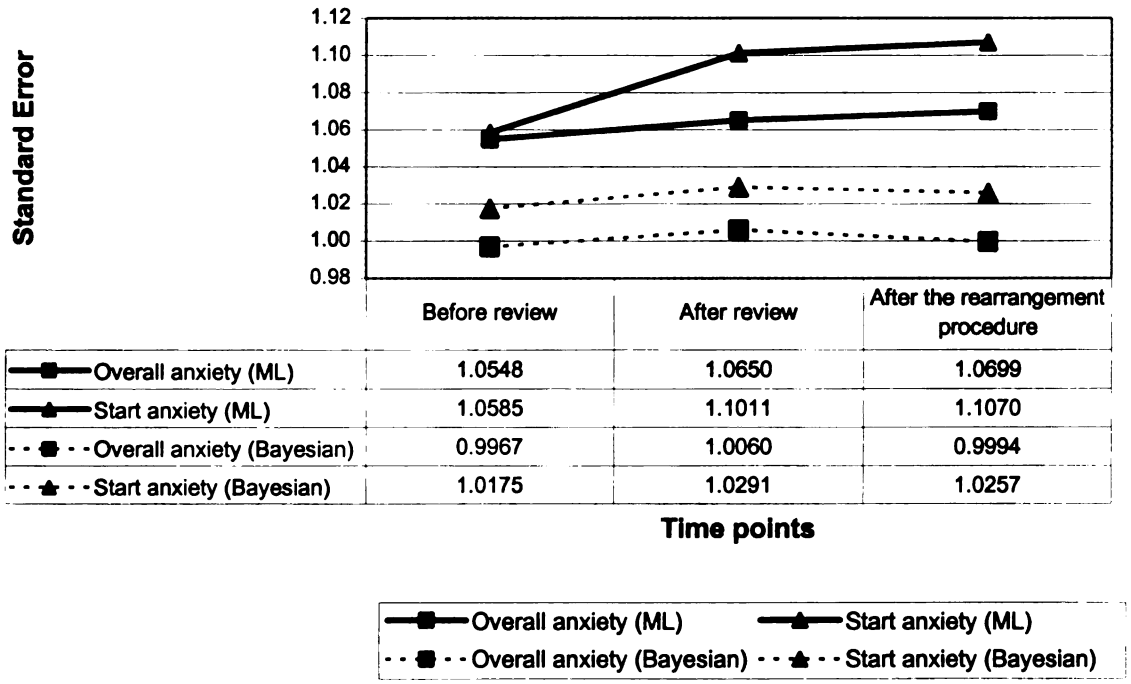


	Before review anxiety effect	After review bias	After the rearrangement procedure bias
Overall anxiety (ML)	-0.4175	-0.2090	-0.2249
Start anxiety (ML)	-0.1484	0.1592	0.1476
Overall anxiety (Bayesian)	-0.5362	-0.3286	-0.3686
Start anxiety (Bayesian)	-0.2575	-0.0104	-0.0401

When comparing the standard errors of the examinee's ability estimates after review, with the before review estimates, it is obvious that under both anxiety conditions and both estimation procedures, the standard error increased after review. After the rearrangement procedure, the SE decreased slightly when the Bayesian estimates were

used. However, these Bayesian ARP standard errors were still larger than the estimates that existed before review. These results are presented in Figure 11.

Figure 11. Standard error estimates of examinees with anxiety (condition 1)



	Before review	After review	After the rearrangement procedure
Overall anxiety (ML)	1.0548	1.0650	1.0699
Start anxiety (ML)	1.0585	1.1011	1.1070
Overall anxiety (Bayesian)	0.9967	1.0060	0.9994
Start anxiety (Bayesian)	1.0175	1.0291	1.0257

Condition 2: 250 Items With 5 Reviews Maximum

The results of condition 2 are very similar to the results of condition 1 which included the same simulated examinees, and the same item pool of 250 items. The only difference between the two conditions was that the examinees in condition 2 were allowed to make up to 5 changes (rather than 3) to their answers on the test.

Overall, 41.66% of the simulated examinees made correct-to-incorrect, or incorrect-to-correct changes to their answers. These types of changes are the only ones that will be discussed since the rearrangement procedure takes place only when such changes have occurred. Table 12 describes the percentage of actual answer changes, that are divided in the four categories that were discussed in the previous condition in the simulation.

As can be seen from Table 12, the majority of the changes that were made (41.66%) were from incorrect to correct ones. In addition, the majority of those changes were from examinees that made one or two such changes throughout their test. There were also 15.51% of examinees that made correct-to-incorrect changes to questions to which they had approximately a 0.50 probability of answering correctly.

Only 6.23% of the simulated examinees had made 'stupid mistakes' that were then changed to correct answers. Finally, there were also 3.58% of the same examinees that changed their answers to incorrect answers to an item that was originally answered correctly just by chance.

Table 12. Percentage of actual answer changing patterns in condition 2

	Number of changes	Number of examinees	Percentage of examinees (out of 26000)
Incorrect-to-correct changes (0.5 probability)	**1**	3729	14.34%
	2	3799	14.61%
	3	2428	9.34%
	4	790	3.04%
	5	85	0.33%
	Sum	**10831**	**41.66%**
Correct-to-incorrect changes (0.5 probability)	**1**	3224	12.40%
	2	711	2.73%
	3	87	0.33%
	4	10	0.04%
	5	0	0.00%
	Sum	**4032**	**15.51%**
Stupid' mistake corrections (incorrect-to-correct)	**1**	1231	4.73%
	2	331	1.27%
	3	54	0.21%
	4	5	0.02%
	5	0	0.00%
	Sum	**1621**	**6.23%**
Unlucky guess' changes (correct-to-incorrect)	**1**	722	2.78%
	2	155	0.60%
	3	44	0.17%
	4	9	0.03%
	5	0	0.00%
	Sum	**930**	**3.58%**

After the examinees reviewed their items on the test in the simulation, the rearrangement procedure was used. Because of the rearrangement procedure, there were 894 examinees (5.89%) that used item review, to which 1 of their items on the test were ignored. There were also 766 simulated examinees (5.05%) that used item review,

to which 2 of their items on the test were ignored. Finally, there were also 7853 examinees (51.77%) of the examinees that used item review, to which 3 of their items on the test were ignored.

When items were rearranged because of the rearrangement procedure, the amount of information that was provided at each ability level was used as an indicator for which item should be selected to be used next. The average amount of information that was gained by rearranging the items in condition 1 was 0.0514 with a standard deviation of 0.0421. The minimum amount of information that was gained was 0.0001, while the maximum information that was gained was 0.3396.

Results Based On Bias (Condition 2)

Table 13 presents the overall pattern of bias that exists at the three time points of before review, after review, and after the rearrangement procedure.

Table 13. Overall bias of the 250 pool with 5 reviews estimates (Condition 2)

Bias	Mean	Standard Deviation
Maximum Likelihood Bias		
Before review	-0.1374	0.3229
After review	0.0718	0.3346
After the rearrangement procedure	0.0610	0.3394
Bayesian Bias		
Before review	-0.2641	0.3099
After review	-0.0659	0.2989
After the rearrangement procedure	-0.1066	0.2955

As described in Table 13, the ML bias estimate before review was -0.1374. After review, the ML bias dropped in magnitude to 0.0718. After the rearrangement procedure,

the bias decreased in magnitude further to 0.0610. This was a 15.6% decrease in the bias when compared to the after review bias. The results that were based on the Bayesian bias showed a different pattern of bias. The Bayesian bias before review was 0.2641. After review, the Bayesian bias dropped in magnitude to -0.0659. After the rearrangement procedure, the bias increased in magnitude to -0.1066. This was a 61.7% increase in bias when compared to the bias after review.

The results of the bias are described more analytically in Table 14, which presents the bias at each of the 13 ability levels from which the examinees were sampled. In most cases the after review bias was smaller than the before review bias. However, at some ability levels such as at the θ level of 1.5, 2.0 and 3.0, the after review bias was larger than the before review bias. The reason for this increase is because in certain cases, the review process eliminated the randomness from the examinee's responses. This resulted in a mismatch between the examinee responses and the IRT model. For example, examinees with an ability of θ=2.0 might have had a 90% probability of answering item i correctly. Consequently, it is expected that 90% of the examinees with a θ of 2.0 would answer item i correctly, and 10% would answer the item incorrectly. However, if 100% of these examinees answer the item correctly, then their response patterns do not match the IRT model, which consequently will increase the after review bias of the ability estimates.

Table 14 shows that there were 9 out of the 13 ability levels where the ML bias decreased in magnitude after the rearrangement procedure. These improvements existed at the θ levels from -1.5 to 2.5. So the effects of the rearrangement procedure were generally more evident at the positive rather than the negative end of the θ scale when the ML estimation procedure is used.

T

s

wa

Ba

ab

Table 14. Conditional Maximum Likelihood bias when 5 reviews are permitted with a 250 sized item pool (Condition 2)

Ability θ	Before Review	After Review	After Rearrangement procedure (ARP)	Improvement from rearrangement procedure
-3.0	-0.1261	0.0613	-0.1105	
-2.5	-0.2234	-0.0476	-0.0722	
-2.0	-0.1765	-0.1073	-0.1151	
-1.5	-0.1284	0.0694	0.0559	Yes
-1.0	-0.1469	0.0437	0.0341	Yes
-0.5	-0.1595	0.0438	0.0333	Yes
0.0	-0.1398	0.1167	0.1041	Yes
0.5	-0.1422	0.0770	0.0648	Yes
1.0	-0.1034	0.0242	0.0204	Yes
1.5	-0.0883	0.1940	0.1765	Yes
2.0	-0.1442	0.2080	0.1985	Yes
2.5	-0.1087	-0.0198	-0.0065	Yes
3.0	-0.0501	-0.0520	0.1053	

Table 15 describes the conditional bias when the Bayesian estimation procedure was used. There were only two ability levels that showed an improvement in the Bayesian bias. Those were the θ levels of -3.0 and -2.5. However, at the rest of the ability levels, the ARP bias was still smaller in magnitude than the before review bias.

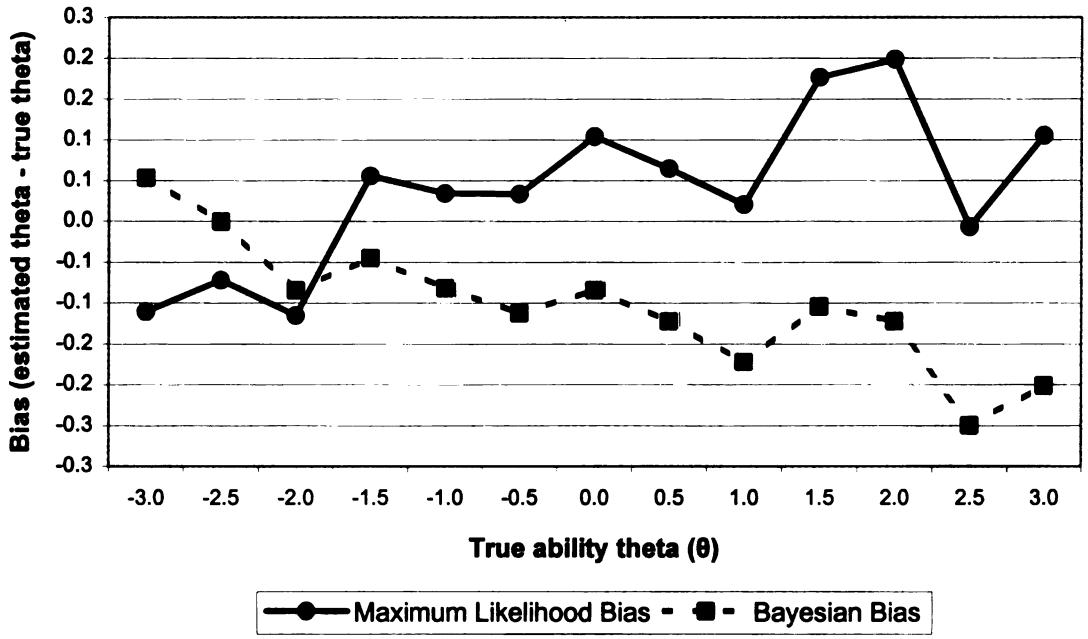
Table 15. Conditional Bayesian bias when 5 reviews are permitted with a 250 sized item pool (Condition 2)

Ability θ	Before Review	After Review	After Rearrangement procedure (ARP)	Improvement from rearrangement Procedure
-3.0	-0.0469	0.0845	0.0534	Yes
-2.5	-0.1237	0.0272	-0.0002	Yes
-2.0	-0.1808	-0.0712	-0.0846	
-1.5	-0.1979	-0.0150	-0.0451	
-1.0	-0.2309	-0.0433	-0.0820	
-0.5	-0.2621	-0.0782	-0.1123	
0.0	-0.2680	-0.0322	-0.0842	
0.5	-0.2839	-0.0660	-0.1224	
1.0	-0.2874	-0.1638	-0.1720	
1.5	-0.2980	-0.0435	-0.1042	
2.0	-0.3577	-0.0539	-0.1223	
2.5	-0.3570	-0.2211	-0.2498	
3.0	-0.3656	-0.1688	-0.2012	

A comparison of the results from Tables 14 and 15 is shown in Figure 12. This figure describes the bias that existed in the final ARP estimates, at each of the 13 θ levels. When the ML estimate was used, the bias of the test scores increased from a negative bias to a positive bias as the examinee's ability estimates increased in the simulation. So the examinees who were at the lower end of the distribution obtained lower ability estimates than their true abilities after review and after the rearrangement procedure. However, the examinees at the higher ends of the θ scale obtained higher

ML ability estimates than their true scores when the rearrangement procedure was used. This is because the ML procedure is more biased towards the extremes of the θ scale (Kim & Nicewander, 1993; Lord, 1986; Wang & Vispoel, 1998). However, the pattern of bias is a function of the ML estimation procedure, rather than a function of the rearrangement procedure.

Figure 12. Condition 2 ML and Bayesian Bias after the rearrangement procedure (250 pool and 5 reviews)

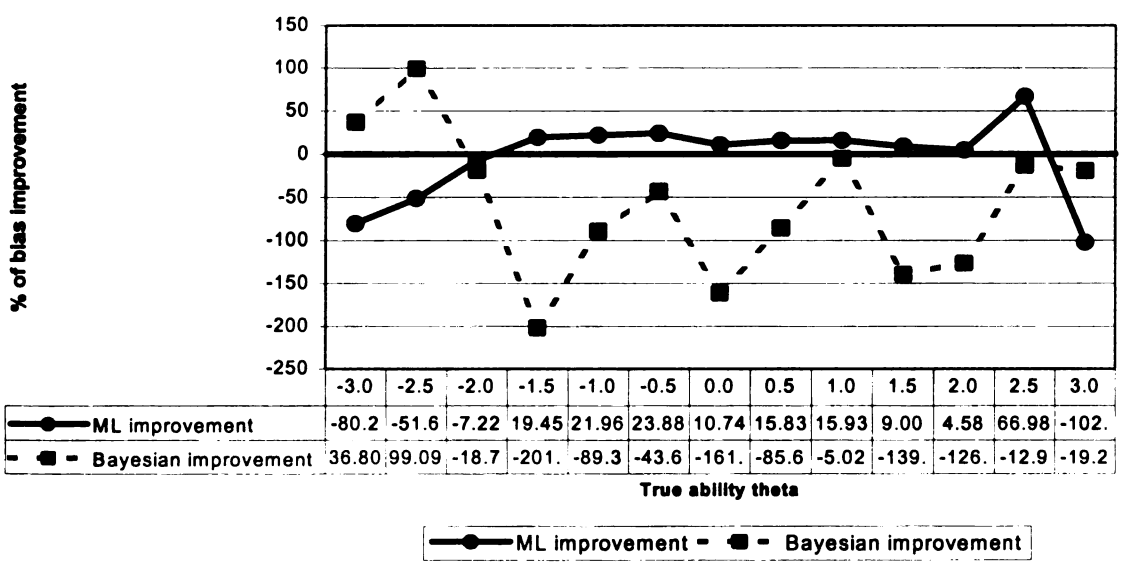


With reference to the simulated examinees at the extreme ends of the score scale, the results of condition 2 show the opposite pattern when the Bayesian rather than the ML estimates were used. For example, examinees at the lower end of the scale have higher Bayesian estimated test scores rather than true test scores after the rearrangement procedure. In contrast, the examinees at the higher end of the scale obtained lower ability Bayesian estimates than their true scores when the rearrangement

procedure was used. This is consistent with prior research that has shown similar patterns of results when the Bayesian estimation procedures are used (Kim & Nicewander, 1993; Lord, 1986; Wang & Vispoel, 1998). However, the bias towards the mean is a function of the Bayesian estimation procedure, rather than a function of the rearrangement procedure.

Figure 13 describes the percentage of bias reduction that has occurred from the after review estimates to the ARP estimates. The ML estimation procedure appears to work well at most ability levels since it has a small positive percentage of improvement at 9 of the 13 true ability levels. This means that the ML bias decreased in magnitude due to the rearrangement procedure. However, the Bayesian bias does not appear to work as well in reducing the ARP bias. The Bayesian estimation is especially problematic at some θ levels such as θ=-1.5 where the bias increases as much as 201.8%.

Figure 13. Percentage of ARP bias improvement with 5 changes and pool of 250 items (condition 2)



Results Based on the Standard Error of the θ Estimate (Condition 2)

Table 16 describes the average standard deviation of both estimation procedures at each of the three time points of the rearrangement procedure. These results are averaged over all of the simulated examinees in the sample, including the ones with test anxiety. According to Table 16, the ML SD before review was 1.0728. After review, the ML SD increased to 1.0896. After the rearrangement procedure, the standard deviation increased further to 1.0945. This was a 0.4% increase in the SD when compared to the SD after review.

Table 16. Overall standard deviation of the θ estimates obtained from the pool of 250 items, when 5 reviews were permitted (Condition 2)

	Standard Deviation of the θ estimates
Maximum Likelihood	
Before review	1.0728
After review	1.0896
After the rearrangement procedure	1.0945
Bayesian	
Before review	1.0187
After review	1.0285
After the rearrangement procedure	1.0232

The results that were based on the Bayesian estimation showed a different pattern of standard deviations. The Bayesian SD before review was 1.0187. After review, the Bayesian SD increased to 1.0285. However, after the rearrangement

procedure, the SD decreased to 1.0232. So the ARP Bayesian SD improved slightly by 0.5% when compared to the SD that existed after review.

Table 17. Conditional Maximum Likelihood standard error when 5 reviews are permitted with a 250 sized item pool (Condition 2)

Ability θ	Before Review	After Review	After Rearrangement procedure (APR)	Improvement from rearrangement procedure
-3.0	0.2331	0.1262	0.2349	
-2.5	0.3313	0.3156	0.3373	
-2.0	0.3570	0.3289	0.3316	
-1.5	0.3204	0.2922	0.2986	
-1.0	0.3014	0.2928	0.2985	
-0.5	0.3334	0.3557	0.3596	
0.0	0.3109	0.3243	0.3282	
0.5	0.3165	0.3403	0.3425	
1.0	0.3562	0.3519	0.3564	
1.5	0.3026	0.3010	0.3058	
2.0	0.2966	0.3461	0.3617	
2.5	0.3702	0.3503	0.3797	
3.0	0.2836	0.1106	0.2229	

The standard error results are described more analytically in Table 17, which presents the ML SE at each of the 13 ability levels from which the examinees were sampled. Table 17 shows that at no ability level did the standard error of the ARP ML decrease when compared to the after review se. In most cases, the final ARP SE was

even larger than SE that existed before review. This was expected since the standard error tends to increase when the length of a test is shortened, which is the case with the rearrangement procedure.

Table 18. Conditional Bayesian standard error when 5 reviews are permitted with a 250 sized item pool (Condition 2)

Ability θ	Before Review	After Review	After Rearrangement procedure (APR)	Improvement from rearrangement procedure
-3.0	0.3086	0.2748	0.2788	
-2.5	0.2766	0.2470	0.2552	
-2.0	0.3049	0.2645	0.2672	
-1.5	0.3024	0.2667	0.2661	Yes
-1.0	0.2984	0.2675	0.2641	Yes
-0.5	0.3124	0.3185	0.3124	Yes
0.0	0.3037	0.3017	0.2991	Yes
0.5	0.3012	0.3043	0.2943	Yes
1.0	0.3232	0.2955	0.3055	
1.5	0.3119	0.2794	0.2815	
2.0	0.3083	0.3034	0.3049	
2.5	0.3565	0.3199	0.3175	Yes
3.0	0.3154	0.2731	0.2752	

Table 18 presents the conditional standard error when the Bayesian estimation procedure was used. The rearrangement procedure was effective in reducing the SE of the Bayesian ability estimates after review at 6 of the 13 ability levels. These were at the θ levels of - 1.5, -1.0, -0.5, 0.0, 0.5, and 2.5. This decrease in the standard error tended

to be quite small. However, even at the rest of the ability levels, the SE after the rearrangement procedure was still smaller in magnitude than the SE that existed before item review. This is in contrast to the conditional ML SE that was larger in magnitude after the rearrangement procedure, than the standard error estimates before review.

Figure 14 provides a comparison of the ML and Bayesian SE at each of the 13 ability levels after the rearrangement procedure has taken place. The overall pattern of the results shows that the ML SE tends to be larger than the Bayesian SE at most of the ability levels. This is consistent with Kim and Nicewander (1993) who concluded that the ML estimator produced the largest standard errors compared to other estimators such as the Bayesian modal estimation. The only exceptions to this pattern are at the extremes of the distribution, with the ML SE being smaller than the Bayesian SE at the θ levels of -3.0 and 3.0.

Figure 14. Condition 2 ML and Bayesian standard error after the rearrangement procedure (250 pool and 5 reviews)

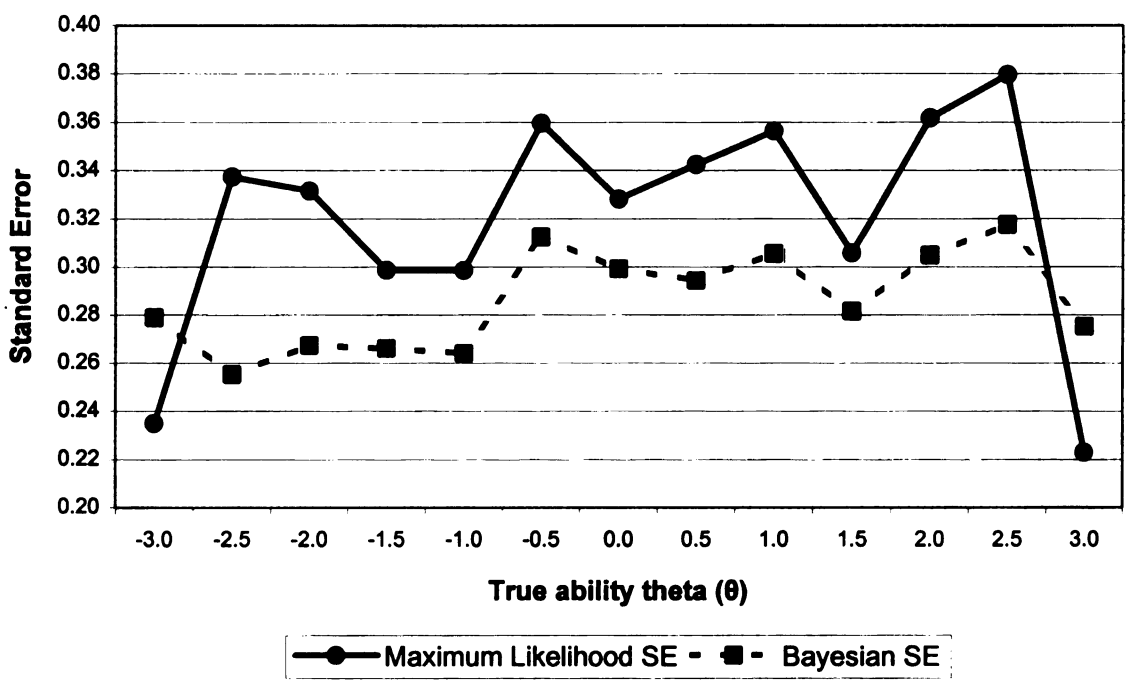
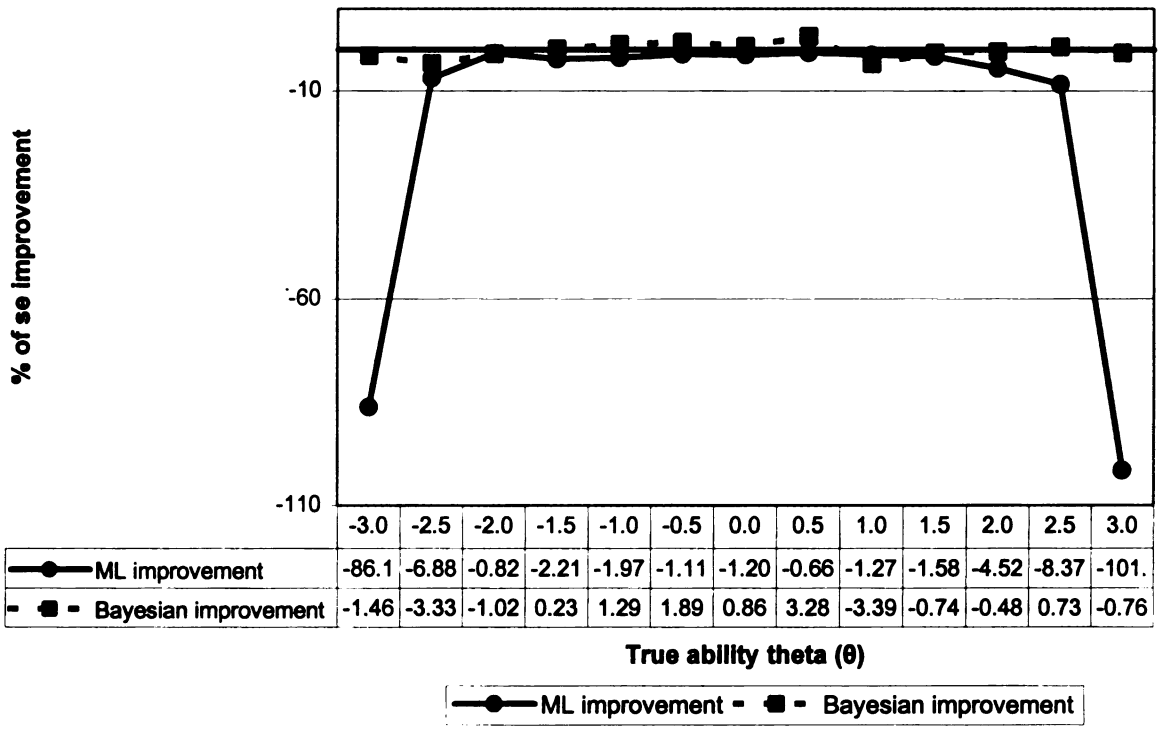


Figure 15 describes the percentage of standard error reduction that has occurred from the after review estimates to the ARP estimates. At a large portion of the distribution, there did not appear to be major differences in the improvement of the ML and Bayesian standard error, although the improvement in the Bayesian standard error appeared to be slightly better than that of the ML SE. However, the ML estimation procedure increased the standard error greatly at the extremes of the θ distribution. At the θ=-3.0 level, the standard error increased by 86.2%, while at θ=3.0 the standard error increased by 101.5%. The increase in the standard errors at the extreme of the distribution is a function of the failure of the ML procedure to converge for examinees whose abilities are at the extremes of the distribution. This occurs when examinees get their answers on the test either all correct, or all wrong. So after review, and after the skipping of items in the rearrangement procedure, it is more likely that the examinees at the extremes of the distribution will get their answers either all wrong, or all correct. Consequently, the ML estimation after the rearrangement procedure will have problems converging for these examinees, which in turn increases the ARP SE of the ability estimates.

Figure 15. Percentage of ARP standard error improvement with 5 changes and pool of 250 items (condition 2)



Reliability Of Test Scores (Condition 2)

The reliability of the ability estimates was also compared when a maximum of 5 reviews were permitted by the examinees. As shown in Table 19, the reliability of the ML estimates before review was 0.817. After the examinees changed their answers on the test, the ML reliability estimate dropped to 0.811. After the rearrangement procedure, the reliability dropped further to 0.806. However, this drop in reliability was too small to have a significantly negative effect on the quality of the examinee's final ability estimates.

When the Bayesian estimation procedure was used, the reliability increased from 0.830 from before review to 0.841 after review. The reliability then increased even further to 0.843 after the rearrangement procedure took place.

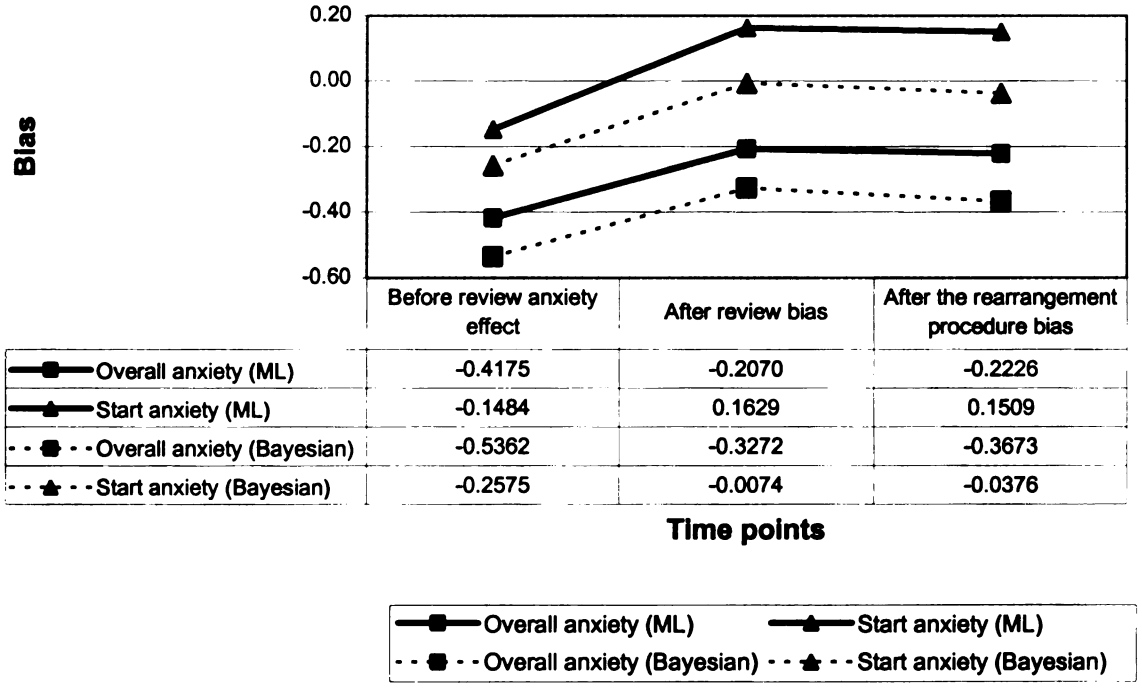
Table 19. Reliability of ability estimates with a pool of 250 items, and 5 permitted reviews (Condition 2)

		Reliability
Maximum Likelihood	Before review	0.817
	After review	0.811
	After the rearrangement procedure	0.806
Bayesian	Before review	0.830
	After review	0.841
	After the rearrangement procedure	0.843

Examinee Anxiety Effects (Condition 2)

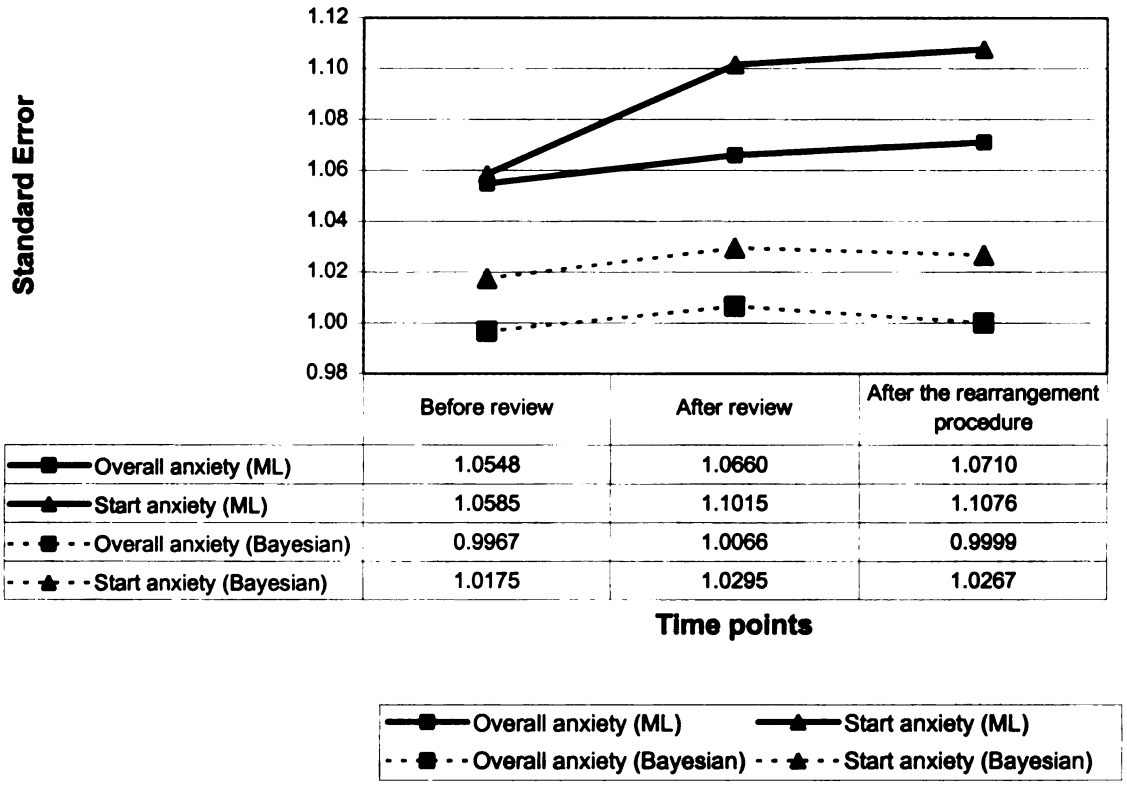
Figure 16 describes how the simulated examinees with anxiety were affected by the rearrangement procedure. Overall, all of the examinees who had anxiety obtained more accurate ability estimates after review when compared to their before review anxiety effects that decreased the precision of their ability estimates. After the rearrangement procedure, the bias tended to increase. However, even after the increase in the bias after the rearrangement procedure, the ARP ability estimates were still more accurate than the before review estimates that contained the anxiety effect. The only exception was for the examinees who had start anxiety, and whose scores were calculated with the ML estimation procedure. The average of that bias showed an increase in the bias after review, and a decrease in the bias after the rearrangement procedure.

Figure 16. Anxiety effects and bias of the ability estimates (condition 2)



When comparing the standard errors of the simulated examinees' ability estimates after review, with the before review estimates, it is obvious that under both anxiety conditions and both estimation procedures, the standard errors increased after review. After the rearrangement procedure, the SE decreased slightly when the Bayesian estimates were used. However, the ARP Bayesian standard errors were still larger than the estimates that existed before review. These results are presented in Figure 17.

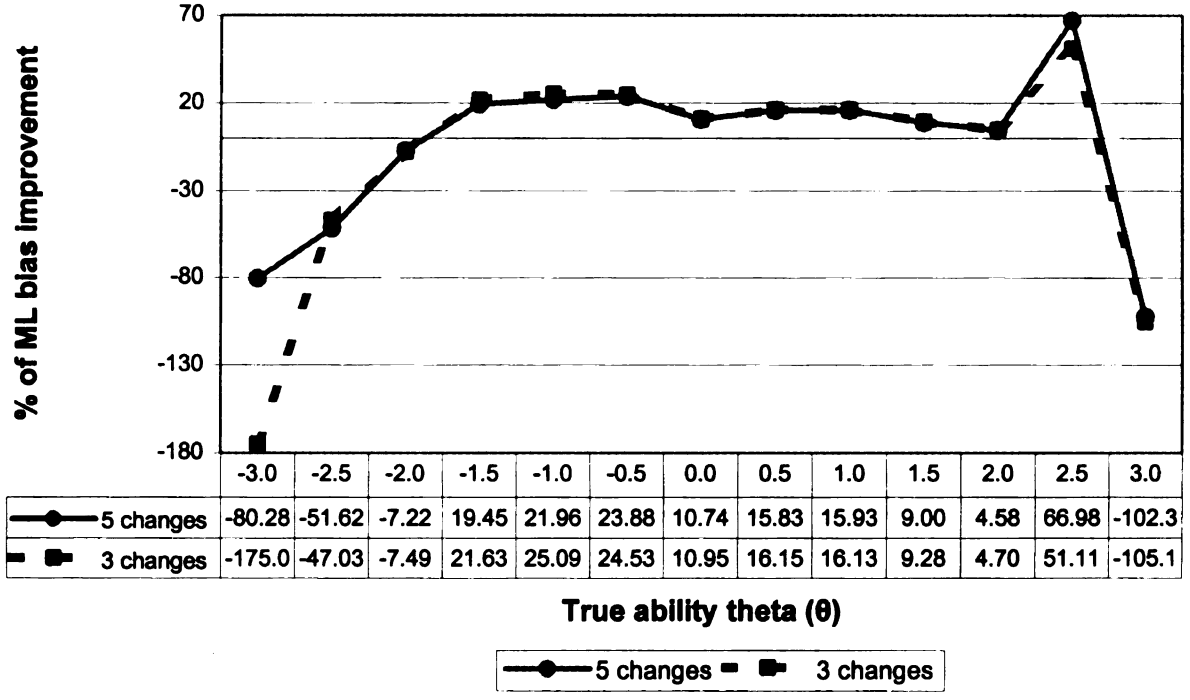
Figure 17. Standard error estimates of examinees with anxiety (condition 2)



Comparison Of Three And Five Changes With A 250 Size Item Pool

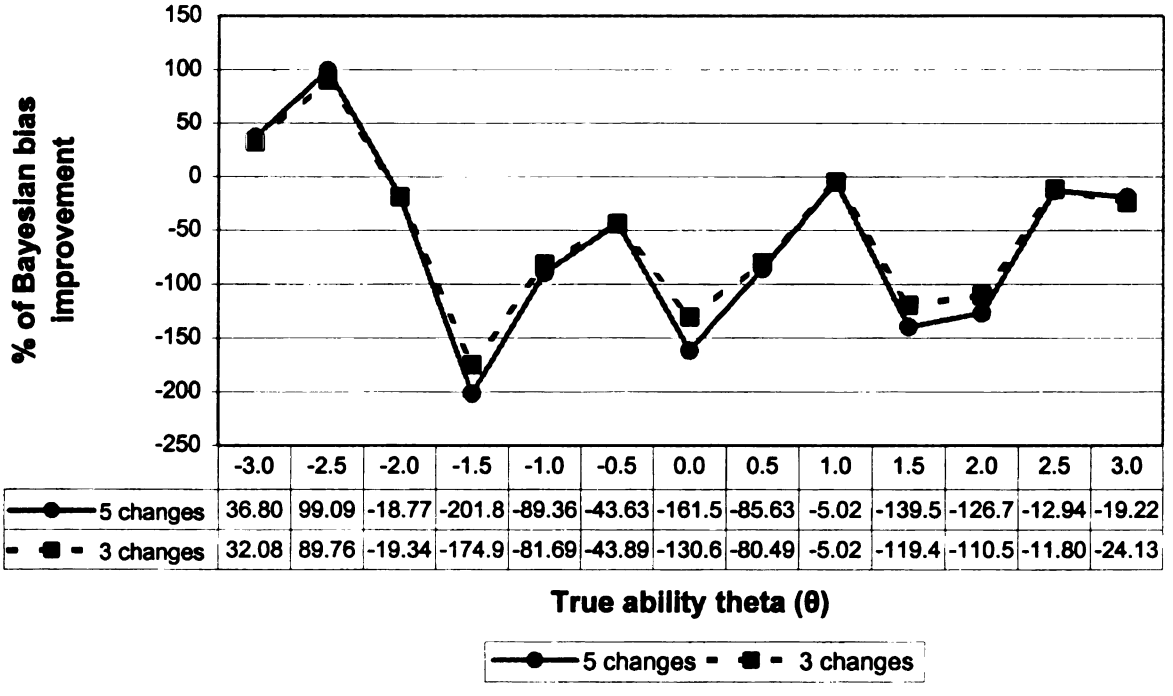
Figures 18 and 19 describe the differences that exist in the amount of item changes on the ARP ML and Bayesian bias. At most θ points, the differences are indistinguishable. The largest difference exists in the ML bias where the bias improvement deteriorates by 175.0% when 3 changes are made, in contrast to 80.3% when 5 changes are made. However, the differences in bias when 3 or 5 changes are made, are more pronounced when the Bayesian estimation procedure is used. With the Bayesian procedure, the improvement in the bias tends to be slightly worse when 5 changes are allowed to be made by the examinees.

Figure 18. The effects of the numbers of items reviewed on the ML bias improvement obtained from a pool of 250 items



	-3.0	-2.5	-2.0	-1.5	-1.0	-0.5	0.0	0.5	1.0	1.5	2.0	2.5	3.0
5 changes	-80.28	-51.62	-7.22	19.45	21.96	23.88	10.74	15.83	15.93	9.00	4.58	66.98	-102.3
3 changes	-175.0	-47.03	-7.49	21.63	25.09	24.53	10.95	16.15	16.13	9.28	4.70	51.11	-105.1

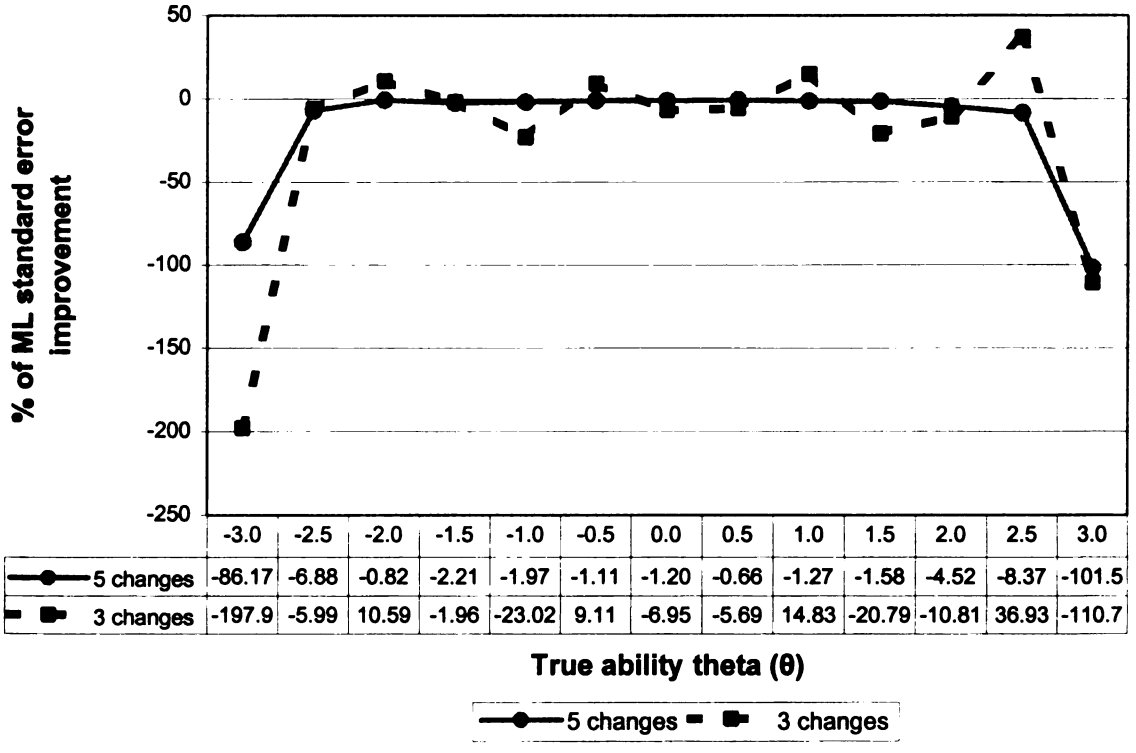
Figure 19. The effects of the numbers of items reviewed on the Bayesian bias improvement obtained from a pool of 250 items



	-3.0	-2.5	-2.0	-1.5	-1.0	-0.5	0.0	0.5	1.0	1.5	2.0	2.5	3.0
5 changes	36.80	99.09	-18.77	-201.8	-89.36	-43.63	-161.5	-85.63	-5.02	-139.5	-126.7	-12.94	-19.22
3 changes	32.08	89.76	-19.34	-174.9	-81.69	-43.89	-130.6	-80.49	-5.02	-119.4	-110.5	-11.80	-24.13

Figures 20 and 21 describe the differences that exist in the improvement of the standard error that occurs after the rearrangement procedure is used. Figure 20 describes the ML SE improvement differences that occur when 3 or 5 changes are made by the examinees to their answers. Figure 21 describes the Bayesian SE improvement differences that occur when 3 or 5 changes are made by the examinees to their answers. Both graphs show that there are no consistent reasons to prefer the allowance of making a maximum 3 or 5 changes to the answers on a test. However, when the Bayesian estimation procedure is used, the increases or decreases in standard errors tend to be smaller in magnitude than when the ML estimation procedure is used.

Figure 20.The effects of the numbers of items reviewed on the ML standard error improvement obtained from a pool of 250 items



	-3.0	-2.5	-2.0	-1.5	-1.0	-0.5	0.0	0.5	1.0	1.5	2.0	2.5	3.0
5 changes	-86.17	-6.88	-0.82	-2.21	-1.97	-1.11	-1.20	-0.66	-1.27	-1.58	-4.52	-8.37	-101.5
3 changes	-197.9	-5.99	10.59	-1.96	-23.02	9.11	-6.95	-5.69	14.83	-20.79	-10.81	36.93	-110.7

Figure 21. The effects of the numbers of items reviewed on the Bayesian standard error improvement obtained from a pool of 250 items

	-3.0	-2.5	-2.0	-1.5	-1.0	-0.5	0.0	0.5	1.0	1.5	2.0	2.5	3.0
5 changes	-1.46	-3.33	-1.02	0.23	1.29	1.89	0.86	3.28	-3.39	-0.74	-0.48	0.73	-0.76
3 changes	-3.35	-3.16	-0.83	0.73	1.51	1.72	0.58	3.11	-3.40	-1.42	-0.70	0.58	-1.23

Condition 3: 500 Items With 3 Reviews Maximum

Condition 3 examined the effects that the rearrangement procedure had on the accuracy of the ability estimates, when an item pool of 500 items was used for the formation of the adaptive tests. This was done to determine if the effects of the rearrangement procedure would be stronger when larger item pools are used. Condition 3 includes the restriction that only up to 3 items are permitted to be reviewed by each examinee.

Overall, 41.66% of the examinees made correct-to-incorrect, or incorrect-to-correct changes to their answers. These types of changes are the only ones that will be discussed since the rearrangement procedure takes place only when such changes have occurred. Table 19 describes the percentage of actual answer changes, that are divided in the four categories that were discussed in the previous condition in the simulation.

As can be seen from Table 20, the majority of the changes that were made (40.37%) were from incorrect to correct ones. In addition, the majority of those changes were from examinees that made one or two such changes throughout their test. There were also 14.94% of examinees that made correct-to-incorrect changes to questions to which they had approximately a 0.50 probability of answering correctly.

Only 5.64% of the examinees had made 'stupid mistakes' that were then changed to correct answers. Finally, there were also 3.03% of the examinees that changed their answers to incorrect answers to an item that was originally answered correctly just by chance.

Table 20. Percentage of actual answer changing patterns in condition 3

	Number of changes	Number of examinees	Percentage of examinees (out of 26000)
Incorrect-to-correct changes (0.5 probability)	**1**	3803	14.63%
	2	3776	14.52%
	3	2204	11.22%
	4	0	0.00%
	5	0	0.00%
	Sum	**10496**	**40.37%**
Correct-to-incorrect changes (0.5 probability)	**1**	3135	12.06%
	2	664	2.55%
	3	86	0.33%
	4	0	0.00%
	5	0	0.00%
	Sum	**3885**	**14.94%**
Stupid' mistake corrections (incorrect-to-correct)	**1**	1156	4.45%
	2	310	1.19%
	3	0	0.00%
	4	0	0.00%
	5	0	0.00%
	Sum	**1466**	**5.64%**
Unlucky guess' changes (correct-to-incorrect)	**1**	644	2.48%
	2	143	0.55%
	3	0	0.00%
	4	0	0.00%
	5	0	0.00%
	Sum	**787**	**3.03%**

After the simulated examinees reviewed their items on the test, the rearrangement procedure was used. Because of the rearrangement procedure, there were 928 examinees (5.99%) that used item review, to which 1 of their items on the test were ignored. There were also 783 examinees (5.05%) that used item review, to which

2 of their items on the test were ignored. Finally, there were also 8048 examinees (51.91%) of the examinees that used item review, to which 3 of their items on the test were ignored.

When items were rearranged because of the rearrangement procedure, the amount of information that was provided at each ability level was used as an indicator for which item should be selected to be used next. The average amount of information that was gained by rearranging the items in condition 1 was 0.0513 with a standard deviation of 0.0426. The minimum amount of information that was gained was 0.00008, while the maximum information that was gained was 0.3396.

Results Based On Bias (Condition 3)

The overall pattern of results from condition 3 are consistent with the results of the previous 2 conditions. Table 21 presents these results in detail.

Table 21. Overall bias of the 500 pool with 3 reviews estimates (Condition 3)

Bias	Mean	Standard Deviation
Maximum Likelihood Bias		
Before review	-0.1374	0.3229
After review	0.0668	0.3301
After the rearrangement procedure	0.0562	0.3348
Bayesian Bias		
Before review	-0.2641	0.3099
After review	-0.0693	0.2956
After the rearrangement procedure	-0.1090	0.2929

As described in Table 21, the ML bias estimate before review was -0.1374. After review, the ML bias dropped in magnitude to 0.0668. After the rearrangement procedure, the bias decreased further in magnitude to 0.0562. This was a 15.8% decrease in the bias when compared to the bias after review. The results that were based on the Bayesian bias showed a different pattern. The Bayesian bias estimate before review was -0.2641. After review, the Bayesian bias dropped in magnitude to -0.0693. After the rearrangement procedure, the Bayesian bias then increased in magnitude by 57.3% to -0.1090. However, this bias was still smaller than the before review bias.

By comparing the overall results in terms of bias, the ML estimates tend to be more accurate. The lowest bias exists with the ML estimate when it is estimated after the rearrangement procedure. This bias of 0.0562 is even lower than the smallest Bayesian bias (-0.0693) that is produced only after item review has taken place.

Table 22 describes the conditional bias at the three points of the rearrangement procedure. At most ability levels, the after review bias was smaller than the before review bias. However, in some cases, such as at the θ level of 1.5 and 2.0, the after review bias was larger than the before review bias. The reason for this increase is because in certain cases, the review process eliminated the randomness from the examinee's responses. This resulted in a mismatch between the examinee responses and the IRT model. For example, examinees with an ability of θ=2.0 might have had a 90% probability of answering item i correctly. Consequently, it is expected that 90% of the examinees with a θ of 2.0 would answer item i correctly, and 10% would answer the item incorrectly. However, if 100% of these examinees answer the item correctly, then their response patterns do not match the IRT model, which consequently will increase the after review bias of the ability estimates.

Table 22. Conditional Maximum Likelihood bias when 3 reviews are permitted with a 500 sized item pool (Condition 3)

Ability θ	Before Review	After Review	After Rearrangement procedure (ARP)	Improvement from rearrangement procedure
-3.0	-0.1261	0.0503	-0.1268	
-2.5	-0.2234	-0.0284	-0.0521	
-2.0	-0.1765	-0.1054	-0.1133	
-1.5	-0.1284	0.0639	0.0497	Yes
-1.0	-0.1469	0.0368	0.0266	Yes
-0.5	-0.1595	0.0429	0.0326	Yes
0.0	-0.1398	0.1075	0.0959	Yes
0.5	-0.1422	0.0703	0.0586	Yes
1.0	-0.1034	0.0241	0.0202	Yes
1.5	-0.0883	0.1857	0.1688	Yes
2.0	-0.1442	0.1970	0.1856	Yes
2.5	-0.1087	-0.0199	-0.0068	Yes
3.0	-0.0501	-0.0527	0.1060	

Table 22 also shows that there were 9 out of the 13 ability levels where the ML bias decreased in magnitude after the rearrangement procedure was used. These improvements existed at the θ levels from -1.5 to 2.5. So the effects of the rearrangement procedure were generally more evident at the positive rather than the negative end of the θ scale when the ML estimation procedure was used.

Table 23. Conditional Bayesian bias when 3 reviews are permitted with a 500 sized item pool (Condition 3)

Ability θ	Before Review	After Review	After Rearrangement procedure (ARP)	Improvement from rearrangement procedure
-3.0	-0.0469	0.0764	0.0490	Yes
-2.5	-0.1237	0.0376	0.0097	Yes
-2.0	-0.1808	-0.0697	-0.0834	
-1.5	-0.1979	-0.0187	-0.0495	
-1.0	-0.2309	-0.0477	-0.0855	
-0.5	-0.2621	-0.0790	-0.1134	
0.0	-0.2680	-0.0379	-0.0875	
0.5	-0.2839	-0.0704	-0.1248	
1.0	-0.2874	-0.1640	-0.1722	
1.5	-0.2980	-0.0503	-0.1090	
2.0	-0.3577	-0.0632	-0.1307	
2.5	-0.3570	-0.2224	-0.2506	
3.0	-0.3656	-0.1624	-0.2003	

Table 23 presents the bias when the conditional Bayesian estimation procedure was used. The rearrangement procedure was not very effective in reducing the bias of the Bayesian ability estimates after review at most of the ability levels. The only exceptions were the θ levels of - 3.0 and -2.5 where the bias decreased in magnitude with the rearrangement procedure. However, even the at the rest of the ability levels, the bias after the rearrangement procedure was still smaller in magnitude than the bias that existed before item review.

A comparison of the results from Tables 22 and 23 is presented in Figure 22. This figure describes the bias that existed in the final ARP estimates, at each of the 13 θ levels. When the ML estimate was used, the bias of the test scores increased from a negative bias to a positive bias as the ability estimates increased. So the examinees who were at the lower end of the distribution obtained lower ability estimates than their true abilities after the rearrangement procedure. However, the examinees at the higher ends of the θ scale obtained higher ML ability estimates than their true scores when the rearrangement procedure was used. This is because the ML procedure is more biased towards the extremes of the θ scale (Kim & Nicewander, 1993; Lord, 1986; Wang & Vispoel, 1998). This pattern of bias is a function of the ML estimation procedure, rather than a function of the rearrangement procedure.

With reference to the simulated examinees at the extreme ends of the score scale, the results of condition 3 showed the opposite pattern when the Bayesian estimates were used. For example, examinees at the lower end of the scale had higher Bayesian estimated test scores rather than true test scores after the rearrangement procedure. In contrast, the examinees at the higher end of the scale obtained lower Bayesian ability estimates than their true scores when the rearrangement procedure was used. This is because Bayesian estimates are more biased towards the mean of the θ scale (Kim & Nicewander, 1993; Lord, 1986; Wang & Vispoel, 1998). However, the bias towards the mean is a function of the Bayesian estimation procedure, rather than a function of the rearrangement procedure.

Figure 22. Condition 3 ML and Bayesian Bias after the rearrangement procedure (500 pool and 3 reviews)

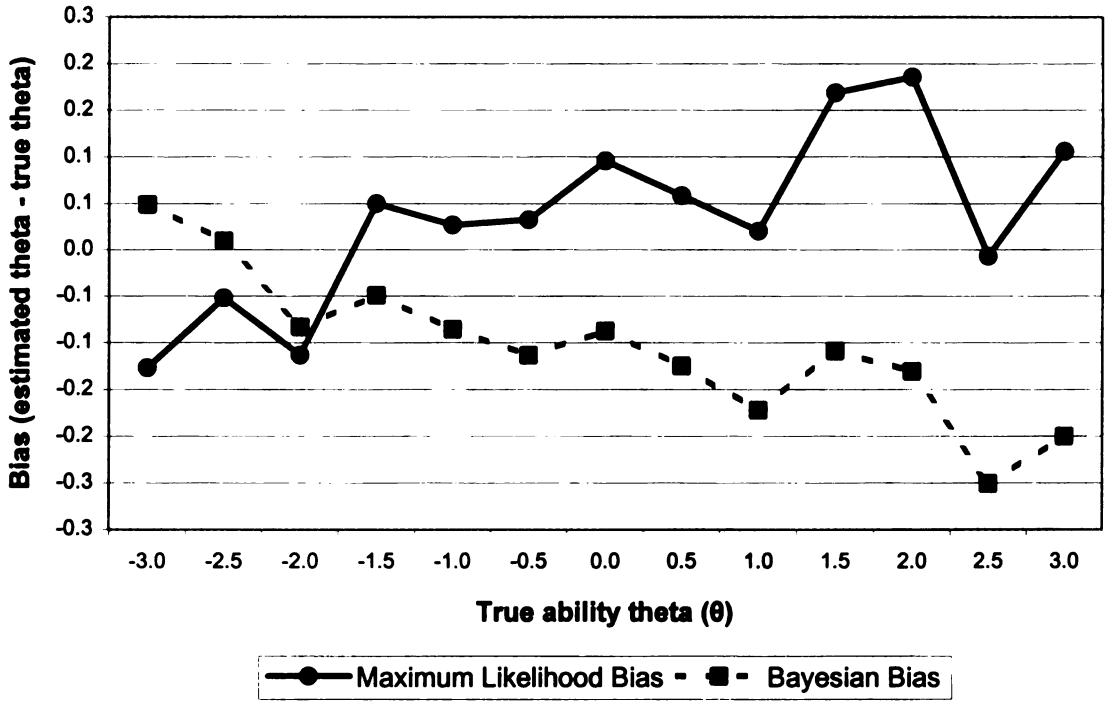
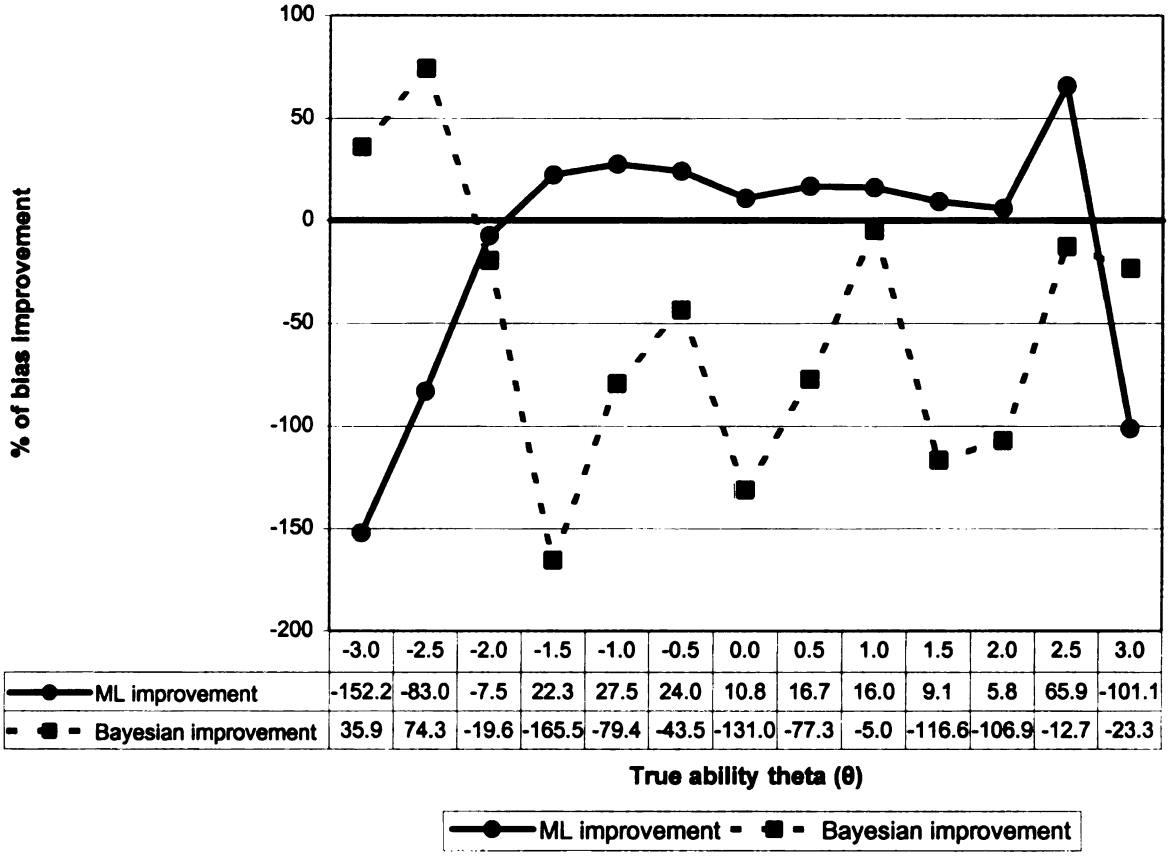


Figure 23 describes the percentage of bias reduction that has occurred from the after review estimates to the ARP estimates. The ML estimation procedure appears to work well at most ability levels since it has a positive percentage of bias improvement 9 of the 13 ability levels. This means that the ML bias decreased due to the rearrangement procedure. The ML estimation procedure was problematic, though for the examinees whose abilities were around -3.0 θ and 3.0 θ. The bias produced by the rearrangement procedure for the examinees at the -3.0 θ was actually 152.2% worse than the bias that existed after review. The ML bias became worse by 101.1% for the examinees whose true θ was 3.0. However, the Bayesian bias tended to increase at all but 2 θ levels. These were the levels of -3.0 and -2.5. So in terms of bias, the ML estimates tend to be

more accurate than the Bayesian estimates when the rearrangement procedure is used in condition 3.

Figure 23. Percentage of ARP bias improvement with 3 changes and pool of 500 items (condition 3)



	-3.0	-2.5	-2.0	-1.5	-1.0	-0.5	0.0	0.5	1.0	1.5	2.0	2.5	3.0
ML improvement	-152.2	-83.0	-7.5	22.3	27.5	24.0	10.8	16.7	16.0	9.1	5.8	65.9	-101.1
Bayesian improvement	35.9	74.3	-19.6	-165.5	-79.4	-43.5	-131.0	-77.3	-5.0	-116.6	-106.9	-12.7	-23.3

Results Based on the Standard Error of the θ Estimate (Condition 3)

Table 24 describes the average standard deviation of both estimation procedures at each of the three time points of the rearrangement procedure. These results are averaged over all of the examinees in the sample, including the ones with test anxiety. According to Table 24, the ML SD before review was 1.0728. After review, the ML SD increased to 1.0873. After the rearrangement procedure, the SD increased even further to 1.0923. So the ARP ML SD increased slightly by 0.4% when compared to the SD that existed after review.

Table 24. Overall standard deviation of the θ estimates obtained from the pool of 500 items, when 3 reviews were permitted (Condition 3)

	Standard Deviation of θ estimates
Maximum Likelihood	
Before review	1.0728
After review	1.0873
After the rearrangement procedure	1.0923
Bayesian	
Before review	1.0187
After review	1.0267
After the rearrangement procedure	1.0220

The results that were based on the Bayesian estimation showed a different pattern of standard deviation. The Bayesian SE before review was 1.0187. After review, the Bayesian SD increased to 1.0267. After the rearrangement procedure, the standard deviation dropped to 1.0220. This was a 3.7% increase in the SD when compared to the SE after review.

The results from Table 24 are described more analytically in Table 25, which presents the conditional ML SE at each of the 13 ability levels from which the examinees were sampled. Table 25 shows that at no ability level did the standard error of the ARP ML decrease when compared to the after review se. In most cases, the final ARP SE was even larger than SE that existed before review. This is expected since standard error tends to decrease when the length of a test is shortened, which is the case with the rearrangement procedure.

Table 25. Conditional Maximum Likelihood standard error when 3 reviews are permitted with a 500 sized item pool (Condition 3)

Ability θ	Before Review	After Review	After Rearrangement procedure (APR)	Improvement from rearrangement procedure
-3.0	0.2331	0.1093	0.2205	
-2.5	0.3313	0.3148	0.3391	
-2.0	0.3570	0.3304	0.3333	
-1.5	0.3204	0.2870	0.2916	
-1.0	0.3014	0.2888	0.2929	
-0.5	0.3334	0.3524	0.3568	
0.0	0.3109	0.3206	0.3246	
0.5	0.3165	0.3340	0.3366	
1.0	0.3562	0.3517	0.3562	
1.5	0.3026	0.2937	0.2978	
2.0	0.2966	0.3341	0.3483	
2.5	0.3702	0.3494	0.3800	
3.0	0.2836	0.1066	0.2231	

Table 26 presents the standard error when the Bayesian estimation procedure was used. The rearrangement procedure was effective in reducing the SE of the Bayesian ability estimates after review at 6 of the 13 ability levels. These were at the θ levels of - 1.5, -1.0, -0.5, 0.0, 0.5, and 2.5. This decrease in the standard error tended to be quite small. However, even at the rest of the ability levels, the SE after the

rearrangement procedure was still smaller in magnitude than the SE that existed before item review.

Table 26. Conditional Bayesian standard error when 3 reviews are permitted with a 500 sized item pool (Condition 3)

Ability θ	Before Review	After Review	After Rearrangement procedure (APR)	Improvement from rearrangement procedure
-3.0	0.3086	0.2648	0.2698	
-2.5	0.2766	0.2483	0.2534	
-2.0	0.3049	0.2655	0.2679	
-1.5	0.3024	0.2632	0.2608	Yes
-1.0	0.2984	0.2642	0.2602	Yes
-0.5	0.3124	0.3158	0.3108	Yes
0.0	0.3037	0.2994	0.2976	Yes
0.5	0.3012	0.2997	0.2914	Yes
1.0	0.3232	0.2952	0.3053	
1.5	0.3119	0.2719	0.2752	
2.0	0.3083	0.2923	0.2954	
2.5	0.3565	0.3177	0.3164	Yes
3.0	0.3154	0.2746	0.2791	

Figure 24 provides a comparison of the ML and Bayesian SE at each of the 13 ability levels after the rearrangement procedure has taken place. The overall pattern of the results shows that the ML SE tends to be larger than the Bayesian SE at most of the ability levels. This is consistent with Kim and Nicewander (1993) who concluded that the ML estimator produced the largest standard errors compared to other estimators such as

the Bayesian modal estimation. The only exceptions are at the extremes of the distribution, with the ML SE being smaller than the Bayesian SE at the θ levels of -3.0 and 3.0.

Figure 24. Condition 3 ML and Bayesian standard error after the rearrangement procedure (500 pool and 3 reviews)

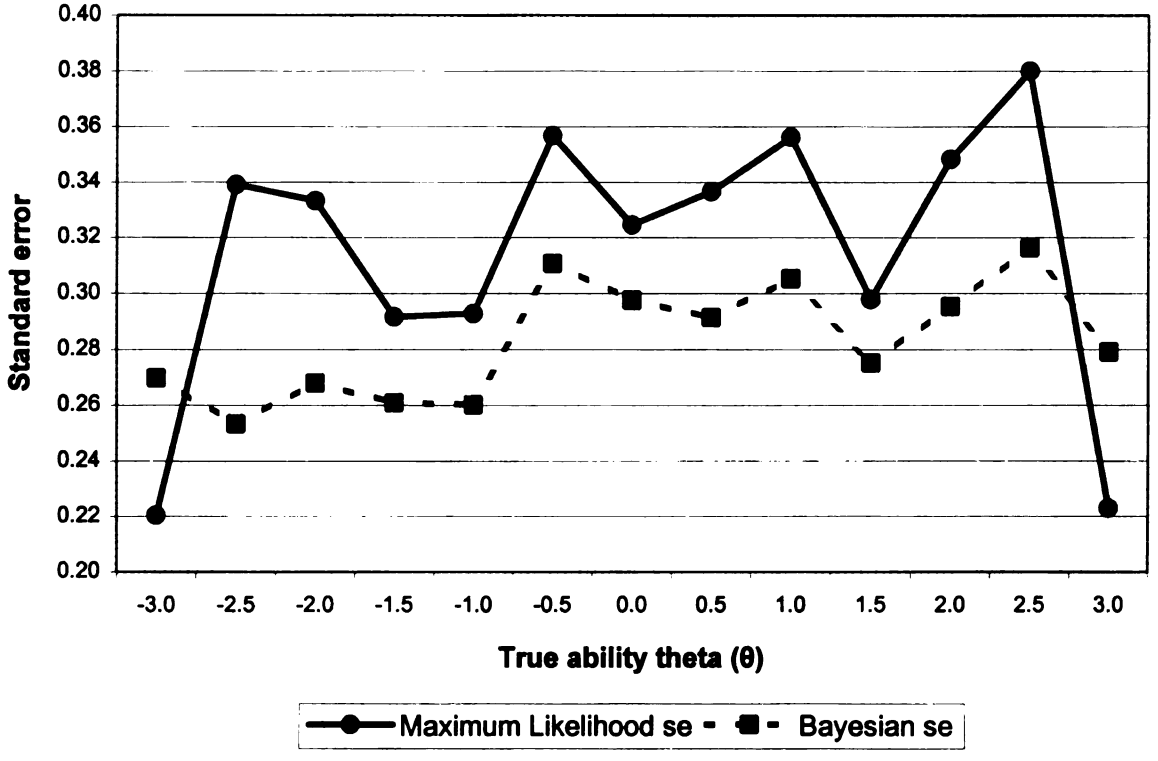
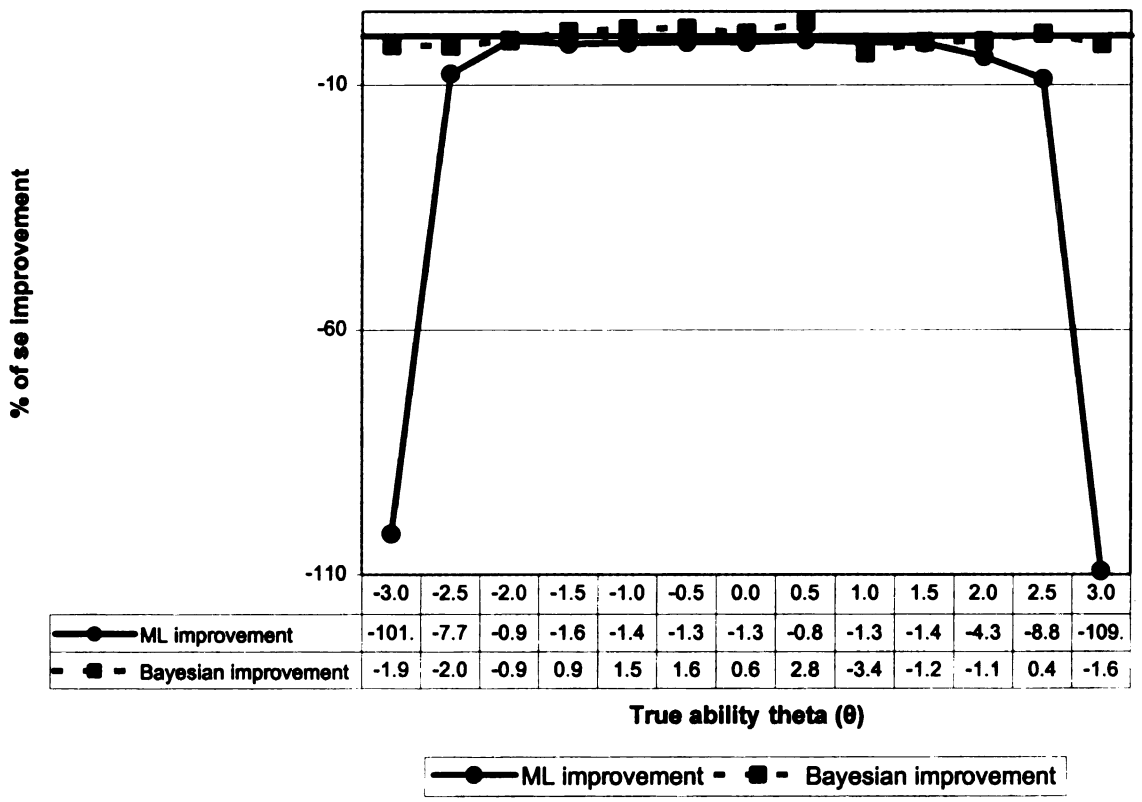


Figure 25 describes the percentage of standard error reduction that has occurred from the after review estimates to the ARP estimates. In the largest portion of the distribution, there do not appear to be major differences in the improvement of the ML and Bayesian standard error, although the improvement in the Bayesian standard error appears to be slightly better than that of the ML. However, the ML estimation procedure appears to increase the standard error greatly at the extremes of the θ distribution. At the θ=-3.0 level, the standard error increases by 101.7%, while at θ=3.0 the standard

error increases by 109.2%. The increase in the standard errors at the extreme of the distribution is a function of the failure of the ML procedure to converge for examinees whose abilities are at the extremes of the distribution. This occurs when examinees get their answers on the test either all correct, or all wrong. So after review, and after the skipping of items in the rearrangement procedure, it is more likely that the examinees at the extremes of the distribution will get their answers either all wrong, or all correct. Consequently, the ML estimation after the rearrangement procedure will have problems converging for these examinees, which in turn increases the ARP SE of the ability estimates.

Figure 25. Percentage of ARP standard error improvement with 3 changes and pool of 500 items (condition 3)



Reliability Of Test Scores (Condition 3)

The reliability of the ability estimates was also compared when a maximum of three reviews were permitted by the examinees in the simulation. As shown in Table 27, the reliability of the ML estimates before review was 0.943. After the examinees changed their answers on the test, the ML reliability estimate dropped to 0.942. After the rearrangement procedure, the reliability dropped further to 0.941. However, this drop in reliability was too small to have a significantly negative effect on the quality of the examinee's final ability estimates.

When the Bayesian estimation procedure was used, the reliability jumped from 0.948 from before review to 0.952 after review. The reliability then increased slightly to 0.953 after the rearrangement procedure took place.

Table 27. Reliability of ability estimates with a pool of 500 items, and 3 permitted reviews (Condition 3)

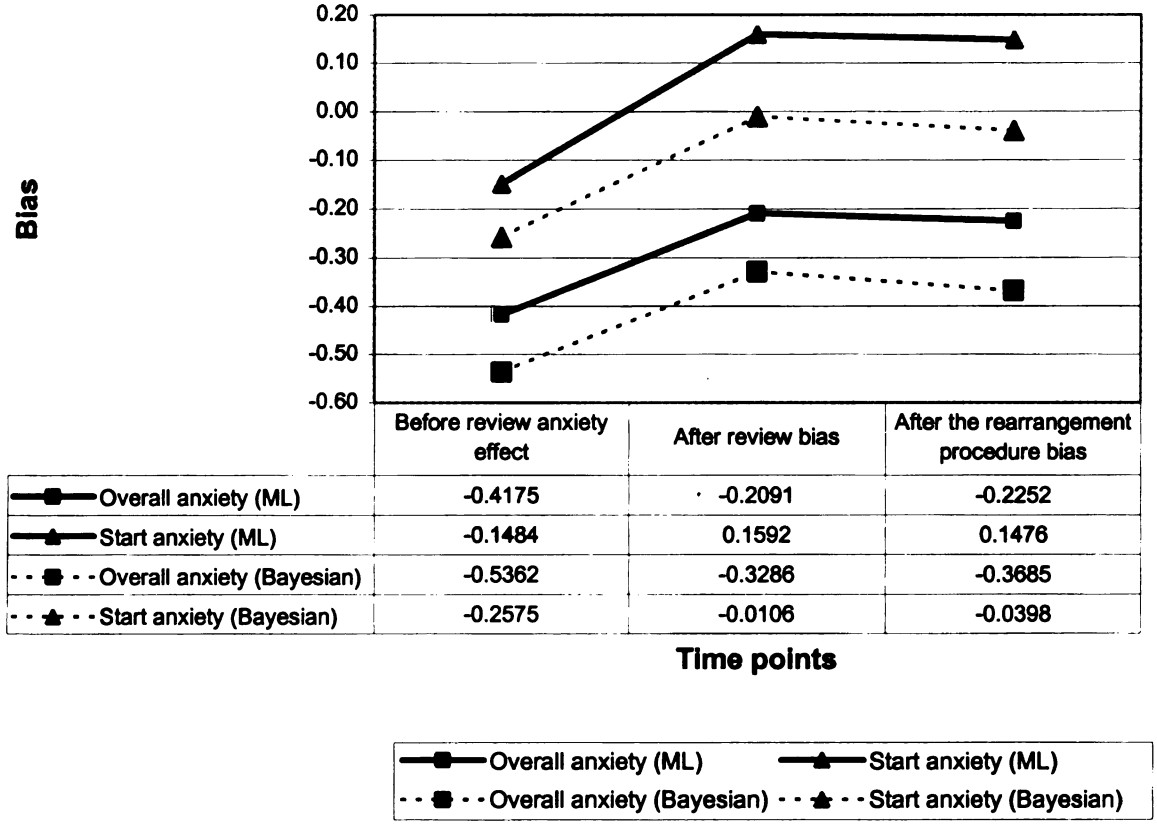
		Reliability
Maximum Likelihood	Before review	0.943
	After review	0.942
	After the rearrangement procedure	0.941
Bayesian	Before review	0.948
	After review	0.952
	After the rearrangement procedure	0.953

Examinee Anxiety Effects (Condition 3)

Figure 26 describes how the simulated examinees with anxiety were affected by the rearrangement procedure in terms of the bias of their score estimates. With the exception of one condition, the examinees who had anxiety obtained more accurate ability estimates after review when compared to their before ability estimates that

contained the anxiety effects. However, the rearrangement procedure was not very effective in reducing the bias of the ability estimates further. With the exception of the ML bias of the examinees with start anxiety, the rest of the ARP bias estimates increased when compared to the bias after review. Even after the increase in the bias after the rearrangement procedure, the ARP ability estimates were more accurate than the before review estimates.

Figure 26. Anxiety effects and bias of the ability estimates (condition 3)

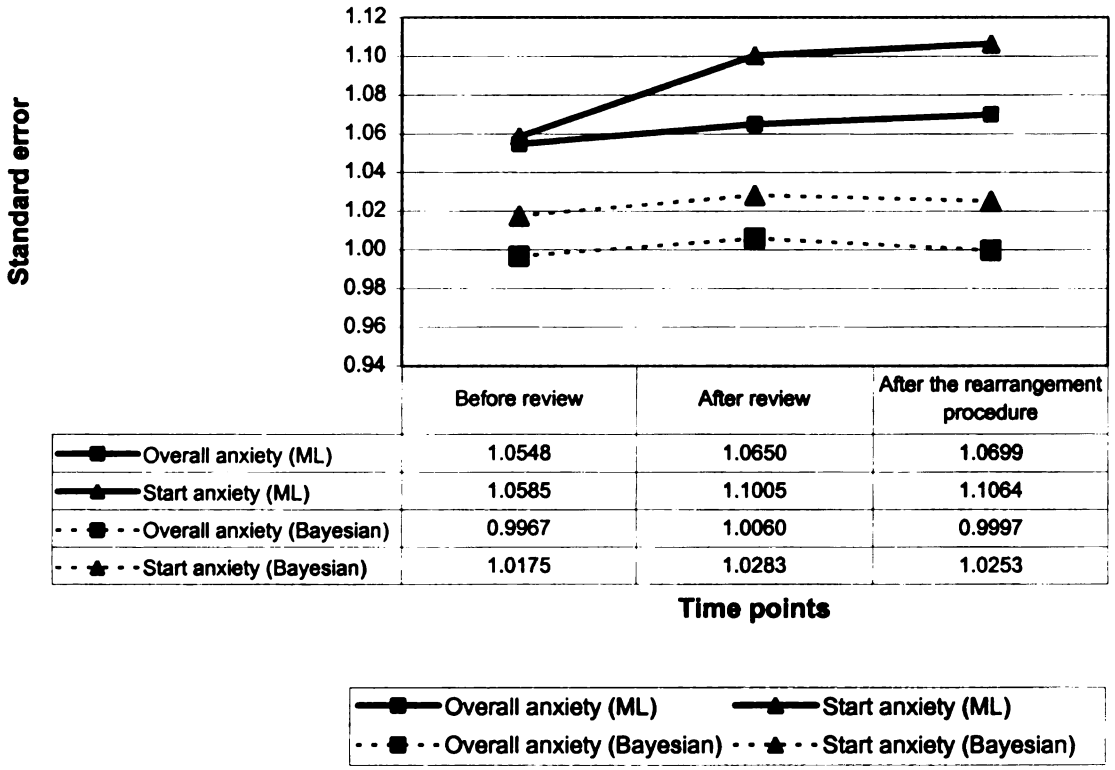


	Before review anxiety effect	After review bias	After the rearrangement procedure bias
Overall anxiety (ML)	-0.4175	-0.2091	-0.2252
Start anxiety (ML)	-0.1484	0.1592	0.1476
Overall anxiety (Bayesian)	-0.5362	-0.3286	-0.3685
Start anxiety (Bayesian)	-0.2575	-0.0106	-0.0398

When comparing the standard errors of the examinee's ability estimates after review, with the before review estimates, it is obvious that under both anxiety conditions and both estimation procedures, the standard error increased after review. After the rearrangement procedure, the SE decreased slightly when the Bayesian estimates were

used. However, these standard errors were still larger than the estimates that existed before review. These results are presented in Figure 27.

Figure 27. Standard error estimates of examinees with anxiety (condition 3)



Condition 4: 500 Items With 5 Reviews Maximum

Condition 4 examined the effects that the rearrangement procedure had on the accuracy of the ability estimates, when an item pool of 500 items was used for the formation of the adaptive tests, and when a maximum of 5 items were permitted to be reviewed by each examinee.

Overall, 41.66% of the simulated examinees made correct-to-incorrect, or incorrect-to-correct changes to their answers. These types of changes are the only ones that will be discussed since the rearrangement procedure takes place only when such changes have occurred. Table 28 describes the percentage of actual answer changes, that are divided in the four categories that were discussed in the previous condition in the simulation.

As can be seen from Table 28, the majority of the changes that were made in the simulation (41.52%) were from incorrect to correct ones. In addition, the majority of those changes were from examinees that made one or two such changes throughout their test. There were also 15.63% of examinees that made correct-to-incorrect changes to questions to which they had approximately a 0.50 probability of answering correctly.

Only 6.08% of the simulated examinees had made 'stupid mistakes' that were then changed to correct answers. Finally, there were also 3.55% of the examinees that changed their answers to incorrect answers to an item that was originally answered correctly just by chance.

After the simulated examinees reviewed their items on the test, the rearrangement procedure was used. Because of the rearrangement procedure, there were 892 examinees (5.87%) that used item review, to which 1 of their items on the test were ignored. There were also 754 examinees (4.96%) that used item review, to which 2 of their items on the test were ignored. Finally, there were also 7907 examinees

(52.01%) of the examinees that used item review, to which 3 of their items on the test were ignored.

Table 28. Percentage of actual answer changing patterns in condition 4 under condition 4

	Number of changes	Number of examinees	Percentage of examinees (out of 26000)
Incorrect-to-correct changes (0.5 probability)	**1**	3640	14.00%
	2	3931	15.12%
	3	2406	9.25%
	4	715	2.75%
	5	104	0.40%
	Sum	**10796**	**41.52%**
Correct-to-incorrect changes (0.5 probability)	**1**	3210	12.35%
	2	734	2.82%
	3	114	0.44%
	4	6	0.02%
	5	0	0.00%
	Sum	**4064**	**15.63%**
Stupid' mistake corrections (incorrect-to-correct)	**1**	1219	4.69%
	2	310	1.19%
	3	42	0.16%
	4	10	0.04%
	5	0	0.00%
	Sum	**1581**	**6.08%**
Unlucky guess' changes (correct-to-incorrect)	**1**	725	2.79%
	2	148	0.57%
	3	46	0.18%
	4	5	0.02%
	5	0	0.00%
	Sum	**924**	**3.55%**

When items were rearranged because of the rearrangement procedure, the amount of information that was provided at each ability level was used as an indicator for which item should be selected to be used next. The average amount of information that was gained by rearranging the items in condition 1 was 0.0514 with a standard deviation of 0.0421. The minimum amount of information that was gained was 0.0001, while the maximum information that was gained was 0.3396.

Results Based On Bias (Condition 4)

Table 29 describes these results in detail. As shown in Table 29, the average ML bias estimates that were -0.1374 before review, decreased in magnitude to 0.0705 after review. The bias decreased in magnitude even further after the rearrangement procedure to 0.0597. This was a 14.4% improvement in the bias estimates when compared to the after review bias. When the Bayesian estimation was used, the bias decreased in magnitude from -0.2641 from before review, to -0.0668 after review. This bias increased after the rearrangement procedure by 61.3% to 0.1078.

Table 29. Overall bias of the 500 pool with 5 reviews estimates (Condition 4)

Bias	Mean	Standard Deviation
Maximum Likelihood Bias		
Before review	-0.1374	0.3229
After review	0.0705	0.3336
After the rearrangement procedure	0.0597	0.3381
Bayesian Bias		
Before review	-0.2641	0.3099
After review	-0.0668	0.2984
After the rearrangement procedure	-0.1078	0.2946

Table 30. Conditional Maximum Likelihood bias when 5 reviews are permitted with a 500 sized item pool (Condition 4)

Ability θ	Before Review	After Review	After Rearrangement procedure (ARP)	Improvement from rearrangement procedure
-3.0	-0.1261	0.0549	-0.1121	
-2.5	-0.2234	-0.0380	-0.0606	
-2.0	-0.1765	-0.1053	-0.1130	
-1.5	-0.1284	0.0666	0.0519	Yes
-1.0	-0.1469	0.0434	0.0339	Yes
-0.5	-0.1595	0.0431	0.0325	Yes
0.0	-0.1398	0.1125	0.0999	Yes
0.5	-0.1422	0.0762	0.0642	Yes
1.0	-0.1034	0.0240	0.0201	Yes
1.5	-0.0883	0.1946	0.1779	Yes
2.0	-0.1442	0.2040	0.1937	Yes
2.5	-0.1087	-0.0223	-0.0089	Yes
3.0	-0.0501	-0.0504	0.1083	

According to Table 30, the after review bias was smaller at most conditional θ levels than the before review bias. However, in some cases, such as at the θ level of 1.5 and 2.0, the after review bias was larger than the before review bias. The reason for this increase is because in certain cases, the review process eliminated the randomness from the examinee's responses. This resulted in a mismatch between the examinee responses and the IRT model. For example, examinees with an ability of θ=2.0 might

have had a 90% probability of answering item i correctly. Consequently, it is expected that 90% of the examinees with a θ of 2.0 would answer item i correctly, and 10% would answer the item incorrectly. However, if 100% of these examinees answer the item correctly, then their response patterns do not match the IRT model, which consequently will increase the after review bias of the ability estimates.

Table 30 also shows that there were 9 out of the 13 ability levels where the ML bias decreased in magnitude after the rearrangement procedure was used. These improvements existed at the θ levels from -1.5 to 2.5. So the effects of the rearrangement procedure were generally more evident at the positive rather than the negative end of the θ scale when the ML estimation procedure is used.

Table 31 presents the bias when the Bayesian estimation procedure was used. The rearrangement procedure was not very effective in reducing the bias of the Bayesian ability estimates after review at most of the ability levels. The only exceptions were the θ levels of - 3.0 and -2.5 where the bias decreased in magnitude with the rearrangement procedure. However, even the at the rest of the ability levels, the bias after the rearrangement procedure was still smaller in magnitude than the bias that existed before item review.

Table 31. Conditional Bayesian bias when 5 reviews are permitted with a 500 sized item pool (Condition 4)

Ability θ	Before Review	After Review	After Rearrangement procedure (ARP)	Improvement from rearrangement procedure
-3.0	-0.0469	0.0826	0.0509	Yes
-2.5	-0.1237	0.0332	0.0056	Yes
-2.0	-0.1808	-0.0699	-0.0841	
-1.5	-0.1979	-0.0166	-0.0476	
-1.0	-0.2309	-0.0435	-0.0816	
-0.5	-0.2621	-0.0787	-0.1139	
0.0	-0.2680	-0.0354	-0.0879	
0.5	-0.2839	-0.0664	-0.1233	
1.0	-0.2874	-0.1641	-0.1723	
1.5	-0.2980	-0.0426	-0.1024	
2.0	-0.3577	-0.0569	-0.1248	
2.5	-0.3570	-0.2234	-0.2497	
3.0	-0.3656	-0.1594	-0.1974	

A comparison of the results from Tables 30 and 31 is presented in Figure 28. This figure describes the bias that existed in the final ARP estimates, at each of the 13 θ levels. When the ML estimate was used, the bias of the test scores increased from a negative bias to a positive bias as the ability estimates increased. So the examinees who were at the lower end of the distribution obtained lower ability estimates than their true abilities after the rearrangement procedure. However, the examinees at the higher ends of the θ scale obtained higher ML ability estimates than their true scores when the

rearrangement procedure was used. This is because the ML procedure is more biased towards the extremes of the θ scale (Kim & Nicewander, 1993; Lord, 1986; Wang & Vispoel, 1998). This bias towards the extremes is a function of the ML estimation procedure, rather than a function of the rearrangement procedure.

With reference to the simulated examinees at the extreme ends of the score scale, the results of condition 4 showed the opposite pattern when the Bayesian estimates were used. For example, simulated examinees at the lower end of the scale had higher Bayesian estimated test scores rather than true test scores after the rearrangement procedure. In contrast, the simulated examinees at the higher end of the scale obtained lower ability Bayesian estimates than their true scores when the rearrangement procedure was used. This is because Bayesian estimates are more biased towards the mean of the θ scale (Kim & Nicewander, 1993; Lord, 1986; Wang & Vispoel, 1998). However, the bias towards the mean is a function of the Bayesian estimation procedure, rather than a function of the rearrangement procedure.

Figure 28. Condition 4 ML and Bayesian Bias after the rearrangement procedure (500 pool and 5 reviews)

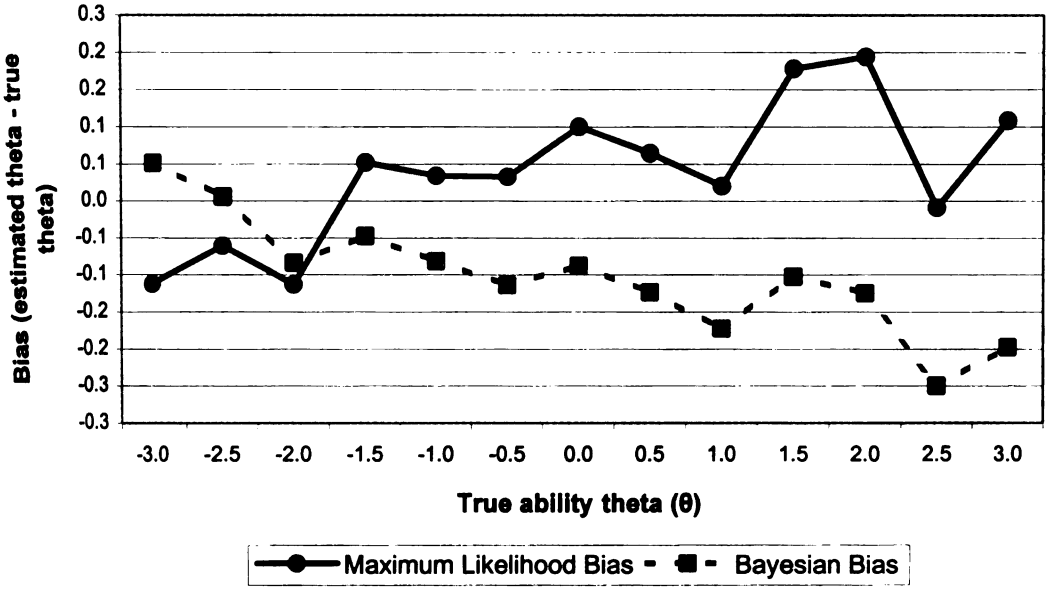
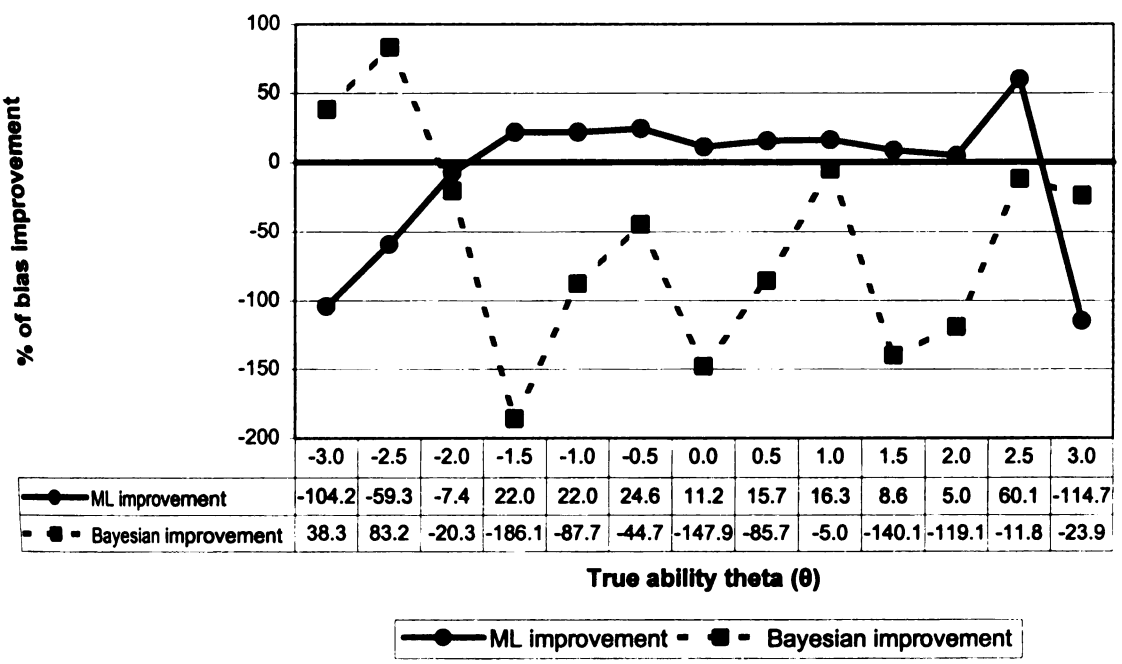


Figure 29 describes the percentage of bias reduction that has occurred from the after review estimates to the ARP estimates. The ML estimation procedure appears to work well at most ability levels since it has a positive percentage of bias improvement 9 of the 13 ability levels. This means that the ML bias decreased due to the rearrangement procedure. The ML estimation procedure was problematic, though for the examinees whose abilities were around -3.0 θ and 3.0. The bias produced by the rearrangement procedure for the examinees at the -3.0 θ was actually 104.2% worse than the bias that existed after review. The ML bias became worse by 114.7% for the examinees whose true θ was 3.0. However, the Bayesian bias tended to increase the bias greatly at all but 2 θ levels. These were the levels of -3.0 and -2.5. So in terms of bias, the ML estimates tend to be more accurate than the Bayesian estimates when the rearrangement procedure is used with condition 4.

Figure 29. Percentage of ARP bias improvement with 5 changes and pool of 500 items (condition 4)



Results Based on the Standard Error of the θ Estimate (Condition 4)

Table 32 describes the average standard deviation of both estimation procedures **at each of the** three time points of the rearrangement procedure. These results are **averaged over** all of the examinees in the sample, including the ones with test anxiety. **According to** Table 32, the ML SD before review was 1.0728. After review, the ML SD **increased to** 1.0891. After the rearrangement procedure, the SD increased further to 1.0940. So the ARP ML SD became slightly worse by 0.4% when compared to the SD that existed after review.

Table 32. Overall standard deviation of the θ estimates obtained from the pool of 500 items, when 5 reviews were permitted (Condition 4)

	Standard Deviation of the θ estimates
Maximum Likelihood	
Before review	1.0728
After review	1.0891
After the rearrangement procedure	1.0940
Bayesian	
Before review	1.0187
After review	1.0282
After the rearrangement procedure	1.0231

The results that were based on the Bayesian estimation showed a different pattern of standard deviation. The Bayesian SD before review was 1.0187. After review, the Bayesian SD increased to 1.0282. After the rearrangement procedure, the standard

deviation decreased to 1.0231.This was a 0.4% improvement in the standard deviation of the Bayesian estimates.

Table 33. Conditional Maximum Likelihood standard error when 5 reviews are permitted with a 500 sized item pool (Condition 4)

Ability θ	Before Review	After Review	After Rearrangement procedure (APR)	Improvement from rearrangement procedure
-3.0	0.2331	0.1110	0.2229	
-2.5	0.3313	0.3200	0.3442	
-2.0	0.3570	0.3302	0.3337	
-1.5	0.3204	0.2904	0.2950	
-1.0	0.3014	0.2938	0.2983	
-0.5	0.3334	0.3541	0.3582	
0.0	0.3109	0.3224	0.3263	
0.5	0.3165	0.3389	0.3404	
1.0	0.3562	0.3516	0.3560	
1.5	0.3026	0.3023	0.3072	
2.0	0.2966	0.3439	0.3607	
2.5	0.3702	0.3464	0.3775	
3.0	0.2836	0.1057	0.2193	

These results from Table 32 are described more analytically in Table 33, which presents the ML SE at each of the 13 ability levels from which the examinees were sampled. Table 33 shows that at no ability level did the standard error of the ARP ML decrease when compared to the after review se. In most cases, the final ARP SE was

even larger than SE that existed before review. This is expected since the standard error tends to decrease when the length of a test is shortened, which is the case with the rearrangement procedure.

Table 34. Conditional Bayesian standard error when 5 reviews are permitted with a 500 sized item pool (Condition 4)

Ability θ	Before Review	After Review	After Rearrangement procedure (APR)	Improvement from rearrangement procedure
-3.0	0.3086	0.2694	0.2721	
-2.5	0.2766	0.2523	0.2545	
-2.0	0.3049	0.2651	0.2674	
-1.5	0.3024	0.2656	0.2635	Yes
-1.0	0.2984	0.2687	0.2640	Yes
-0.5	0.3124	0.3175	0.3118	Yes
0.0	0.3037	0.3001	0.2974	Yes
0.5	0.3012	0.3036	0.2928	Yes
1.0	0.3232	0.2951	0.3051	
1.5	0.3119	0.2822	0.2849	
2.0	0.3083	0.3015	0.3019	
2.5	0.3565	0.3174	0.3154	Yes
3.0	0.3154	0.2759	0.2722	Yes

Table 34 presents the standard error when the Bayesian estimation procedure was used. The rearrangement procedure was effective in reducing the SE of the Bayesian ability estimates after review at 7 of the 13 ability levels. These were at the θ

levels of - 1.5, -1.0, -0.5, 0.0, 0.5, 2.5 and 3.0. This decrease in the standard error tended to be quite small. However, even at the rest of the ability levels, the SE after the rearrangement procedure was still smaller in magnitude than the SE that existed before item review.

Figure 30 provides a comparison of the conditional ML and Bayesian SE at each of the 13 ability levels after the rearrangement procedure has taken place. The overall pattern of the results shows that the ML SE tends to be larger than the Bayesian SE at most of the ability levels. This is consistent with Kim and Nicewander (1993) who concluded that the ML estimator produced the largest standard errors compared to other estimators such as the Bayesian modal estimation. The only exceptions to this pattern are at the extremes of the distribution, with the ML SE being smaller than the Bayesian SE at the θ levels of -3.0 and 3.0.

Figure 30. Condition 4 ML and Bayesian standard error after the rearrangement procedure (500 pool and 5 reviews)

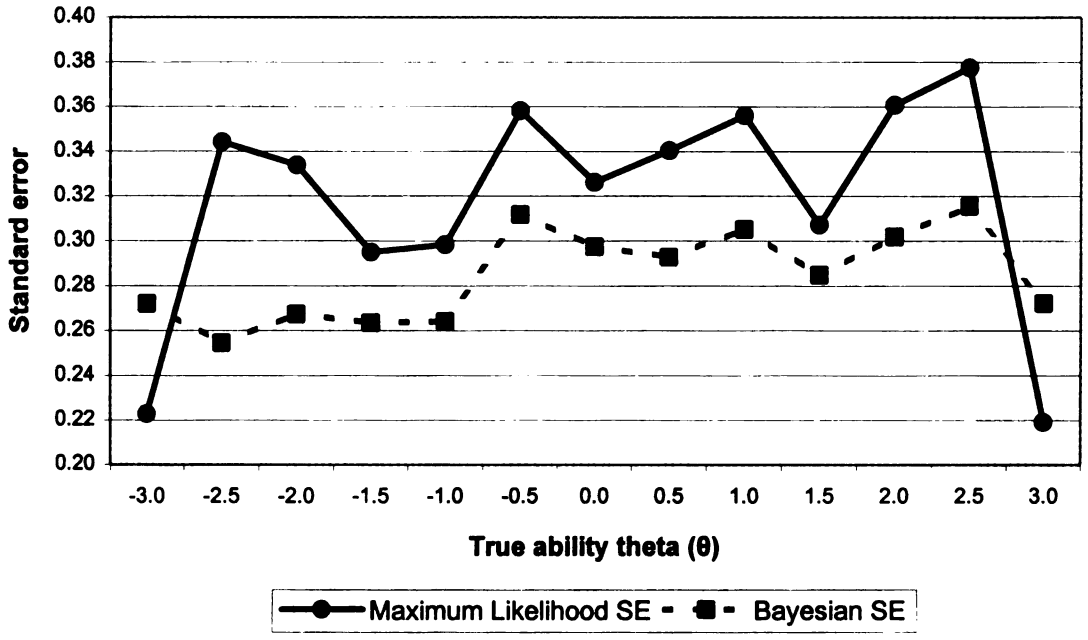


Figure 31 describes the percentage of standard error reduction that has occurred from the after review estimates to the ARP estimates. In the largest portion of the distribution, there do not appear to be major differences in the improvement of the ML and Bayesian standard error, although the improvement in the Bayesian standard error appears to be slightly better than that of the ML se. However, the ML estimation procedure appears to increase the standard error greatly at the extremes of the θ distribution. At the θ=-3.0 level, the standard error increases by 100.8%, while at θ=3.0 the standard error increases by 107.3%. The increase in the standard errors at the extreme of the distribution is a function of the failure of the ML procedure to converge for examinees whose abilities are at the extremes of the distribution. This occurs when examinees get their answers on the test either all correct, or all wrong. So after review, and after the skipping of items in the rearrangement procedure, it is more likely that the examinees at the extremes of the distribution will get their answers either all wrong, or all correct. Consequently, the ML estimation after the rearrangement procedure will have problems converging for these examinees, which in turn increases the ARP SE of the ability estimates.

Reliability Of Test Scores (Condition 4)

The reliability of the ability estimates was finally compared under condition 4 where five reviews were permitted by the examinees with an item pool of 500 items. As shown in Table 35, the reliability of the ML estimates before review was 0.943. After the examinees changed their answers on the test, the ML reliability estimate dropped to 0.941. However, after the rearrangement procedure was used, the reliability of the test scores dropped further to 0.939.

When the Bayesian estimation procedure was used, the reliability jumped from 0.948 from before review to 0.952 after review. The reliability then increased by 0.001 to

0.953 after the rearrangement procedure. This shows that like in condition 4, the Bayesian estimation procedure might be slightly more effective in terms of reliability when used with the rearrangement procedure.

Figure 31. Percentage of ARP standard error improvement with 5 changes and pool of 500 items (condition 4)

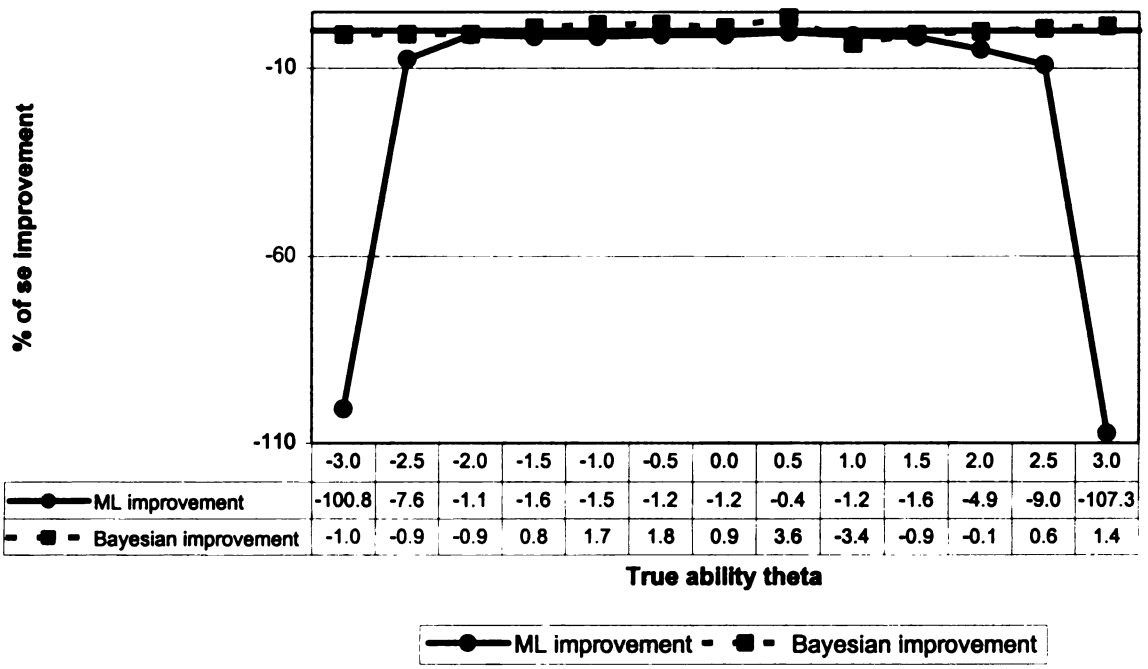


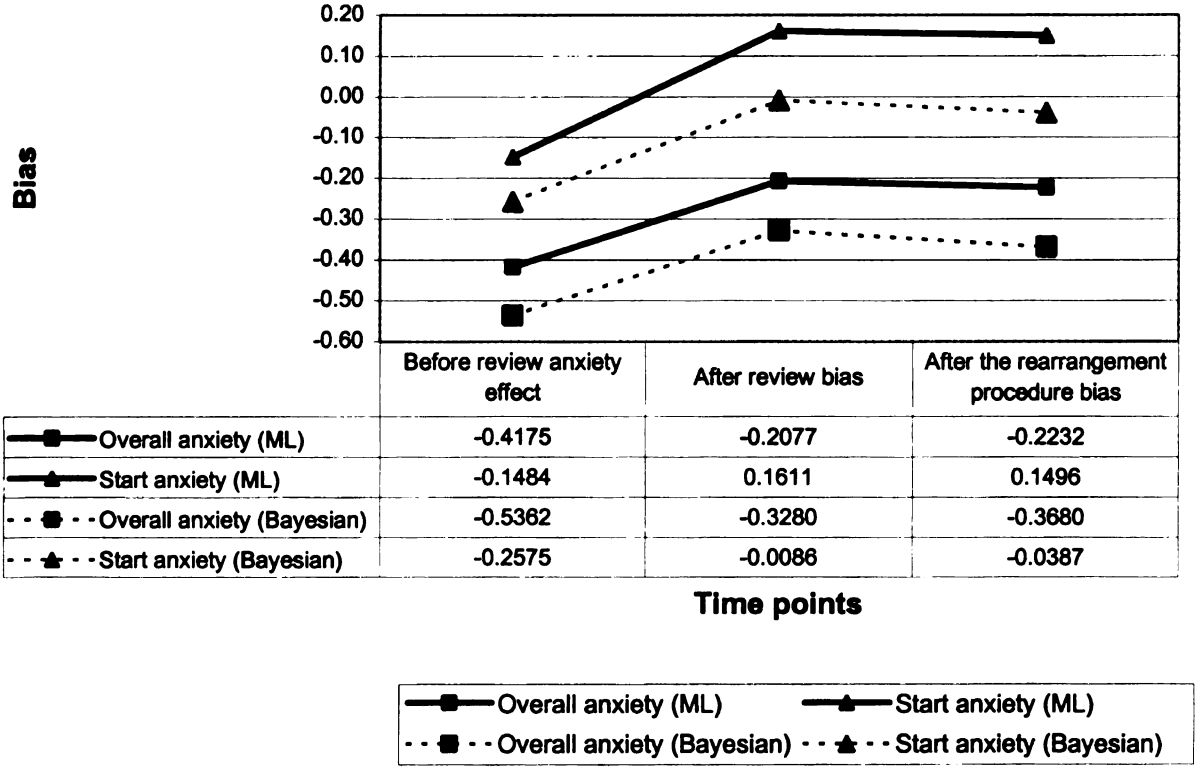
Table 35. Reliability of ability estimates with a pool of 500 items, and 5 permitted reviews (Condition 4)

		Reliability
Maximum Likelihood	Before review	0.943
	After review	0.941
	After the rearrangement procedure	0.939
Bayesian	Before review	0.948
	After review	0.952
	After the rearrangement procedure	0.953

Examinee Anxiety Effects (Condition 4)

Figure 32 describes how the simulated examinees with anxiety were affected by the rearrangement procedure in terms of the bias of their score estimates. With the exception of one condition, the examinees who had anxiety obtained more accurate ability estimates after review when compared to their before review estimates that contained the anxiety effects. However, the rearrangement procedure was not very effective in reducing the bias of the ability estimates further. With the exception of the ML bias of the examinees with start anxiety, the rest of the ARP bias estimates increased when compared to the bias after review. Even after the increase in the bias after the rearrangement procedure, that accuracy of the ARP ability estimates was still better than the before review estimates that contained the anxiety effects.

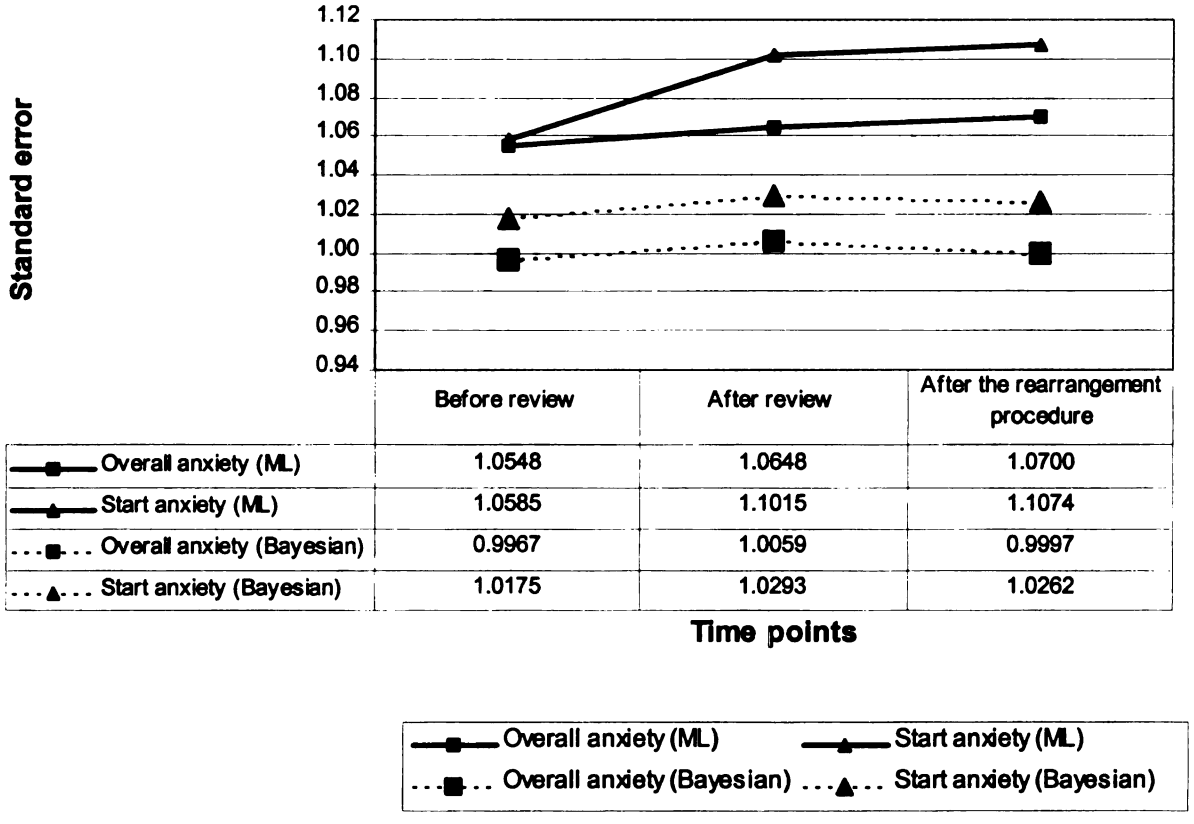
Figure 32. Anxiety effects and bias of the ability estimates (condition 4)



	Before review anxiety effect	After review bias	After the rearrangement procedure bias
Overall anxiety (ML)	-0.4175	-0.2077	-0.2232
Start anxiety (ML)	-0.1484	0.1611	0.1496
Overall anxiety (Bayesian)	-0.5362	-0.3280	-0.3680
Start anxiety (Bayesian)	-0.2575	-0.0086	-0.0387

When comparing the standard errors of the simulated examinee's ability estimates after review, with the before review estimates, it is obvious that under both anxiety conditions and both estimation procedures, the standard error increased after review. However, the SE decreased slightly after the rearrangement procedure when the Bayesian estimates were used. However, these standard errors were still larger than the estimates that existed before review. These results are presented in Figure 33.

Figure 33. Standard error estimates of examinees with anxiety (condition 4)

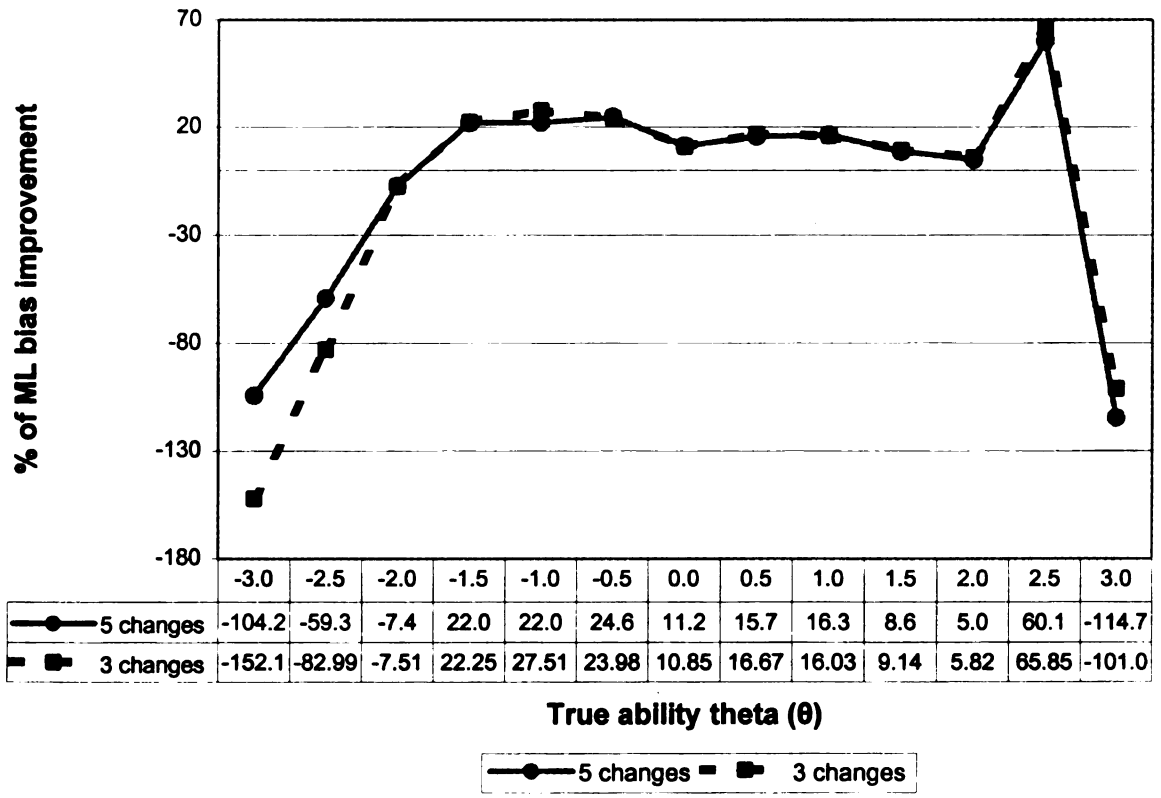


	Before review	After review	After the rearrangement procedure
Overall anxiety (ML)	1.0548	1.0648	1.0700
Start anxiety (ML)	1.0585	1.1015	1.1074
Overall anxiety (Bayesian)	0.9967	1.0059	0.9997
Start anxiety (Bayesian)	1.0175	1.0293	1.0262

Comparison of three and five changes with a 500 size item pool

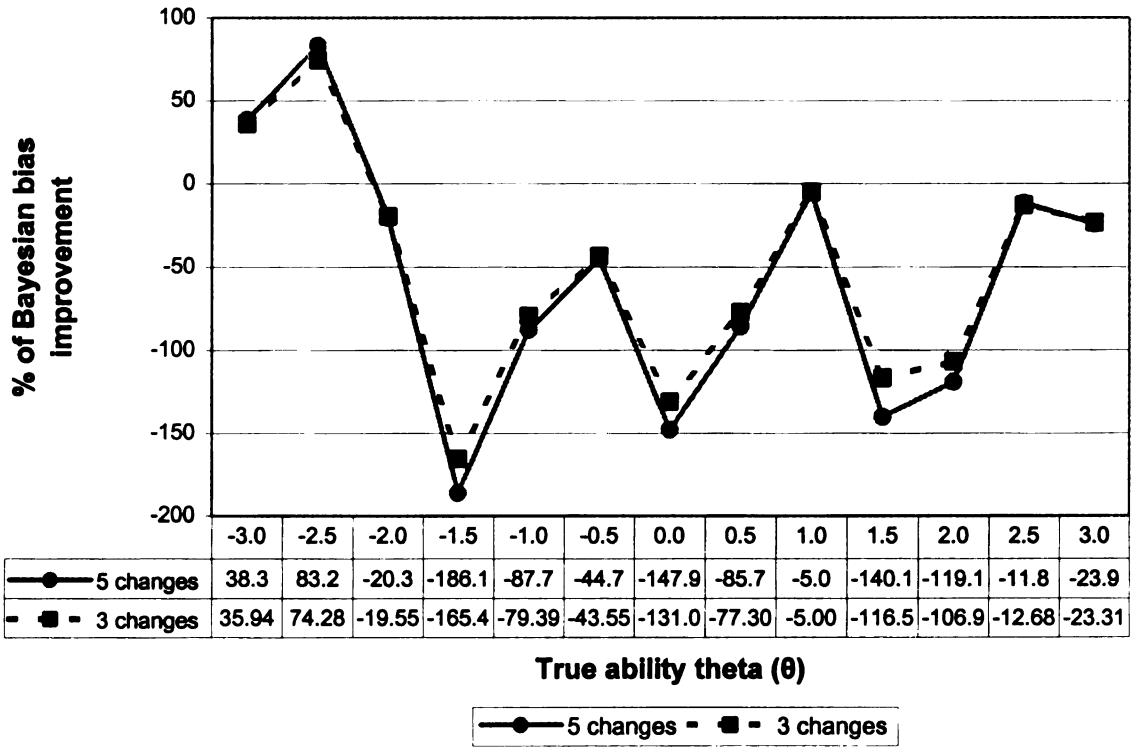
Figures 34 and 35 describe the differences that exist in the amount of item changes have on the improvements on the ARP ML and Bayesian bias. At most θ points, the differences are indistinguishable. The largest difference exists in the ML bias where the bias improvement deteriorates by 152.2% when 3 changes are made, in contrast to 104.2% when 5 changes are made. However, the differences are a bit more pronounced when the Bayesian estimation procedure is used. With the Bayesian procedure, the improvement in the bias tends to be slightly worse when 5 rather than 3 changes are allowed to be made by the examinees.

Figure 34. The effects of the numbers of items reviewed on the ML bias improvement obtained from a pool of 500 items



	-3.0	-2.5	-2.0	-1.5	-1.0	-0.5	0.0	0.5	1.0	1.5	2.0	2.5	3.0
5 changes	-104.2	-59.3	-7.4	22.0	22.0	24.6	11.2	15.7	16.3	8.6	5.0	60.1	-114.7
3 changes	-152.1	-82.99	-7.51	22.25	27.51	23.98	10.85	16.67	16.03	9.14	5.82	65.85	-101.0

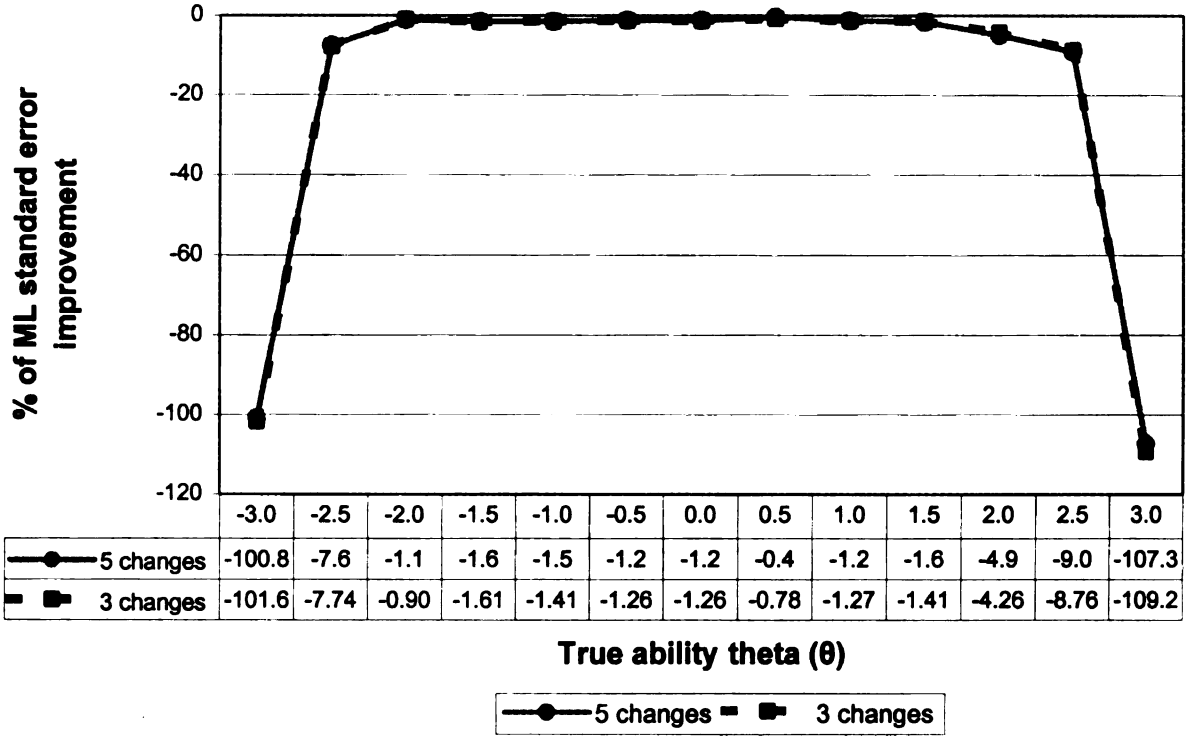
Figure 35. The effects of the numbers of items reviewed on the Bayesian bias improvement obtained from a pool of 500 items



	-3.0	-2.5	-2.0	-1.5	-1.0	-0.5	0.0	0.5	1.0	1.5	2.0	2.5	3.0
5 changes	38.3	83.2	-20.3	-186.1	-87.7	-44.7	-147.9	-85.7	-5.0	-140.1	-119.1	-11.8	-23.9
3 changes	35.94	74.28	-19.55	-165.4	-79.39	-43.55	-131.0	-77.30	-5.00	-116.5	-106.9	-12.68	-23.31

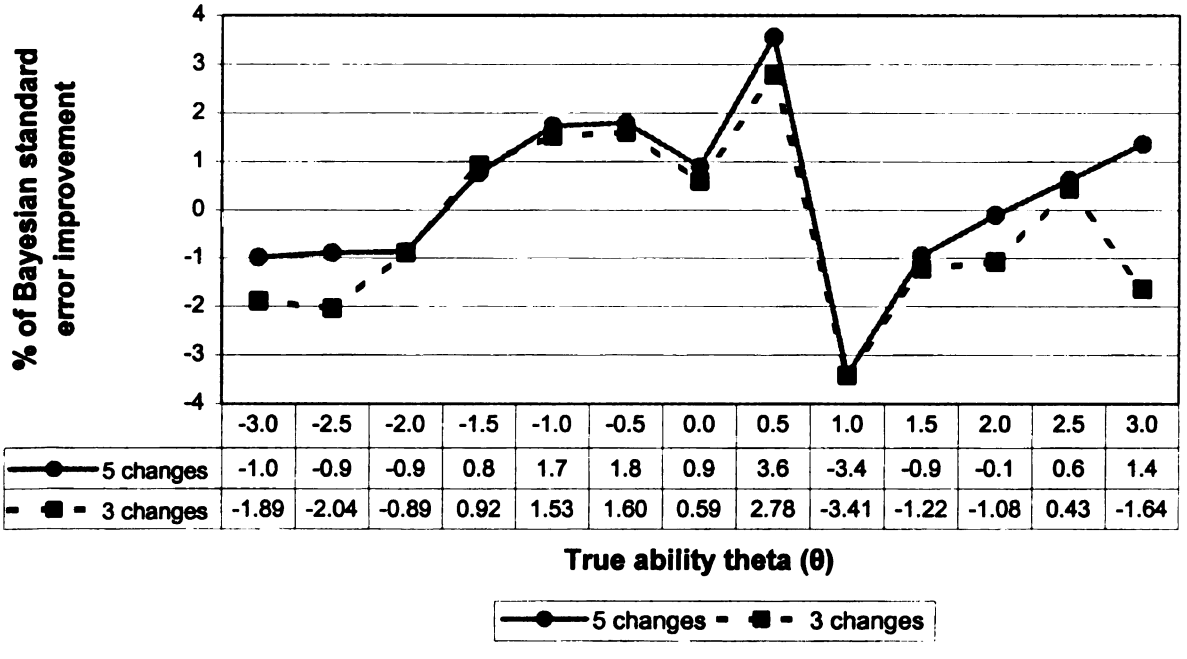
Figures 36 and 37 describe the differences that exist in the improvement of the standard error that occurs after the rearrangement procedure is used. Figure 36 describes the ML standard error improvement differences that occur when 3 or 5 changes are made by the examinees to their answers. Figure 37 describes the Bayesian standard error improvement differences that occur when 3 or 5 changes are made by the examinees to their answers. Both graphs show that are no consistent reasons to prefer the allowance of making a maximum 3 or 5 changes to the answers on a test in terms of the standard error estimates. However, when the Bayesian estimation procedures are used, the increases or decreases in standard errors tend to be smaller in magnitude than when the ML estimation procedure is used.

Figure 36. The effects of the numbers of items reviewed on the ML standard error improvement obtained from a pool of 500 items



	-3.0	-2.5	-2.0	-1.5	-1.0	-0.5	0.0	0.5	1.0	1.5	2.0	2.5	3.0
5 changes	-100.8	-7.6	-1.1	-1.6	-1.5	-1.2	-1.2	-0.4	-1.2	-1.6	-4.9	-9.0	-107.3
3 changes	-101.6	-7.74	-0.90	-1.61	-1.41	-1.26	-1.26	-0.78	-1.27	-1.41	-4.26	-8.76	-109.2

Figure 37. The effects of the numbers of items reviewed on the Bayesian standard error improvement obtained from a pool of 250 items



	-3.0	-2.5	-2.0	-1.5	-1.0	-0.5	0.0	0.5	1.0	1.5	2.0	2.5	3.0
5 changes	-1.0	-0.9	-0.9	0.8	1.7	1.8	0.9	3.6	-3.4	-0.9	-0.1	0.6	1.4
3 changes	-1.89	-2.04	-0.89	0.92	1.53	1.60	0.59	2.78	-3.41	-1.22	-1.08	0.43	-1.64

Item Pool Size Differences

Since there were very slight differences in the conditional bias and standard errors when 5 rather than 3 changes were permitted on the test, the item pool size effects will be discussed with the conditions that permit up to 5 items to be reviewed. Figure 39 describes the ML and Bayesian bias for the item pools that contained either 250 or 500 items. Based on Figure 38, the differences that are produced from the ML estimates because of the item pool size are very minimal. After review, the bias of the ability ML estimates was smaller by 0.0013 when the item pool of 500 items was used instead of the 250 sized item pool. With the Bayesian after review estimates, the bias was smaller with the item pool of 250 items, by 0.0009. After the rearrangement procedure, the pool of 500 items improved the ML bias by 0.0013 compared to the pool of 250 items. However, with the Bayesian estimation procedure, the 250 sized item pool had an ARP bias that was smaller than the 500 pool bias by 0.0012. So based on the results of the bias, using item pools that are larger than 250 items do not necessarily improve the bias of the ability estimates when the rearrangement procedure is used.

Figure 38. Item pool size effects on the estimation bias when 5 reviews are permitted

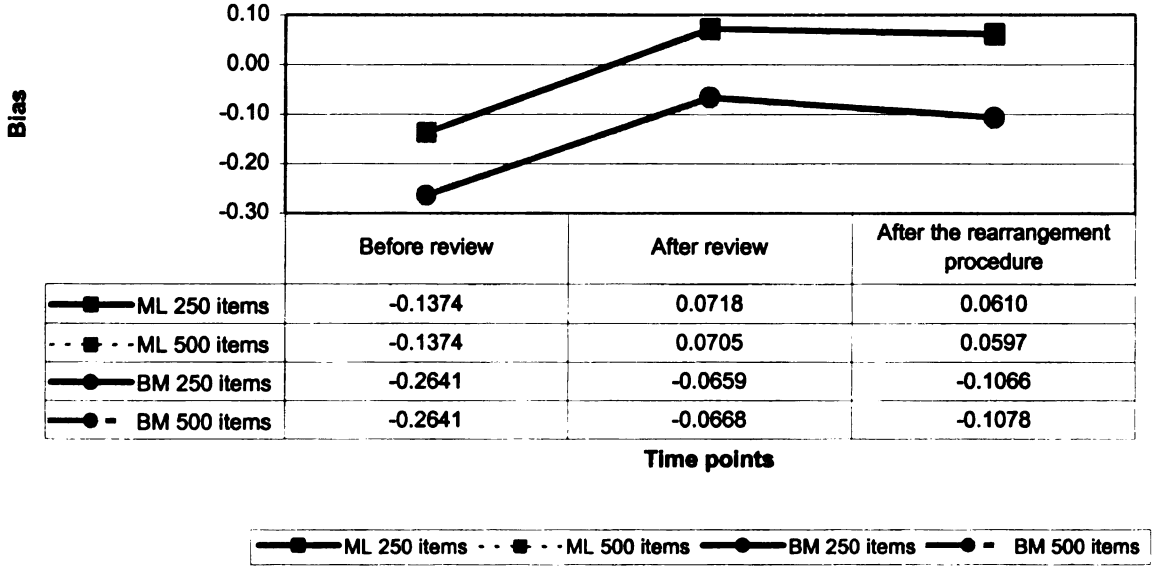
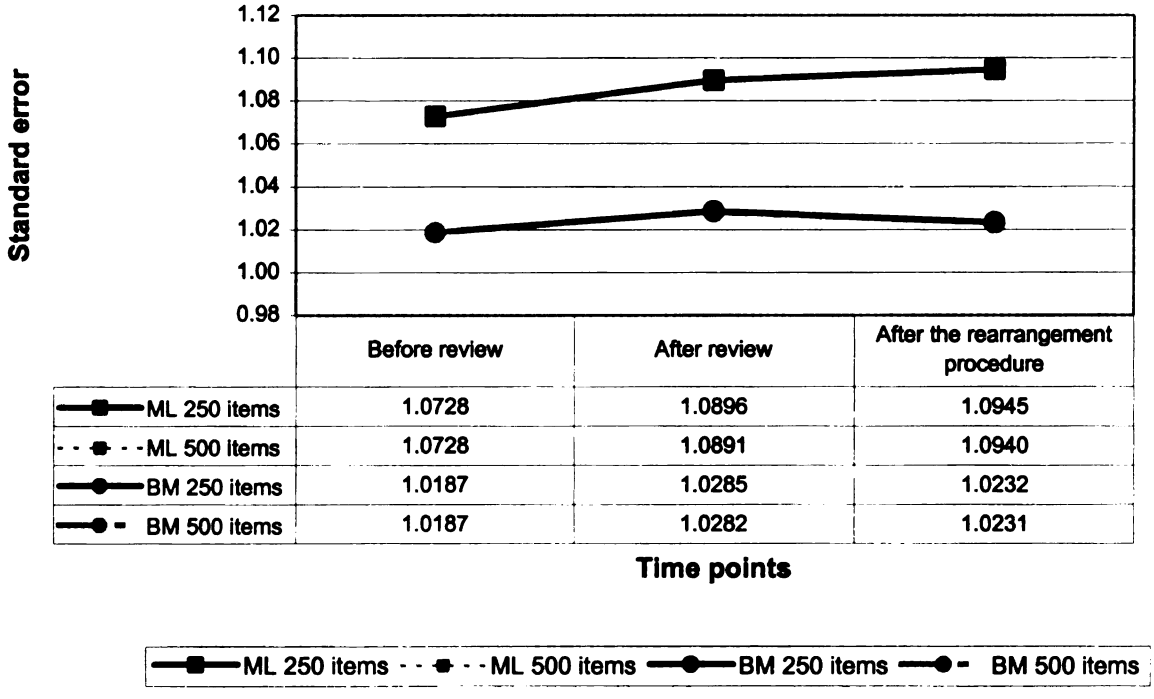


Figure 39 describes the differences that occurred because of the item pool size on the standard errors of the ability estimates. Overall, the item pools with 500 items managed to provide slightly smaller standard errors after review and after the rearrangement procedure, than the pool that contained 250 items. However, these differences are as small as 0.0001. For this reason, again, there is no reason to require larger item pools than 250 items to be used with the rearrangement procedure.

Figure 39. Item pool size effects on the standard error when 5 reviews are permitted



	Before review	After review	After the rearrangement procedure
ML 250 items	1.0728	1.0896	1.0945
ML 500 items	1.0728	1.0891	1.0940
BM 250 items	1.0187	1.0285	1.0232
BM 500 items	1.0187	1.0282	1.0231

Overall Final Comparison

Table 36 provides an overall comparison of the effects that the rearrangement procedure had on the bias, the standard errors, and the reliability of the simulated examinee's ability estimates in all 4 conditions that were examined in this study. In terms of the amount of items that were permitted to be changed, there were no large or consistent differences in the final ability estimates, when 3 or 5 changes were made to the test items. In some conditions the estimates improved slightly when 3 changes were permitted, while in other conditions, 5 changes were more beneficial for the accuracy of the final ability estimates. For example, the ML bias estimates were smaller when 3 items were permitted to be changed rather than 5. However, the Bayesian bias was smaller when 5 reviews were permitted rather than 3.

When the item pool sizes were compared, there were also no strong reasons to prefer an item pool of 250 or 500 items to be used with the rearrangement procedure. However, in terms of the reliability of the final ability estimates, the reliability was consistently higher when the tests were developed from item pools of 500, rather than from 250 items. However, the rearrangement procedure was able to work equally as well with both sizes of item pools.

When comparing the effects of the ML to the Bayesian estimates, the results are also mixed. For example, the ML estimates provide more accurate ability estimates with the rearrangement procedure, when the bias is considered as an indicator of quality of the rearrangement procedure. When the standard error was used as an indicator of quality of the rearrangement procedure, the Bayesian estimates appear to consistently work better than the ML estimates. Finally, with the reliability was the indicator of the effectiveness of the rearrangement procedure, the Bayesian estimates tended to show more of an effect of the rearrangement procedure.

Table 36. Overall results comparison

		250 item pool		500 item pool	
		3 changes	5 changes	3 changes	5 changes
ML Bias	Before review	-0.1374	-0.1374	-0.1374	-0.1374
	After review	0.0673	0.0718	0.0668	0.0705
	After rearrangement procedure	0.0567	0.0610	0.0562	0.0597
	% improvement from rearrangement procedure	15.75%	14.98%	15.90%	15.27%
Bayesian Bias	Before review	-0.2641	-0.2641	-0.2641	-0.2641
	After review	-0.0687	-0.0659	-0.0693	-0.0668
	After rearrangement procedure	-0.1087	-0.1066	-0.1090	-0.1078
	% improvement from rearrangement procedure	-58.14%	-61.78%	-57.33%	-61.34%
ML Standard Error	Before review	1.0728	1.0728	1.0728	1.0728
	After review	1.0878	1.0896	1.0873	1.0891
	After rearrangement procedure	1.0928	1.0945	1.0923	1.0940
	% improvement from rearrangement procedure	-0.45%	-0.45%	-0.46%	-0.45%
Bayesian Standard Error	Before review	1.0187	1.0187	1.0187	1.0187
	After review	1.0273	1.0285	1.0267	1.0282
	After rearrangement procedure	1.0223	1.0232	1.0220	1.0231
	% improvement from rearrangement procedure	0.48%	0.51%	0.46%	0.50%
ML Reliability	Before review	0.821	0.817	0.943	0.943
	After review	0.818	0.811	0.942	0.941
	After rearrangement procedure	0.810	0.806	0.941	0.939
	% improvement from rearrangement procedure	-0.93%	-0.63%	-0.19%	-0.19%
Bayesian Reliability	Before review	0.834	0.830	0.948	0.948
	After review	0.847	0.841	0.952	0.952
	After rearrangement procedure	0.849	0.843	0.953	0.953
	% improvement from rearrangement procedure	0.20%	0.24%	0.05%	0.09%

CHAPTER 5

CONCLUSIONS

Due to the increased popularity of computerized adaptive testing, many high stakes tests, certification, or even achievement tests are now being converted to a computer adaptive format. However, as researchers are trying to improve many of the components of CAT, students are trying to familiarize themselves with the new testing format and processes of CAT (Pommerich & Burden, 2000; Reckase, 2000). One of the components of adaptive tests that creates some tension between students and CAT researchers, and which has not been conclusively resolved yet, is that of item review. On the one hand, examinees prefer to have the option of changing their answers on adaptive tests (Bowles & Pommerich, 2001). They argue that item review allows them to perform to the maximum extent of their abilities since they are able to rethink over their answers, as well as to correct questions that might have been misread, miskeyed, or miscalculated. This is especially important for examinees that have test taking anxiety (Wise, Roos, Plake & Nebelsick-Gullett, 1994; Wise, 1997) and who are very likely to make careless errors on such tests.

However, some researchers believe that item review should not be permitted on CAT since it does not follow the logic on which adaptive tests are based on (Wise, 1996), and since item review might actually hurt the accuracy of the examinee's ability estimates. For this reason, a rearrangement procedure was proposed in this study. This rearrangement procedure that is used after item review takes place, was hypothesized to improve the accuracy of the examinee ability estimates, without allowing the examinee's to artificially inflate their ability estimates. This procedure was examined under the following conditions: a) different sized item pools, b) different number of items

reviewed, c) with the Maximum Likelihood and the Bayesian Mean estimation procedures, and d) with examinees that have two forms of test anxiety. The major conclusions that have been reached based on the research questions of this study are discussed below.

How Does The Rearrangement Procedure Affect The Bias Of The Ability Estimates?

When the ML estimation was used for obtaining the examinee ability estimates, the ML bias became smaller after the rearrangement procedure. Although this was consistent with all 4 conditions of the simulation, the smallest bias was obtained under condition 3 of the simulation study. This was the condition where examinees could make up to 3 changes to their tests, that was created from an item pool of 500 items. This condition had a decreased of 15.9% in the bias when compared to the bias that existed after review.

When looking at the conditional bias of the ML estimates, the overall pattern of bias showed that the ML bias tends to increase from a negative to a positive bias as the ability of the examinees increases. Therefore, examinees with lower ability estimates tend to have a negative ML bias, while examinees with higher abilities tend to have a positive ML bias. This means that examinees at the lower end of the distribution have lower estimated scores than true scores when the ML is used with the rearrangement procedure. The situation is the opposite for examinees at the higher end of the ability distribution where a positive bias exists. This is expected since the ML estimator is more biased towards the extremes (Kim & Nicewander, 1993; Lord, 1986; Wang & Vispoel, 1998). However, this is a reflection of the properties of the estimator rather than a reflection of the rearrangement procedure.

When the Bayesian estimation procedure was used for estimating the examinee's abilities, the results were quite different. Although the Bayesian bias decreased with review, the bias tended to increase in magnitude with the rearrangement procedure. This increase could be as large as 61.8% when five reviews were permitted in item pools of 250 items. However, even with that increase, the bias tended to be smaller in magnitude than the bias before review. So the ML estimates appear to reflect the examinee's true score more accurately than the Bayesian estimates.

In contrast to the ML bias, the Bayesian estimator is biased towards the mean (Kim & Nicewander, 1986; Wang & Vispoel, 1998). So examinees at the lower end of the ability distribution tend to have higher Bayesian estimates than true scores with the rearrangement procedure. This is because the Bayesian bias is positive at the lower end of the distribution. Examinees at the higher end of the distribution who have negative bias, have lower Bayesian ability estimates than their true scores after the rearrangement procedure is used. Again, this is a reflection of the properties of the estimator rather than a reflection of the rearrangement procedure.

The size of the conditional bias at each ability level can also be compared to the effect size indices that were proposed by Cohen (1992). According to Cohen's standards, all of the biases before review, after review, and after the rearrangement procedure can be considered as small effect sizes for all conditional ability levels, and for all three points of the rearrangement procedure. Therefore, the differences in the bias between the three points of the rearrangement procedure are even smaller. However, since these differences have appeared in a simulation from a dataset of 26000 examinees, that means that they are real differences, and that they cannot be attributed solely to sampling error.

What Are The Effects Of The Rearrangement Procedure On The Reliability Of The Examinee's Final Ability Estimates?

The changes in reliability because of the rearrangement procedure were very small and they probably do not reflect significant effects on the ability estimates. These changes also differed based on the ability estimation procedure that was used. So when the ML procedure was used, the reliability tended to decrease with item review, and with the rearrangement procedure. When the Bayesian estimation was used, the reliability increased with item review. The reliability continued to slightly increase when the rearrangement procedure was used. So the Bayesian estimation appears to work better in terms of increasing the reliability of the ARP estimates. However, this increase is too small to have a significant effect on the overall results of the study.

How Does The Rearrangement Procedure Affect The Ability Estimates Of The Examinees That Have Computerized-Test-Anxiety?

The overall results of this study have shown that the accuracy of the ML and Bayesian ability estimates tend to increase with item review for the examinees that have test anxiety. However, the rearrangement procedure is not very effective in reducing the bias of the estimates further with the rearrangement procedure. A possible reason for that is because examinees are limited to the amount of items that they can change. Consequently, they do not have the opportunity to change all of the answers that they might have answered incorrectly because of their anxiety. For example, an examinee with start anxiety might have made errors in answering the first 7 items on the test. However, the rearrangement procedure only allowed 3 or 5 revisions. For this reason, there would still be items with aberrant responses for this examine that were not revised. So these items would decrease the accuracy of the ability estimates. However, the

accuracy of their final ability estimates after the rearrangement procedure was still better than the before review ability estimates that contained the anxiety effects.

The ARP standard errors of the ability for the examinees with anxiety were consistently larger than the standard errors before item review. This was not surprising since the standard errors tend to increase when the length of a test is shortened, which is the case with the rearrangement procedure.

Since this is a simulation study, however, it is very difficult to determine if the anxiety behavior that has been simulated by the examinees is comparable to the behavior of real examinees. For this reason, these results should be interpreted very cautiously.

How Does The Choice Of The Ability Estimation Procedure Affect The Results From The Rearrangement Procedure?

The Maximum Likelihood procedure appears to be more accurate than the Bayesian in terms of reducing the estimation bias when the rearrangement procedure is used. The final ML bias produced after the rearrangement procedure is actually smaller than the Bayesian bias that exists under any condition in any of the three time points of the rearrangement process. However, the Bayesian procedure is more effective in reducing the standard error of the ability estimates, and in increasing the reliability of the test scores after the rearrangement procedure. It is not surprising that the ML estimates have more standard error and lower reliability than the Bayesian estimates. This is a function of the ML estimator, that tends to have larger standard error (Lord, 1986) and lower reliability when compared to the Bayesian procedure (Kim & Nicewander, 1993).

Consequently, it is very difficult to determine which estimation procedure is more appropriate to be used with the rearrangement procedure. That would depend on what dependent variable is considered as more appropriate to reflect effects of the

rearrangement procedure. On the one hand, it was expected that the standard error of the ability would increase when items were removed from the estimation of the final test score. This can be remedied, however, by adding additional items to the test to compensate for the items that have been ignored by the rearrangement procedure.

On the other hand, the changes in the reliability estimates are so small, that they should not have any significant effect on the results of the final ability estimates for any individual examinee. However, the results tend to be more pronounced when the bias is used as an indicator of the effects of the rearrangement procedure. In this case, the ML procedure is more effective in estimating the examinee's abilities more accurately.

How Does The Maximum Number Of Item Changes Affect The Examinee's Final Ability Estimates?

Overall, there were very slight differences in the ARP biases when a maximum of 3 or 5 reviews were permitted. These differences were too insignificant to have a strong rationale for permitting 3 or 5 reviews. Consequently, there is no reason to deny the examinees from changing 5 of their answers on a 30-item test, especially since not all the examinees choose to change that many items. In addition, since the examinees will not be able to perform the Wainer strategy with only 5 item reviews, then this cheating strategy should not be an issue for prohibiting examinees from changing their answers on adaptive tests.

Implications For Practice

Overall, the rearrangement procedure has shown some positive and promising results. On the one hand, it is associated with item review that permits examinees to change any mistakes that they might have made, such as miskeyed, miscalculated, or misread, items. These corrections will make the examinees' final ability estimates more

valid since the careless errors will be removed from the final test scores (Vispoel, 1998a). In addition, many examinees will have less anxiety when they realize that they can go back and change some of their answers on the test (Wise, Roos, Plake & Nebelsick-Gullett, 1994; Wise, 1997). This will also allow them to pace themselves better throughout the test when they know that they can come back to an item and spend more time on it after they have reached the end of the test.

However, item review is also associated with two main cheating strategies, the Wainer and the Kingsbury strategy. The Kingsbury strategy should not be a major issue for item review since the current research has shown that examinees are not able to use this strategies effectively to artificially inflate their test scores, even when they are taught to do so (Vispoel, Clough, Bleiler, Henderickson & Ihrig, 2001). The examinees that want to cheat will not be able to perform the Wainer strategy either, since that requires changing the answers to all of the items on a test. Since the rearrangement procedure only permits up to five item reviews, then this cheating strategy cannot be effectively used.

In addition, the rearrangement procedure itself does not reduce the efficiency of CAT. It does not require extra testing time for the examinees, and it does not require the administration of additional items either. The rearrangement procedure is just an algorithm that can be used with the ability estimation procedures after the test has ended.

In terms of the rearrangement procedure itself, it is effective in the sense that it can reduce the overall ML bias of the ability estimates, and the Bayesian standard error estimates. The ARP reliability also increased slightly with the Bayesian estimates. For this reason, if the rearrangement procedure were adopted, the ability estimation procedure that would be used, would have to depend on the dependant variable that would be used as an index of the effectiveness of this procedure. So if the standard error

or the reliability are considered more accurate indicators of the quality of the estimates, then the Bayesian estimation should be used with the rearrangement procedure. If the bias is considered as a more accurate indicator of the quality of the estimated, then the ML estimation procedure should be used.

Since there were very small differences that the size of the two item pools had on the rearrangement procedure, then a regular item pool of about 250 items is adequate for developing tests that can be used with the rearrangement procedure. In addition, since there were so small differences when 3 or 5 items were permitted to be reviewed, then there is no reason to prohibit examinees from making up to 5 changes to their test answers if they choose to do so.

Limitations

A large component that is missing from this study is the use of real data. Since very few adaptive tests allow review, obtaining real data to base this model on was very difficult. For this reason, the specifications of this study were based on prior studies that have dealt with CAT and with item review. The only aspect of this simulated CAT, on which no prior research has been done, was on the characteristics of the items that are reviewed by the examinees. For this reason, more research needs to be done with real data from adaptive tests that permitted review, to ensure that the positive effects of the rearrangement procedure can be replicated.

It might also be argued that it is difficult for testing organizations to explain why certain items have been omitted from the examinee's final ability estimates. However, many CATs administer seed items in their tests to pilot them and judge their quality. Such items are not used for the estimation of the examinee's abilities either. So the examinees can just be informed at the beginning of the test that some additional items might be omitted from their test scores in order to improve their ability estimates. In case

that it is too risky to omit items on high stakes tests, this procedure could still be used for general achievement and aptitude tests.

Another possible limitation of the rearrangement procedure, is that omitting items from the test will reduce the total amount of test information (Lunz, Bergstrom and Wright, 1992; Wainer, 1993). This is correct. A possible solution to overcome this problem would be to add as many additional items at the end of the test, as the number of items that have been omitted. This would create a large increase in the test information since these additional items would be perfectly targeted to the examinee's 'corrected' ability estimate after review. However, testing organizations would have to judge the feasibility of this solution since administering more items would be more costly to them.

One way in which the item pool was different from a real item pool, is that the distribution of the b-parameters was uniform in the simulation study. In reality, however, the distribution of b-parameters is approximately normal. The reason why a uniform distribution was used in this simulation, was to ensure that the item pool would contain an adequate amount of items that are appropriate for examinees of all ability levels. So it is possible that the results of the simulation would have been different if the distribution of b-parameters was normal. It is hypothesized that with such a distribution, the item pool that contained 500 items would have produced more accurate ability estimates, than the pool with 250 items. This difference would be more pronounced for the examinees at the extremes of the distribution, since the pool of 500 items would have a larger variety of very easy and very difficult items that it could administer to the very low and very high ability examinees, correspondingly. However, more research needs to be done to examine the effects that the item parameter distributions have on the rearrangement procedure.

The scaling factor of 1.7 was not used anywhere in the simulation procedure. Since I did not have to compare the distribution of the normal ogive model to the logistic model's distribution, there was no need to include the scaling factor in the simulation process. However, it would be interesting to examine if there would be significant differences in the rearrangement procedure results if this scaling factor were used. My hypothesis is that the difference in the results would be minimal.

Finally, it is essential for test developers to pilot the use of the rearrangement procedure before applying it to CAT. It is possible that this procedure might have different results when real item pools are used. Consequently, the rearrangement procedure needs to be examines with a) real data, b) variable length adaptive tests, c) varying item selection procedures, d) with item selection constraints such as item exposure controls, and e) with polytomous items..

APPENDIX

APPENDIX A

SAS PROGRAM CODE

```
*THIS IS THE CODE FOR CONDITION2 OF THE SIMULATION;
libname dat 'c:/windows/desktop/elena' ;
*************************************************************;
*PART 1. DATA SETUP;
*************************************************************;

*1. INPUTS THE A B C PARAMETERS FROM THE FILE;
data dat1;
filename itt 'c:/windows/desktop/elena/fn25n.itt';
infile itt missover;
input a b c;
run;

data data1;
set dat1;
item=_n_;
proc sort; by item;
run;


*2. INPUTS THE SEM FROM THE FILE;
data data2;
filename ase 'c:/windows/desktop/elena/fn25n.ase';
infile ase missover;
input
thetagroup theta thetahat finalsem junk1 junk2
sem1 sem2 sem3 sem4 sem5
sem6 sem7 sem8 sem9 sem10
sem11 sem12 sem13 sem14 sem15
sem16 sem17 sem18 sem19 sem20
sem21 sem22 sem23 sem24 sem25
sem26 sem27 sem28 sem29 sem30;
run;


*3. NUMBERS EACH EXAMINEE;
data data3 (drop=junk1 junk2);
set data2;
case=_n_;
proc sort; by case;
run;
```

```
*4. INPUTS THE ITEMS FROM THE FILE;
data data4;
filename trs 'c:/windows/desktop/elena/fn25n.trs';
*filename trs 'c:/cbts/fn25n.trs';
infile trs missover;
input
status thetagroup theta bias length  finscore allresponses $32.
item1  item2 item3  item4   item5
item6  item7 item8  item9  item10
item11 item12 item13  item14  item15
item16 item17 item18 item19 item20
item21  item22 item23  item24  item25
item26 item27 item28  item29  item30;
run;


*5. CHANGES THE STRING VARIABLE THAT INCLUDED ALL OF THE ITEM
RESPONSES IN ONE; *VARIABLE TO 30 INDIVIDUAL RESPONSES IN 30
SEPARATE VARIABLES;
data data5;
set data4;
array response(30) ;
do i=1 to 30;
response(i)= substr (allresponses,i,1);
end;
run;


*6. SORTS CASES AND DELETES UNECESSARY DATASETS;
data data6 (drop=  thetagroup length);
set data5;
case=_n_;
proc sort; by case;
run;

          proc datasets;
          delete dat1 data2 data4 data5;
          run;


*8. MERGES THE FILES;
data data8;
merge data6 data3;
by case;
proc sort; by case;
run;

          proc datasets;
          delete data3 data6;
          run;
```

```
*9. CREATES A VARIABLE THAT ORDERS THE ITEMS THAT WERE PRESENTED
TO EACH EXAMINEE;
data data9;
set data8;
itemord1=0; itemord2=0; itemord3=0; itemord4=0; itemord5=0;
itemord6=0; itemord7=0; itemord8=0; itemord9=0; itemord10=0;
itemord11=0; itemord12=0; itemord13=0; itemord14=0; itemord15=0;
itemord16=0; itemord17=0; itemord18=0; itemord19=0; itemord20=0;
itemord21=0; itemord22=0; itemord23=0; itemord24=0; itemord25=0;
itemord26=0; itemord27=0; itemord28=0; itemord29=0; itemord30=0;
if item1>0 then itemord1=1; if item2>0 then itemord2=2;
if item3>0 then itemord3=3; if item4>0 then itemord4=4;
if item5>0 then itemord5=5; if item6>0 then itemord6=6;
if item7>0 then itemord7=7; if item8>0 then itemord8=8;
if item9>0 then itemord9=9; if item10>0 then itemord10=10;
if item11>0 then itemord11=11; if item12>0 then itemord12=12;
if item13>0 then itemord13=13; if item14>0 then itemord14=14;
if item15>0 then itemord15=15; if item16>0 then itemord16=16;
if item17>0 then itemord17=17; if item18>0 then itemord18=18;
if item19>0 then itemord19=19; if item20>0 then itemord20=20;
if item21>0 then itemord21=21; if item22>0 then itemord22=22;
if item23>0 then itemord23=23; if item24>0 then itemord24=24;
if item25>0 then itemord25=25; if item26>0 then itemord26=26;
if item27>0 then itemord27=27; if item28>0 then itemord28=28;
if item29>0 then itemord29=29; if item30>0 then itemord30=30;
order=itemord1; output; order=itemord2; output;
order=itemord3; output; order=itemord4; output;
order=itemord5; output; order=itemord6; output;
order=itemord7; output; order=itemord8; output;
order=itemord9; output; order=itemord10; output;
order=itemord11; output; order=itemord12; output;
order=itemord13; output; order=itemord14; output;
order=itemord15; output; order=itemord16; output;
order=itemord17; output; order=itemord18; output;
order=itemord19; output; order=itemord20; output;
order=itemord21; output; order=itemord22; output;
order=itemord23; output; order=itemord24; output;
order=itemord25; output; order=itemord26; output;
order=itemord27; output; order=itemord28; output;
order=itemord29; output; order=itemord30; output;
proc sort; by case;
run;
```

```
*10. CREATES ONE VARIABLE THAT CONTAINS THE SPECIFIC ITEMS THAT
WERE ;
*ADMINISTERED TO EACH EXAMINEE;
data data10;
set data8;
item=item1; output; item=item2; output;
item=item3; output; item=item4; output;
item=item5; output; item=item6; output;
item=item7; output; item=item8; output;
item=item9 ; output; item=item10; output;
item=item11; output; item=item12; output;
item=item13; output; item=item14; output;
item=item15; output; item=item16; output;
item=item17; output; item=item18; output;
item=item19; output; item=item20; output;
item=item21; output; item=item22; output;
item=item23; output; item=item24; output;
item=item25; output; item=item26; output;
item=item27; output; item=item28; output;
item=item29; output; item=item30; output;
run; proc sort;  by case; run;


*11. PUTS ALL OF THE ITEM RESPONSES FOR EACH EXAMINEE IN ONE
VARIABLE;
data data11;
set data8;
itresponse=response1; output; itresponse=response2; output;
itresponse=response3; output; itresponse=response4; output;
itresponse=response5; output; itresponse=response6; output;
itresponse=response7; output; itresponse=response8; output;
itresponse=response9; output; itresponse=response10; output;
itresponse=response11; output; itresponse=response12; output;
itresponse=response13; output; itresponse=response14; output;
itresponse=response15; output; itresponse=response16; output;
itresponse=response17; output; itresponse=response18; output;
itresponse=response19; output; itresponse=response20; output;
itresponse=response21; output; itresponse=response22; output;
itresponse=response23; output; itresponse=response24; output;
itresponse=response25; output; itresponse=response26; output;
itresponse=response27; output; itresponse=response28; output;
itresponse=response29; output; itresponse=response30; output;
run;
proc sort; by case;
run;
```

```
*12. PUTS ALL OF THE SEM VARIABLES IN ONE VARIABLE FOR EACH
EXAMINEE;
data data12;
set data8;
sem=sem1; output; sem=sem2; output; sem=sem3; output; sem=sem4;
output;
sem=sem5; output; sem=sem6; output; sem=sem7; output; sem=sem8;
output;
sem=sem9; output; sem=sem10; output;
sem=sem11; output; sem=sem12; output; sem=sem13; output;
sem=sem14; output;
sem=sem15; output; sem=sem16; output; sem=sem17; output;
sem=sem18; output;
sem=sem19; output; sem=sem20; output;
sem=sem21; output; sem=sem22; output; sem=sem23; output;
sem=sem24; output;
sem=sem25; output; sem=sem26; output; sem=sem27; output;
sem=sem28; output;
sem=sem29; output; sem=sem30; output;
proc sort; by case;
run;


*13. MERGES ALL OF THE DATA FILES TOGETHER;
data data13;
merge data12 data11 data10 data9;
by case;
proc sort; by item;
run;

          proc datasets;
          delete data9 data10 data11 data12;
          run;


*14. DELETES THE ITEMS THAT WERE NOT ADMINISTERED;
data dat.data14items250;
merge data1 data13 ;
by item;
IF theta=. then delete;
proc sort; by case;
run;
```

```
*15. DROPS A LOT OF THE UNECCESSARY VARIABLES;
data data15 (drop=item1 item2 item3 item4 item5 item6 item7 item8
item9 item10 item11 item12 item13 item14 item15 item16 item17
item18  item19 item20 item21 item22 item23 item24 item25 item26
item27  item28 item29 item30
itemord1 itemord2 itemord3 itemord4 itemord5 itemord6 itemord7
itemord8 itemord9 itemord10 itemord11 itemord12 itemord13
itemord14 itemord15 itemord16 itemord17 itemord18  itemord19
itemord20 itemord21 itemord22 itemord23 itemord24 itemord25
itemord26 itemord27  itemord28 itemord29 itemord30
response1 response2 response3 response4 response5
response6 response7 response8 response9 response10
response11 response12 response13 response14 response15
response16 response17 response18 response19 response20
response21 response22 response23 response24 response25
response26 response27 response28 response29 response30
sem1 sem2 sem3 sem4 sem5 sem6 sem7 sem8 sem9 sem10
sem11 sem12 sem13 sem14 sem15 sem16 sem17 sem18 sem19 sem20
sem21 sem22 sem23 sem24 sem25 sem26 sem27 sem28 sem29 sem30
response1 response2 response3 response4 response5 response6
response7 response8 response9 response10
response11 response12 response13 response14 response15 response16
response17 response18 response19 response20
response21 response22 response23 response24 response25 response26
response27 response28 response29 response30);

set dat.data14items250;
run;
```

```
*16. CREATES THE P-VALUES AND FLAGS THE ITEMS THAT MIGHT BE
REVIEWED;
*guess=1 indicates that the item was answered correctly just by
chance;
*guess=2 indicates that the item was answered incorrectly because
of a 'stupid' mistake;

data data16;
set data15;
by case;

L=a*(theta-b);
      p=c + (1-c)*(exp(L))/(1+exp(L));
      review=0;
      count=0;
      change=0;
      guess=0;
if p<0.33 and itresponse=1 then guess=1;
if p>0.8 and itresponse=0 then guess=2;
if first.case then count=0;
if first.case then totcount=0;
      if (p<0.53 and p>0.47) or guess=1 or guess=2 then do;
      review=1 ;
      totcount+1;
      count=totcount;
end;
run;



*17. Sorts data and drops unnecessary datasets;
data data17(drop=i);
set data16;
proc sort;
by case order;
run;

           proc datasets;
           delete data1 data15 data16;
           run;
```

```
*18. CHANGES THE FORMAT OF THE DATA SO THAT EACH EXAMINEE IS ON
ONE LINE;

data data18;
retain case a1-a30 b1-b30 c1-c30 order1-order30
p1-p30 L1-L30 item1-item30 bias finscore finalsem
itresponse1-itresponse30  review1-review30
change1-change30 sem1-sem30 thetahat theta
count totcount change guess guess1-guess30;

array items[30] item1-item30;
array aas[30] a1-a30;
array bs[30] b1-b30;
array cs[30] c1-c30;
array ps[30] p1-p30;
array Ls[30] L1-L30;
array orders[30] order1-order30;
array sems[30] sem1-sem30;
array itresponses[30] itresponse1-itresponse30;
*the reviews variables identify the items that are eligible for
review;
array reviews[30] review1-review30;
*describes the patterns of changes;
array changes[30] change1-change30;
array Iis(30) Ii1-Ii30;
array guesses(30) guess1-guess30;

set data17; by case;
     if first.case then do;
     i=1;
          do j=1 to 30;
          items(j)=. ;
          aas(j)=. ;
          bs(j)=. ;
          cs(j)=. ;
          ps(j)=. ;
          Ls(j)=. ;
          orders(j)=. ;
          sems(j)=. ;
          itresponses(j)=. ;
          reviews(j)=.;
          changes(j)=.;
          guesses(j)=.;
          end;
     end;
```

```
*now I include the old  variables;
items(i)=item;
aas(i)=a;
bs(i)=b;
cs(i)=c;
ps(i)=p;
Ls(i)=L;
orders(i)=order;
sems(i)=sem;
itresponses(i)=itresponse;
reviews(i)=review;
changes(i)=change;
guesses(i)=guess;

if last.case then output;
i+1;
run;
```

```
      ********************************************************;
      *PART 2. BEFORE REVIEW;
      ********************************************************;

      *19. CREATES THE OWENS BAYESIAN ESTIMATES BEFORE REVIEW;
      data data19;
      set data18;
      array aas[30] a1-a30;
      array bs[30]  b1-b30;
      array cs[30]  c1-c30;
      array ds[30]  d1-d30;
      array ls[30]  l1-l30;
      array nds[30] nd1-nd30;
      array epriors(30) eprior1-eprior30;
      array vpriors(30) vprior1-vprior30;
      array an2s[30] an2s1-an2s30;
      array phismallD(30) phismallD1-phismallD30;
      array phibigD(30) phibigD1-phibigD30;
      array phibigND(30) phibigND1-phibignd30;
      array alpha(30) alpha1-alpha30;
      array varpart(30)  varpart1-varpart30;
      array varpartE(30) varpartE1-varpartE30;
      array itresponses[30] itresponse1-itresponse30;
      array guesses(30) guess1-guess30;
      array ps(30) p1-p30;
      array Is(30) I1-I30;

     biasbay1=0;
     errorbay1=0;
     thetabays1=0;

    if a1=0 then a1=0.1;

    I1=(a1**2)*(1-c1)/((c1+exp(L1))*(1+exp(-L1))**2);
    an2s1=a1**(-2);
    D1=(b1/(sqrt(an2s1+1)));
    nD1=-D1;
    eprior1=-1;
    vprior1=1;
   phismallD1=PDF('normal',d1);
  phibigD1=cdf('normal', d1);
 phibigND1=cdf('normal',nD1);

  Alpha1=c1+((1-c1)*phibigND1);
  varpart1=(1+(an2s1)  )**(-1);
  varpartE1=1/(sqrt(an2s1+1));

        if itresponse1=1 then do;
        vprior1=(1-(1-c1)*varpart1*phismallD1*((((1-
c1)*phismallD1/Alpha1) -D1)/Alpha1));
```

```
      eprior1= eprior1+ ((1-c1)*varpartE1*phismallD1)/(c1+(1-
c1)*phibigND1);
      end;

            if itresponse1=0 then do;
            vprior1=1-
varpart1*phismallD1*((phismallD1/phibigD1)+D1)/phibigD1 ;
            eprior1= eprior1 - (
varpartE1*phismallD1)/(phibigD1);
            end;

do i=2 to 30;
j=i-1;
if aas(i)=0 then aas(i)=0.1;
an2s(i)=((aas(i))**(-2));
Is(i)=((aas(i))**2)*(1-cs(i))/((cs(i)+exp(Ls(i)))*(1+exp(-
Ls(i)))**2);

IF  itresponses(i)=1 THEN DO;
      Ds(i)=((bs(i)-epriors(j))/(sqrt(an2s(i)+vpriors(j))));
      nDs(i)=-Ds(i);
      phismallD(i)=PDF('normal',Ds(i));
      phibigD(i)=  CDF('normal',Ds(i));
      phibigND(i)= CDF('normal',nDs(i));

      Alpha(i)=cs(i)+(1-cs(i))*phibigND(i);
      varpart(i)=(1+(an2s(i)*vpriors(j)**(-1)))**(-1);
      varpartE(i)=1/(sqrt(an2s(i)+vpriors(j)));
      vpriors(i)=vpriors(j)*(1-(1-
cs(i))*varpart(i)*phismallD(i)*((((1-
cs(i))*phismallD(i)/Alpha(i)) )));
      epriors(i)= epriors(j)+ ((1-
cs(i))*vpriors(j)*varpartE(i)*phismallD(i))/(cs(i)+(1-
cs(i))*phibigND(i));
END;

else if itresponses(i)=0 then do;
      Ds(i)=((bs(i)-epriors(j))/(sqrt(an2s(i)+vpriors(j))));
      nDs(i)=-Ds(i);
      phismallD(i)=PDF('normal',Ds(i));
      phibigD(i)=cdf('normal', Ds(i));
      phibigND(i)=cdf('normal',nDs(i));

      Alpha(i)=cs(i)+(1-cs(i))*phibigND(i);
      varpart(i)=(1+(an2s(i)*vpriors(j)**(-1)))**(-1);
      varpartE(i)=1/(sqrt(an2s(i)+vpriors(j)));
      vpriors(i)=vpriors(j)* (1-
varpart(i)*phismallD(i)*((phismallD(i)/phibigD(i))+Ds(i))/phibigD
(i)) ;
      epriors(i)= epriors(j)- (
vpriors(j)*varpartE(i)*phismallD(i))/(phibigD(i));
end;
```

```
end;

thetabays1=eprior30;
biasbay1=thetabays1-theta;
errorbay1=abs(thetabays1-theta);
RUN;

*19B. MAXUMUM LIKELIHOOD ESTIMATION BEFORE ANY CHANGES ARE MADE
BIAS1;
data dat.data19b250;
set data19;
array aas[30] a1-a30;
array bs[30] b1-b30;
array cs[30] c1-c30;
array ls[30] l1-l30;
array phs[30] ph1-ph30;
array pts[30] pt1-pt30;
array ws[30] w1-w30;
array vs[30] v1-v30;
array psps[30] psp1-psp30;
array itresponses[30] itresponse1-itresponse30;
array guesses(30) guess1-guess30;

mltheta=thetabays1;
sumn=0;
sumd=0;
delta=0;
bigt=.5;
temp=0;
biasml1=0; errorml1=0;

*k is the number of iterations;

      do i=1 to 30;
      L=aas(i)*(mltheta-bs(i));
      phs(i)=(   1+exp(-L)   )**(-1);
      pts(i)=cs(i)+ (1-cs(i))*phs(i);
            if (pts(i)<.00001) then pts(i)=.00001;
            if (pts(i)>.99999) then pts(i)=.99999;
      ws(i)=pts(i)*(1-pts(i));
      vs(i)=itresponses(i)-pts(i);
      psps(i)=phs(i)/pts(i);
      sumn=sumn+ aas(i)*vs(i)*psps(i);
      sumd=sumd+(aas(i)**2)*ws(i)*psps(i)*psps(i);
      if sumd<.0001 then sumd=.0001;
      delta=sumn/sumd;
      end;

      if abs(delta)=<bigt then do;
            mltheta=mltheta+delta;
            end;
      if abs(delta)>bigt and delta>0.0 then do;
```

```
            delta=bigt;
            mltheta=mltheta+delta;
            end;
      if abs(delta)>bigt and delta=<0.0 then do;
            delta=-bigt;
            mltheta=mltheta+delta ;
            end;

do k=1 to 100;
IF abs(delta)>0.0001  THEN DO;
sumn=0; sumd=0;

      do i=1 to 30;
      L=aas(i)*(mltheta-bs(i));
      phs(i)=(   1+exp(-L)  )**(-1);
      pts(i)=cs(i)+ (1-cs(i))*phs(i);
            if (pts(i)<.00001) then pts(i)=.00001;
            if (pts(i)>.99999) then pts(i)=.99999;
      ws(i)=pts(i)*(1-pts(i));
      vs(i)=itresponses(i)-pts(i);
      psps(i)=phs(i)/pts(i);
      sumn=sumn+ aas(i)*vs(i)*psps(i);
      sumd=sumd+(aas(i)**2)*ws(i)*psps(i)*psps(i);
      if sumd<.0001 then sumd=.0001;
      delta=sumn/sumd;
      if i<2 then mltheta=mltheta-.15;
      end;

      if abs(delta)=<bigt then do;
            mltheta=mltheta+delta;
            end;
      if abs(delta)>bigt and delta>0.0 then do;
            delta=bigt;
            mltheta=mltheta+delta;
            end;
      if abs(delta)>bigt and delta=<0.0 then do;
            delta=-bigt;
            mltheta=mltheta+delta ;
            end;

END;
end;
if mltheta>3.3 then mltheta=3.3;
if mltheta<-3.3 then mltheta=-3.3;

biasml1=mltheta-theta;
errorml1=abs(mltheta- theta);
run;

            proc datasets;
            delete data17 data18 data19;
            run;
```

```
*20. ALLOWS ONLY UP TO 5 ANSWERS TO BE CHANGED;
*chose is the variable that shows the number of answers changed;
data data20;
set dat.data19b250;
maxcount=5;
chose=0;
array reviews[30] review1-review30;
array guesses(30) guess1-guess30;

      DO G=1 TO 30;
            IF REVIEWS(G)=1 AND GUESSES(G)=1 AND chose<5 THEN DO;
            CHOSE=CHOSE+1;
            reviews[G]=2;
            END;
            IF REVIEWS(G)=1 AND GUESSES(G)=2 AND chose<5 THEN DO;
            CHOSE=CHOSE+1;
            reviews[G]=2;
            END;
      END;
      totcount=totcount-chose;
      maxcount=maxcount-chose;

do i=1 to 30;
    random1=ranuni(0);
      *we do this so that we have no division by 0;
      if totcount>0 and maxcount>0 then  do;
            if random1= <(maxcount/totcount)then do;
                  if chose<5 then do;
                        if reviews(i)=1 then do;
                        chose=chose+1;
                        reviews[i]=2;
                        j=i-1;
                        totcount=totcount-chose;
                        maxcount=maxcount-chose;
                        end;
                  end;
            end;
      end;
do j= 2 to 30;
      random2=ranuni(0);
      if reviews(i)=1 then do; *(if the previous variable was
      changed then do;
            *we do this so to make sure that we have no more than
      5 revisions;
            if chose<5 then do;
                  *we do this so that we have no division by 0;
                  if totcount>0 and maxcount>0 then do;
                        if random2=<(maxcount/totcount) and
                        reviews(j)=1 and chose<5 then do;
                        chose=chose+1;
                        reviews(j)=2;
```

```
                    totcount=totcount-chose;
                    maxcount=maxcount-chose;
                    end;
                end;
            end;
        end;
end;
end;
run;
```

```
*****************************************************************;
*PART 3. REVIEW PROCEDURE AND ABILITY ESTIMATION AFTER REVIEW;
*****************************************************************;

*21. HERE I ACTUALLY CHANGE THE ANSWERS WHERE REVIEW=2;
data data21;
set data20;
array reviews[30] review1-review30;
array ps[30] p1-p30;
array itresponses[30] itresponse1-itresponse30;
array changes[30] change1-change30;
array aas[30] a1-a30;
array bs[30] b1-b30;
array cs[30] c1-c30;
array ds[30] d1-d30;
array ls[30] l1-l30;
array nds[30] nd1-nd30;
array epriors(30) eprior1-eprior30;
array vpriors(30) vprior1-vprior30;
array an2s[30] an2s1-an2s30;
array phismallD(30) phismallD1-phismallD30;
array phibigD(30) phibigD1-phibigD30;
array phibigND(30) phibigND1-phibignd30;
array alpha(30) alpha1-alpha30;
array  varpart(30)  varpart1-varpart30;
array  varpartE(30)  varpartE1-varpartE30;
array guesses(30) guess1-guess30;
*array Is(30) I1-I30;
array Iis(30) Ii1-Ii30;

changes1=0;
changes2=0;
changes3=0;
changes4=0;
changes5=0;
changes6=0;
test1=0;
test2=0;
do i=1 to 30;
Iis(i)=((aas(i))**2)*(1-cs(i))/((cs(i)+exp(Ls(i)))*(1+exp(-
Ls(i)))**2);

*RANDOM4 IS USED TO RANDOMLY DETERMINE IF THE QUESTION WILL BE
ANSWERED CORRECTLY BASED ON 72%;
*RANDOM5 IS USED TO RANDOMLY DETERMINE IF THE QUESTION WILL BE
ANSWERED INCORRECTLY BASED ON 28%;
*11=1, 01=2, 01=6 00=3, 10=4 10=5;
*test2=1 means that only changes of 10 or 01 were made;
*test1=1 means that any changes of 10, 01, 00 or even 11 were
made;
```

```
      random4=ranbin(0,1, 0.72);
            *if p=1 (72%) then W--> R and change=2;
            *if p=0 (28%) then W--> W and change=3;
      random5=ranbin(0,1,0.28);
            *if p=1 (21%) then R--> W and change=4;
            *if p=0 (79%) then R--> R and change=1;
      if reviews(i)=2 then do;
                  if (itresponses(i)=1 and random5=1) then do;
                  itresponses(i)=0 ; changes(i)=4;
changes4=changes4+1;
                  test2=1;
                  test1=1;
                  end;
                        if (itresponses(i)=1 and random5=0) then
do;
                        itresponses(i)=1 ; changes(i)=1;
changes1=changes1+1;
                        test2=0;
                        test1=1;
                        end;
                              if (itresponses(i)=0 and random4=1)
then do;
                              itresponses(i)=1 ; changes(i)=2;
changes2=changes2+1;
                              test2=1;
                              test1=1;
                              end;
                                    if ( itresponses(i)=0 and
random4=0) then do;
                                    itresponses(i)=0 ; changes(i)=3;
changes3=changes3+1;
                                    test2=0;
                                    test1=1;
                                    end;

      end;

      IF REVIEWS(I)=2 AND GUESSES(I)=1 THEN DO;
      ITRESPONSES(I)=0;
      CHANGES(I)=5;
      CHANGES5=CHANGES5+1;
      TEST2=1;
      test1=1;
      END;
      IF REVIEWS(I)=2 AND GUESSES(I)=2 THEN DO;
      ITRESPONSES(I)=1;
      CHANGES(I)=6;
      CHANGES6=CHANGES6+1;
      TEST2=1;
      test1=1;
      END;
 end;
```

```
totinfo1=sum(ii1-ii30;
se1=1/(sqrt(totinfo1));
run;

 *22. OWENS BAYESIAN ESTIMATION AFTER REVIEW;
data data22;
set data21;
array aas[30] a1-a30;
array bs[30] b1-b30;
array cs[30] c1-c30;
array ds[30] d1-d30;
array ls[30] l1-l30;
array nds[30] nd1-nd30;
array epriors(30) eprior1-eprior30;
array vpriors(30) vprior1-vprior30;
array an2s[30] an2s1-an2s30;
array phismallD(30) phismallD1-phismallD30;
array phibigD(30) phibigD1-phibigD30;
array phibigND(30) phibigND1-phibignd30;
array alpha(30) alpha1-alpha30;
array  varpart(30)  varpart1-varpart30;
array  varpartE(30)  varpartE1-varpartE30;
array itresponses[30] itresponse1-itresponse30;
array guesses(30) guess1-guess30;
*array Is(30) I1-I30;
biasbay2=0;
errorbay2=0;
thetabays2=0;
if a1=0 then a1=0.1;

*I1=(a1**2)*(1-c1)/((c1+exp(L1))*(1+exp(-L1))**2);
an2s1=a1**(-2);
D1=(b1/(sqrt(an2s1+1)));
nD1=-D1;
eprior1=0;
vprior1=1;
phismallD1=PDF('normal',d1);
phibigD1=cdf('normal', d1);
phibigND1=cdf('normal',nD1);

 Alpha1=c1+((1-c1)*phibigND1);
 varpart1=(1+(an2s1)  )**(-1);
 varpartE1=1/(sqrt(an2s1+1));

      if itresponse1=1 then do;
      vprior1=(1-(1-c1)*varpart1*phismallD1*((((1-
c1)*phismallD1/Alpha1) -D1)/Alpha1));
      eprior1=  ((1-c1)*varpartE1*phismallD1)/(c1+(1-
c1)*phibigND1);
      end;
```

```
          if itresponse1=0 then do;
          vprior1=1-
varpart1*phismallD1*((phismallD1/phibigD1)+D1)/phibigD1 ;
          eprior1= - (      varpartE1*phismallD1)/(phibigD1);
          end;

do i=2 to 30;
j=i-1;
if aas(i)=0 then aas(i)=0.1;
an2s(i)=((aas(i))**(-2));
*Is(i)=((aas(i))**2)*(1-cs(i))/((cs(i)+exp(Ls(i)))*(1+exp(-
Ls(i)))**2);

if itresponses(i)=1 then do;
     Ds(i)=((bs(i)-epriors(j))/(sqrt(an2s(i)+vpriors(j))));
     nDs(i)=-Ds(i);
     phismallD(i)=PDF('normal',Ds(i));
     phibigD(i)=  CDF('normal',Ds(i));
     phibigND(i)= CDF('normal',nDs(i));

     Alpha(i)=cs(i)+(1-cs(i))*phibigND(i);
     varpart(i)=(1+(an2s(i)*vpriors(j)**(-1)))**(-1);
     varpartE(i)=1/(sqrt(an2s(i)+vpriors(j)));
     vpriors(i)=vpriors(j)*(1-(1-
cs(i))*varpart(i)*phismallD(i)*((((1-
cs(i))*phismallD(i)/Alpha(i)) )));
     epriors(i)= epriors(j)+ ((1-
cs(i))*vpriors(j)*varpartE(i)*phismallD(i))/(cs(i)+(1-
cs(i))*phibigND(i));
end;

else if itresponses(i)=0 then do;
     Ds(i)=((bs(i)-epriors(j))/(sqrt(an2s(i)+vpriors(j))));
     nDs(i)=-Ds(i);
     phismallD(i)=PDF('normal',Ds(i));
     phibigD(i)=cdf('normal', Ds(i));
     phibigND(i)=cdf('normal',nDs(i));

     Alpha(i)=cs(i)+(1-cs(i))*phibigND(i);
     varpart(i)=(1+(an2s(i)*vpriors(j)**(-1)))**(-1);
     varpartE(i)=1/(sqrt(an2s(i)+vpriors(j)));
     vpriors(i)=vpriors(j)* (1-
varpart(i)*phismallD(i)*((phismallD(i)/phibigD(i))+Ds(i))/phibigD
(i)) ;
     epriors(i)= epriors(j)- (
vpriors(j)*varpartE(i)*phismallD(i))/(phibigD(i));
end;
end;

*totinfo2=sum(i1-i30);
*se2=1/(sqrt(totinfo2));
thetabays2=eprior30;
```

```
biasbay2=thetabays2-theta;
errorbay2=abs(thetabays2-theta);
run;

          proc datasets;
          delete data20 data21;
          run;

*23. MAXIMUM  LIKELIHOOD ESTIMATION AFTER REVIEW;
data data23;
set data22;
array aas[30] a1-a30;
array bs[30] b1-b30;
array cs[30] c1-c30;
array ls[30] l1-l30;
array phs[30] ph1-ph30;
array pts[30] pt1-pt30;
array ws[30] w1-w30;
array vs[30] v1-v30;
array psps[30] psp1-psp30;
array itresponses[30] itresponse1-itresponse30;
array guesses(30) guess1-guess30;
array changes(30) change1-change30;

mltheta2=thetabays2;
sumn=0;
sumd=0;
delta=0;
bigt=.5;
temp=0;
biasml2=0;
errorml2=0;
*test2=0;
*do the same for all variables s1-s30;
*k is the number of iterations;

     do i=1 to 30;
     L=aas(i)*(mltheta2-bs(i));
     phs(i)=(   1+exp(-L)  )**(-1);
     pts(i)=cs(i)+ (1-cs(i))*phs(i);
          if (pts(i)<.00001) then pts(i)=.00001;
          if (pts(i)>.99999) then pts(i)=.99999;
     ws(i)=pts(i)*(1-pts(i));
     vs(i)=itresponses(i)-pts(i);
     psps(i)=phs(i)/pts(i);
     sumn=sumn+ aas(i)*vs(i)*psps(i);
     sumd=sumd+(aas(i)**2)*ws(i)*psps(i)*psps(i);
     if sumd<.0001 then sumd=.0001;
     delta=sumn/sumd;
     if changes(i)=2 or changes(i)=4 then test2=1;
     end;
```

```
      if abs(delta)=<bigt then do;
            mltheta2=mltheta2+delta;
            end;
      if abs(delta)>bigt and delta>0.0 then do;
            delta=bigt;
            mltheta2=mltheta2+delta;
            end;
      if abs(delta)>bigt and delta=<0.0 then do;
            delta=-bigt;
            mltheta2=mltheta2+delta ;
            end;

do k=1 to 100;
if abs(delta)>0.0001  then do;
sumn=0; sumd=0;

      do i=1 to 30;
      L=aas(i)*(mltheta2-bs(i));
      phs(i)=(   1+exp(-L)  )**(-1);
      pts(i)=cs(i)+ (1-cs(i))*phs(i);
            if (pts(i)<.00001) then pts(i)=.00001;
            if (pts(i)>.99999) then pts(i)=.99999;
      ws(i)=pts(i)*(1-pts(i));
      vs(i)=itresponses(i)-pts(i);
      psps(i)=phs(i)/pts(i);
      sumn=sumn+ aas(i)*vs(i)*psps(i);
      sumd=sumd+(aas(i)**2)*ws(i)*psps(i)*psps(i);
      if sumd<.0001 then sumd=.0001;
      delta=sumn/sumd;
      end;

      if abs(delta)=<bigt then do;
            mltheta2=mltheta2+delta;
            end;
      if abs(delta)>bigt and delta>0.0 then do;
            delta=bigt;
            mltheta2=mltheta2+delta;
            end;
      if abs(delta)>bigt and delta=<0.0 then do;
            delta=-bigt;
            mltheta2=mltheta2+delta;
            end;

end; end;
if mltheta2>3 then mltheta2=3;
if mltheta2<-3 then mltheta2=-3;

biasml2=mltheta2- theta;
errorml2=abs(mltheta2-theta);
run;
```

```
*************************************************************;
*PART 4. REARRANGEMENT PROCEDURE;
*************************************************************;

data data24;
set   data23;
array aas[30] a1-a30;
array bs[30] b1-b30;
array cs[30] c1-c30;
array ds[30] d1-d30;
array ls[30] l1-l30;
array nds[30] nd1-nd30;
array epriors(30) eprior1-eprior30;
array vpriors(30) vprior1-vprior30;
array an2s[30] an2s1-an2s30;
array phismallD(30) phismallD1-phismallD30;
array phibigD(30) phibigD1-phibigD30;
array phibigND(30) phibigND1-phibignd30;
array alpha(30) alpha1-alpha30;
array varpart(30)  varpart1-varpart30;
array varpartE(30)  varpartE1-varpartE30;
array itresponses[30] itresponse1-itresponse30;
array changes(30) change1-change30;
array Is(30) I1-I30;
array skipped(30) skipped1-skipped30;
array Iis(30) Ii1-Ii30;
array diff(30) diff1-diff30;
array bdiff(30) bdiff1-bdiff30;
*array ignores(30) ignore1-ignore30;
f=0; z=0;
do ii=1 to 30;
bdiff(ii)=0;
end;

*part 1;
*ignore is the variable that shows how many items have been left
behind;
ignore=0;
skipped1=0; skipped2=0; skipped3=0; skipped4=0; skipped5=0;
skipped6=0; skipped7=0; skipped8=0;
skipped9=0; skipped10=0;
skipped11=0; skipped12=0; skipped13=0; skipped14=0; skipped15=0;
skipped16=0; skipped17=0; skipped18=0;
skipped19=0; skipped20=0;
skipped21=0; skipped22=0; skipped23=0; skipped24=0; skipped25=0;
skipped26=0; skipped27=0; skipped28=0;
skipped29=0; skipped30=0;
totinfo3=0;
include=0;
flag=0;
switched=0;
```

```
erx
bia
the
epr
vpr
Il=
i=1
*TE
ans
*re

*pa
if
vpr
cl)
epr
inc
if
ski
ign
i=3
epr
vpr
end
end

*pa
if
vpr
;
epr
inc
if
ski
ign
i=3
epr
vpr
End
end
tot
dif

***
***
***
DO I

j=i-
over
if s
if i
```

```
errorml3=0;
biasml3=0;
thetabays3=0;
eprior1=0;
vprior1=1;
I1=(a1**2)*(1-c1)/((c1+exp(L1))*(1+exp(-L1))**2);
i=1;
*TEST is the variable that shows that this examinee changed their
answers;
*review are the qns that were eligible to be changed;

*part2a;
if itresponse1=1 then do;
vprior1=(1-(1-c1)*varpart1*phismallD1*((((1-
c1)*phismallD1/Alpha1) -D1)/Alpha1));
eprior1=  ((1-c1)*varpartE1*phismallD1)/(c1+(1-c1)*phibigND1);
include=include+1;
if change1=2 or change1=6 then do;
skipped2=1;
ignore=ignore+1;
i=3;
eprior2=eprior1;
vprior2=vprior1;
end; else i=2;
end;

*part2b;
if itresponse1=0 then do;
vprior1=1-varpart1*phismallD1*((phismallD1/phibigD1)+D1)/phibigD1
;
eprior1= - (      varpartE1*phismallD1)/(phibigD1);
include=include+1;
if change1=4 or changes1=5 then do;
skipped2=1;
ignore=ignore+1;
i=3;
eprior2=eprior1;
vprior2=vprior1;
End; else i=2;
end;
totinfo3=totinfo3+I1;
diff1=i1-ii1;

*****************************************************************;
*****************************************************************;
*****************************************************************;
DO I=I TO 30;

j=i-1;
over=0;
if skipped(i)=1 then Is(i)=0 ;
if i>30  or include>30 then leave;
```

```
*part3a;************************************************************
*we want to find the next correct answer to replace the previous
incorrect answer since;
*we had changes from a W to R 01;

    if (changes(i)=2 or changes(i)=6) and ignore<3 and i<28 then
do;
      Is(i)=((aas(i))**2)*(1-cs(i))/((cs(i)+exp(Ls(i)))*(1+exp(-
Ls(i)))**2);
    Ds(i)=((bs(i)-epriors(j))/(sqrt(an2s(i)+vpriors(j))));
    nDs(i)=-Ds(i);
    phismallD(i)=PDF('NORMAL',Ds(i));
    phibigD(i)=  CDF('NORMAL',Ds(i));
    phibigND(i)= CDF('NORMAL',nDs(i));
    Alpha(i)=cs(i)+(1-cs(i))*phibigND(i);
    varpart(i)=(1+(an2s(i)*vpriors(j)**(-1)))**(-1);
    varpartE(i)=1/(sqrt(an2s(i)+vpriors(j)));
    vpriors(i)=vpriors(j)*(1-(1-
cs(i))*varpart(i)*phismallD(i)*((((1-
cs(i))*phismallD(i)/Alpha(i)) )));
    epriors(i)= epriors(j)+ ((1-
cs(i))*vpriors(j)*varpartE(i)*phismallD(i))/(cs(i)+(1-
cs(i))*phibigND(i));
      include=include+1;
      totinfo3=totinfo3+Is(i);
    diff(i)=is(i)-iis(i);
      i=i+2;
      epriors(i-1)=epriors(i-2);
      vpriors(i-1)=vpriors(i-2);
      ignore=ignore+1;
      skipped(i-1)=1;
            if itresponses(i)=1 and ignore<3 and i<29 then do;
            *if ignore>3 or i>29 then leave;
            ignore=ignore+1;
            skipped(i-1)=1;
            i=i+1;
            epriors(i-1)=epriors(i-2);
            vpriors(i-1)=vpriors(i-2);
                  if ignore>3 then leave;
                  do while (itresponses(i)=1 and ignore<3 and
i<29);
                  ignore=ignore+1;
                  skipped(i-1)=1;
                  i=i+1;
                  epriors(i-1)=epriors(i-2);
                  vpriors(i-1)=vpriors(i-2);
                  end;
            end;

                  if ignore>1 then do;
                  i=i-1;
```

```
            F=i; * we do this to use F as an indicator later
on;
               DO Z=1 TO 30;
                     if skipped(Z)=1 and z<31 then do;
                     Is(Z)=((aas(Z))**2)*(1-
cs(Z))/((cs(Z)+exp(Ls(Z)))*(1+exp(-Ls(Z)))**2);
                     *this is the new info;
                     Is(f)=((aas(f))**2)*(1-
cs(f))/((cs(f)+exp(Ls(f)))*(1+exp(-Ls(f)))**2);
                                          if Is(z)>=Is(f) then do;
                                          skipped(f)=1;
                                    skipped(z)=0;
                                  F=z;
                                    end;
                                    if Is(z)<Is(f) then do;
                                          F=F;
                                    skipped(Z)=1;
                                    skipped(f)=0;
                                    end;
                  end;
                  END;
                     if (i ne f) then switched=1;
                  Ds(f)=((bs(f)-
epriors(j))/(sqrt(an2s(f)+vpriors(j))));
                  nDs(f)=-Ds(f);
                  phismallD(f)=PDF('NORMAL',Ds(f));
                  phibigD(f)=  CDF('NORMAL',Ds(f));
                  phibigND(f)= CDF('NORMAL',nDs(f));
                  Alpha(f)=cs(f)+(1-cs(f))*phibigND(f);
                  varpart(f)=(1+(an2s(f)*vpriors(j)**(-1)))**(-
1);
                  varpartE(f)=1/(sqrt(an2s(f)+vpriors(j)));
                     bdiff(i)=bs(f)-bs(i);

                     if itresponses(f)=1 then do;
                     vpriors(f)=vpriors(j)*(1-(1-
cs(f))*varpart(f)*phismallD(f)*((((1-
cs(f))*phismallD(f)/Alpha(f)) )));
                     epriors(f)= epriors(j)+ ((1-
cs(f))*vpriors(j)*varpartE(f)*phismallD(f))/(cs(f)+(1-
cs(f))*phibigND(f));
                     include=include+1;
                     end;

                     if itresponses(f)=0 then do;
                     vpriors(f)=vpriors(j)* (1-
varpart(f)*phismallD(f)*((phismallD(f)/phibigD(f))+Ds(f))/phibigD
(f)) ;
                     epriors(f)= epriors(j)- (
vpriors(j)*varpartE(f)*phismallD(f))/(phibigD(f));
                     include=include+1;
                     end;
```

```
                    Is(f)=((aas(f))**2)*(1-
cs(f))/((cs(f)+exp(Ls(f)))*(1+exp(-Ls(f)))**2);
                    totinfo3=totinfo3+Is(f);
                    diff(i)=is(f)-iis(i);
                         over=1;
             if i<30 then i=i+1;
             Ds(i)=((bs(i)-epriors(j))/(sqrt(an2s(i)+vpriors(j))));
           nDs(i)=-Ds(i);
           phismallD(i)=PDF('NORMAL',Ds(i));
           phibigD(i)=  CDF('NORMAL',Ds(i));
           phibigND(i)= CDF('NORMAL',nDs(i));
           Alpha(i)=cs(i)+(1-cs(i))*phibigND(i);
           varpart(i)=(1+(an2s(i)*vpriors(j)**(-1)))**(-1);
           varpartE(i)=1/(sqrt(an2s(i)+vpriors(j)));
                  if itresponses(i)=1  and include<31 then do;
                vpriors(i)=vpriors(f)*(1-(1-
cs(i))*varpart(i)*phismallD(i)*((((1-
cs(i))*phismallD(i)/Alpha(i)) )));
                epriors(i)= epriors(f)+ ((1-
cs(i))*vpriors(f)*varpartE(i)*phismallD(i))/(cs(i)+(1-
cs(i))*phibigND(i));
                include=include+1;
                  totinfo3=totinfo3+Is(i);
                  diff(i)=is(i)-iis(i);
                end;

                  if itresponses(i)=0   and include<31  then do;
                vpriors(i)=vpriors(f)* (1-
varpart(i)*phismallD(i)*((phismallD(i)/phibigD(i))+Ds(i))/phibigD
(i)) ;
                epriors(i)= epriors(f)- (
vpriors(f)*varpartE(i)*phismallD(i))/(phibigD(i));
                include=include+1;
                  totinfo3=totinfo3+Is(i);
                  diff(i)=is(i)-iis(i);
                end;
             over=1;
           end;
            *the over=1 variable shows that we are done with the
counting that i question;
    end;


*part3b;*****************************************************;
    *we want to find the next incorrect answer to replace the
previous correct answer since we had;
    *changes from a R to W;

    if (changes(i)=4 or changes(i)=5) and ignore<3 and i<28 then
do;
     Is(i)=((aas(i))**2)*(1-cs(i))/((cs(i)+exp(Ls(i)))*(1+exp(-
Ls(i)))**2);
```

```
    Ds(i)=((bs(i)-epriors(i-1))/(sqrt(an2s(i)+vpriors(i-1))));
    nDs(i)=-Ds(i);
    phismallD(i)=PDF('NORMAL',Ds(i));
    phibigD(i)=  CDF('NORMAL',Ds(i));
    phibigND(i)= CDF('NORMAL',nDs(i));
    Alpha(i)=cs(i)+(1-cs(i))*phibigND(i);
    varpart(i)=(1+(an2s(i)*vpriors(j)**(-1)))**(-1);
    varpartE(i)=1/(sqrt(an2s(i)+vpriors(j)));
    vpriors(i)=vpriors(j)* (1-
varpart(i)*phismallD(i)*((phismallD(i)/phibigD(i))+Ds(i))/phibigD
(i)) ;
    epriors(i)= epriors(j)- (
vpriors(j)*varpartE(i)*phismallD(i))/(phibigD(i));
     include=include+1;
     totinfo3=totinfo3+Is(i);
    diff(i)=is(i)-iis(i);
     i=i+2;
     epriors(i-1)=epriors(i-2);
     vpriors(i-1)=vpriors(i-2);
     ignore=ignore+1;
     skipped(i-1)=1;
          if itresponses(i)=0 and ignore<3 and i<29 then do;
          ignore=ignore+1;
          skipped(i-1)=1;
          i=i+1;
          epriors(i-1)=epriors(i-2);
          vpriors(i-1)=vpriors(i-2);
               do while (itresponses(i)=0 and ignore<3 and
i<29);
               ignore=ignore+1;
               skipped(i-1)=1;
               i=i+1;
               epriors(i-1)=epriors(i-2);
               vpriors(i-1)=vpriors(i-2);
               end;
          end;

              if IGNORE>1 then do;
               i=i-1;
               F=i; * we do this to use F as an indicator later
on;
               DO Z=1 TO 30;
               if skipped(Z)=1 and z<31 then do;
               Is(Z)=((aas(Z))**2)*(1-
cs(Z))/((cs(Z)+exp(Ls(Z)))*(1+exp(-Ls(Z)))**2);
               *this is the new info;

               Is(f)=((aas(f))**2)*(1-
cs(f))/((cs(f)+exp(Ls(f)))*(1+exp(-Ls(f)))**2);
                              if Is(z)>=Is(f) then do;
                              skipped(f)=1;
                              skipped(z)=0;
```

```
                                F=z;
                                end;
                                if Is(z)<Is(f) then do;
                                 F=F;
                                skipped(Z)=1;
                                skipped(f)=0;
                                * f is the variable index with the
highest information;
                                  end;
                end;
               END;

                     if (i ne f) then switched=1;
                *the f variable is assigned to the item that
will be used next;
                Ds(f)=((bs(f)-
epriors(j))/(sqrt(an2s(f)+vpriors(j))));
                nDs(f)=-Ds(f);
                phismallD(f)=PDF('NORMAL',Ds(f));
                phibigD(f)=  CDF('NORMAL',Ds(f));
                phibigND(f)= CDF('NORMAL',nDs(f));
                Alpha(f)=cs(f)+(1-cs(f))*phibigND(f);
                varpart(f)=(1+(an2s(f)*vpriors(j)**(-1)))**(-
1);
                varpartE(f)=1/(sqrt(an2s(f)+vpriors(j)));
                    bdiff(i)=bs(f)-bs(i);

                if itresponses(f)=1 then do;
                vpriors(f)=vpriors(j)*(1-(1-
cs(f))*varpart(f)*phismallD(f)*((((1-
cs(f))*phismallD(f)/Alpha(f)) )));
                epriors(f)= epriors(j)+ ((1-
cs(f))*vpriors(j)*varpartE(f)*phismallD(f))/(cs(f)+(1-
cs(f))*phibigND(f));
                include=include+1;
                end;

                   if itresponses(f)=0 then do;
                   vpriors(f)=vpriors(j)* (1-
varpart(f)*phismallD(f)*((phismallD(f)/phibigD(f))+Ds(f))/phibigD
(f)) ;
                   epriors(f)= epriors(j)- (
vpriors(j)*varpartE(f)*phismallD(f))/(phibigD(f));
                   include=include+1;
                   end;

                Is(f)=((aas(f))**2)*(1-
cs(f))/((cs(f)+exp(Ls(f)))*(1+exp(-Ls(f)))**2);
                totinfo3=totinfo3+Is(f);
                diff(i)=is(f)-iis(i);
          over=1;
          if i<30 then i=i+1;
```

```
          Ds(i)=((bs(i)-epriors(j))/(sqrt(an2s(i)+vpriors(j))));
         nDs(i)=-Ds(i);
         phismallD(i)=PDF('NORMAL',Ds(i));
         phibigD(i)=  CDF('NORMAL',Ds(i));
         phibigND(i)= CDF('NORMAL',nDs(i));
         Alpha(i)=cs(i)+(1-cs(i))*phibigND(i);
         varpart(i)=(1+(an2s(i)*vpriors(j)**(-1)))**(-1);
         varpartE(i)=1/(sqrt(an2s(i)+vpriors(j)));

              if itresponses(i)=1  and include<31 then do;
            vpriors(i)=vpriors(f)*(1-(1-
cs(i))*varpart(i)*phismallD(i)*((((1-
cs(i))*phismallD(i)/Alpha(i)) )));
            epriors(i)= epriors(f)+ ((1-
cs(i))*vpriors(f)*varpartE(i)*phismallD(i))/(cs(i)+(1-
cs(i))*phibigND(i));
              include=include+1;
              totinfo3=totinfo3+Is(i);
              diff(i)=is(i)-iis(i);
            end;

              if itresponses(i)=0   and include<31 then do;
            vpriors(i)=vpriors(f)* (1-
varpart(i)*phismallD(i)*((phismallD(i)/phibigD(i))+Ds(i))/phibigD
(i)) ;
            epriors(i)= epriors(f)- (
vpriors(f)*varpartE(i)*phismallD(i))/(phibigD(i));
              include=include+1;
              totinfo3=totinfo3+Is(i);
              diff(i)=is(i)-iis(i);
            end;
          over=1;
          end;
    end;

*IN CASE A CHANGE WAS NOT MADE:;
     if over=0  then do;
    Is(i)=((aas(i))**2)*(1-cs(i))/((cs(i)+exp(Ls(i)))*(1+exp(-
Ls(i)))**2);
     Ds(i)=((bs(i)-epriors(i-1))/(sqrt(an2s(i)+vpriors(i-1))));
    nDs(i)=-Ds(i);
    phismallD(i)=PDF('NORMAL',Ds(i));
    phibigD(i)=  CDF('NORMAL',Ds(i));
    phibigND(i)= CDF('NORMAL',nDs(i));
    Alpha(i)=cs(i)+(1-cs(i))*phibigND(i);
    varpart(i)=(1+(an2s(i)*vpriors(j)**(-1)))**(-1);
    varpartE(i)=1/(sqrt(an2s(i)+vpriors(j)));

       if itresponses(i)=1  then do;
       vpriors(i)=vpriors(j)*(1-(1-
cs(i))*varpart(i)*phismallD(i)*((((1-
cs(i))*phismallD(i)/Alpha(i)) )));
```

```
        epriors(i)= epriors(j)+ ((1-
cs(i))*vpriors(j)*varpartE(i)*phismallD(i))/(cs(i)+(1-
cs(i))*phibigND(i));
           include=include+1;
           totinfo3=totinfo3+Is(i);
           diff(i)=is(i)-iis(i);
        end;

        if itresponses(i)=0  then do;
        vpriors(i)=vpriors(j)* (1-
varpart(i)*phismallD(i)*((phismallD(i)/phibigD(i))+Ds(i))/phibigD
(i)) ;
        epriors(i)= epriors(j)- (
vpriors(j)*varpartE(i)*phismallD(i))/(phibigD(i));
           include=include+1;
           totinfo3=totinfo3+Is(i);
           diff(i)=is(i)-iis(i);
        end;
     end;
     over=1;
END;

thetabays3=eprior30;
biasbay3=thetabays3 - theta;
biasbay3old=theta-eprior30;
errorbay3=abs(thetabays3 -theta);
se3=1/(sqrt(totinfo3));
total=ignore+include; if total>30 then flag=1;
run;
```

```
*25. MAXIMUM LIKELIHOOD AFTER REARRANGEMENT PROCEDURE;
data dat.data25ITEMS250;
set   data24;
array aas[30] a1-a30;
array bs[30] b1-b30;
array cs[30] c1-c30;
array ds[30] d1-d30;
array ls[30] l1-l30;
array nds[30] nd1-nd30;
array epriors(30) eprior1-eprior30;
array vpriors(30) vprior1-vprior30;
array an2s[30] an2s1-an2s30;
array phismallD(30) phismallD1-phismallD30;
array phibigD(30) phibigD1-phibigD30;
array phibigND(30) phibigND1-phibignd30;
array alpha(30) alpha1-alpha30;
array varpart(30)  varpart1-varpart30;
array varpartE(30)  varpartE1-varpartE30;
array itresponses[30] itresponse1-itresponse30;
array changes(30) change1-change30;
array Is(30) I1-I30;
array skipped(30) skipped1-skipped30;
array Iis(30) Ii1-Ii30;
array diff(30) diff1-diff30;
array ps[30] p1-p30;
*array l[30] l1-l30;
array phs[30] ph1-ph30;
array pts[30] pt1-pt30;
array ws[30] w1-w30;
array vs[30] v1-v30;
array psps[30] psp1-psp30;

mltheta3=thetabays3;
sumn=0;
sumd=0;
delta=0;
bigt=.5;
temp=0;
*k is the number of iterations;

      do i=1 to 30;
      if skipped(i)=1 then i=i+1;
      L=aas(i)*(mltheta3-bs(i));
      phs(i)=(   1+exp(-L)  )**(-1);
      pts(i)=cs(i)+ (1-cs(i))*phs(i);
            if (pts(i)<.00001) then pts(i)=.00001;
            if (pts(i)>.99999) then pts(i)=.99999;
      ws(i)=pts(i)*(1-pts(i));
      vs(i)=itresponses(i)-pts(i);
      psps(i)=phs(i)/pts(i);
      sumn=sumn+ aas(i)*vs(i)*psps(i);
```

```
      sumd=sumd+(aas(i)**2)*ws(i)*psps(i)*psps(i);
      if sumd<.0001 then sumd=.0001;
      delta=sumn/sumd;
      end;

      if abs(delta)=<bigt then do;
            mltheta3=mltheta3+delta;
            temp=1;
            end;
      if abs(delta)>bigt and delta>0.0 then do;
            delta=bigt;
            mltheta3=mltheta3+delta;
            temp=2;
            end;
      if abs(delta)>bigt and delta=<0.0 then do;
            delta=-bigt;
            mltheta3=mltheta3+delta ;
            temp=3;
            end;

do k=1 to 100;
if abs(delta)>0.0001 then do;
sumn=0; sumd=0;

      do i=1 to 30;
      if skipped(i)=1 then i=i+1;
      L=aas(i)*(mltheta3-bs(i));
      phs(i)=(   1+exp(-L)  )**(-1);
      pts(i)=cs(i)+ (1-cs(i))*phs(i);
            if (pts(i)<.00001) then pts(i)=.00001;
            if (pts(i)>.99999) then pts(i)=.99999;
      ws(i)=pts(i)*(1-pts(i));
      vs(i)=itresponses(i)-pts(i);
      psps(i)=phs(i)/pts(i);
      sumn=sumn+ aas(i)*vs(i)*psps(i);
      sumd=sumd+(aas(i)**2)*ws(i)*psps(i)*psps(i);
      if sumd<.0001 then sumd=.0001;
      delta=sumn/sumd;
      end;

      if abs(delta)=<bigt then do;
            mltheta3=mltheta3+delta;
            end;
      if abs(delta)>bigt and delta>0.0 then do;
            delta=bigt;
            mltheta3=mltheta3+delta;
            end;
      if abs(delta)>bigt and delta=<0.0 then do;
            delta=-bigt;
            mltheta3=mltheta3+delta ;
            end;;
```

```
if mltheta3>3.3 then mltheta3=3.3;
if mltheta3<-3.3 then mltheta3=-3.3;
biasml3=mltheta3-theta;
errorml3=abs(mltheta3-theta);
end;
end;
RUN;

          proc datasets;
          delete data24;
          run;


data dat.final250 (drop= a1-a30 b1-b30 c1-c30 order1-order30  L1-
L30
item item1-item30
 finscore finalsem  sem1-sem30 thetahat  phismallD1-phismallD30
phibigD1-phibigD30
phibignD1-phibignD30
 alpha1-alpha30 varpart1-varpart30 varparte1-varparte30  totcount
change vprior1-vprior30
d1-d30 l1-l30 an2s1-an2s30 nd1-nd30 flag p f z i j  g z count
total
error1 chose random1
random2 random4 random5  include over  review1-review30 w1-w30
v1-v30 ph ph1-ph30 pt
pt1-pt30 psp psp1-psp30
 sumn sumd  sums temp over delta bigt a b c guess guess1-guess30
1 allresponses delta ii1-ii30 diff1-diff30);
 set dat.data25ITEMS250;
run;
```

```
*ANALYSES RESULTS;
proc sort; by status;

data analys;
set dat.final250;
proc means;
title1 'OVERALL RESULTS 5 changes 250 items';
var biasml1-biasml3 biasbay1-biasbay3 bias errorml1-errorml3
errorbay1-errorbay3 theta thetabays1-thetabays3 mltheta mltheta2
mltheta3;
run;
proc means;
title1 'OVERALL RESULTS    BY STATUS ';
var biasml1-biasml3 biasbay1-biasbay3 bias errorml1-errorml3
errorbay1-errorbay3 theta thetabays1-thetabays3 mltheta mltheta2
mltheta3;
by  status;
run;

data analys;
set analys;
proc sort; by test2;
proc means;
title1 'OVERALL RESULTS by test2 250 ITEMS';
var biasml1-biasml3 biasbay1-biasbay3 bias errorml1-errorml3
errorbay1-errorbay3 theta thetabays1-thetabays3 mltheta mltheta2
mltheta3;
by test2;
run;

data analys;
set analys;
proc sort; by theta;
proc means;
title1 'OVERALL RESULTS by theta 250 ITEMS 5 reviews';
var biasml1-biasml3 biasbay1-biasbay3 bias errorml1-errorml3
errorbay1-errorbay3 theta thetabays1-thetabays3 mltheta mltheta2
mltheta3;
by theta;
run;

data decribe;
set analys;
     proc freq;
     title1 '250 5 changes ';
     Tables test2;
     run;
          proc sort; by test2;
          run;
          proc freq;
          title1 '250 5 changes ';
```

```
          Tables ignore; by test2;
          run;

          proc freq;
          title1 '250 5 changes ';
          where test2=1;
          Tables changes1-changes6;
          run;


data decribeinfo;
set analys;
array diffs[30] diff1-diff30;
where test2=1;
dn=0;
     do i=1 to 30;
     if diffs(i)= . then diffs(i)=0;
     if diffs(i)>0 then dn=dn+1;
     end;
infodiff=sum(diff1-diff30);
if dn>0 then infodiff1=infodiff/dn;
     proc means;
     title1 '250 5 changes ';
     var infodiff1;
     run;.
```

BIBLIOGRAPHY

BIBLIOGRAPHY

Ban, J. C., Wang, T., Yi, Q. & Harris, D. J. (2000). Effects of nonequivalence of item pools on ability estimates in CAT. Paper presented at the annual meeting of the National Council of Educational Measurement, April, New Orleans, LA.

Bowles, R. & Pommerich, M. (2001). An examination of item review on a CAT using the specific information item selection algorithm. Paper presented at the annual meeting of the National Council of Measurement in Education, Seattle, WA.

Camilli, G. & Penfield, D. A. (1997). Variance estimation for differential test functioning based on Mantel-Haenzel statistic. Journal of educational measurement, 34 (2), 123-139.

Camilli, G., Wang, M. & Fesq, J. (1995). The effects of dimensionality on equating the Law School Admissions Test. Journal of educational research, 32 (1), 79-96.

Cohen, J. (1992). A power primer. Psychological bulletin, 112(1), 155-159.

Eignor, D. R., Stocking, M. L., Way, W. D. & Steffen, M. (1993). Case studies in computer adaptive test design through simulation. (Research report RR-93-56). Princeton, NJ: Educational Testing Service.

Geiger, M. A. (1991). Changing multiple-choice answers: Do students accurately perceive their performance? Journal of experimental education, 59(3), 250-257.

Gershon, R., & Bergstrom, B. (1995). Does cheating on CAT pay: NOT! Paper presented at the annual meeting of the American Educational Research Association, San Francisco, CA. (ERIC document reproduction service No. ED 392 844).

Green, B. F., Bock, R. D., Humphreys, R. L., & Reckase, M. D. (1984). Technical guidelines for assessing computerized adaptive tests. Journal of educational measurement, 21 (4), 347-360.

Harvill, L. M., & Davis III, G. (1997). Medical student's reasons for changing answers on multiple-choice tests. Academic medicine, 72(10), 97-99.

Harwell, M., Stone, C. A., Hsu, T. & Kirisci, L. (1996). Monte Carlo studies in item response theory. Applied psychological measurement, 20(2), 101-125.

Kim, J. K., & Nicewander, W. A. (1993). Ability estimation for conventional tests. Psychometrika, 58 (4), 587-599.

Kingsbury, G. G. (1996) Item review and adaptive testing. Paper presented at the annual meeting of the National Council of Measurement in Education, NY.

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

Lord, F. M. (1986). Maximum Likelihood and Bayesian parameter estimation in item response theory. Journal of educational measurement. 23 (2), 157-162.

Luecht, R. M. & Hirsch, T. M. (1992). Item selection using an average growth approximation of target information functions. Applied psychological measurement, 16(1), 41-51.

Lunz, M. E., Bergstrom, B. A., & Wright, B. D. (1992). The effect of review on student ability and test efficiency for computerized adaptive tests. Applied psychological measurement, 16 (1), 33-40.

McBride, J. R., Wetzel, C. D & Hetter, R. D (1997). Preliminary psychometric research for CAT-ASVAB: Selecting an adaptive testing strategy. . In W. A. Sands, B. K. Waters, & J. R. McBride (Eds.), Computerized adaptive testing. From inquiry to operation (pp. 83-95). Washington, DC: American psychological association.

Mills, C. N., & Stocking, M. L. (1995). Practical issues in large-scale high-stakes computerized adaptive testing. (Research Report 95-23). Princeton, NJ: Educational Testing Service.

Moreno, K. E. (1997). CAT-ASVAB Operational test and evaluation. . In W. A. Sands, B. K. Waters, & J. R. McBride (Eds.), Computerized adaptive testing. From inquiry to operation (pp. 199-205). Washington, DC: American psychological association.

Owen, R. J. (1975). A Bayesian sequential procedure for quantal response in the context of adaptive mental testing. Journal of the American statistical association, 70(350), 251-355.

Patsula, L. N. & McLeod. L. D. (2000). Detecting test-takers who have memorized items in computerized-adaptive testing and multi-stage testing: A comparison. Paper presented at the annual meeting of the National Council of Measurement in Education, New Orleans, LA. .

Pommerich, M & Burden, T. (2000). From simulation to application: Examinees react to computerized testing. Paper presented at the annual meeting of the National Council of Measurement in Education, New Orleans, April 2000.

Powers, D. E. (1999). Test anxiety and test performance: Comparing paper based and computer-adaptive versions of the GRE general test. (Research Report 99-15). Princeton, NJ: Educational Testing Service.

Reckase, M. D. (1975). The effect of item choice on ability estimation when using a simple logistic tailored testing model. Paper presented at the annual meeting of the American Educational Research Association, Washington, D.C.

Reckase, M. D. (2000). Computerized testing- The adolescent years: Juvenile Delinquent or positive role model? Paper presented at the annual meeting of the National Council of Measurement in Education, New Orleans, April 2000.

Robin, F. (1999). CBTS: Computer-based testing simulation and analyses [computer program]. Amherst, MA: University of Massachusetts, Laboratory of Psychometric and Evaluative Research.

Robin, F., Xing, D., Scrams, D. & Potenza, M. (2000). Classification accuracy and test security for a computerized adaptive mastery test calibrated with different IRT models. Paper presented at the annual meeting of the National Council of Measurement in Education, New Orleans, LA.

SAS Institute Inc. (1999). Language Reference: Concepts, Version 8. Cary, NC.

Schwartz, S. P., McMorris, R. F., & DeMers, L. P. (1991). Reasons for changing answers: An evaluation using personal interviews. Journal of educational measurement, 28(2), 163-171.

Stocking, M. L. (1997). Revising item responses in computerized adaptive tests: A comparison of three models. Applied psychological measurement, 21 (2), 129-142.

Stone, G. E., & Lunz, M. E. (1994). The effect of review on the psychometric characteristics of computerized adaptive tests. Applied measurement in education, 7(3), 211-222.

van der Linden, W. & van Krimpen-Stoop, E. M.L.A. (2001). Using response times to detect aberrant behavior in computerized adaptive testing. Paper presented at the annual meeting of the National Council of Measurement in Education, Seattle, WA.

Vispoel, W. P., Clough, S. J., Bleiler, T., Henderickson, A. B. & Ihrig, D. (2001). Can examinees use judgments of item difficulty to improve proficiency estimates on computerized adaptive vocabulary tests? Paper presented at the annual meeting of the National Council of Measurement in Education, Seattle, WA.

Vispoel, W. P. (1998a).Psychometric characteristics of computer-adaptive and self-adaptive vocabulary tests: The role of answer feedback and text anxiety. Journal of educational measurement, 35 (2), 155-167.

Vispoel, W. P. (1998b). Review and changing answers on computerized adaptive and self-adaptive vocabulary tests. Journal of educational measurement, 35 (4), 328-347.

Vispoel, W. P., Henderickson, A. B. & Bleiler, T. (2000). Limiting answer review and change on computerized adaptive vocabulary tests: Psychometric and attitudinal results. Journal of educational measurement, 37(1), 21-38.

Vispoel, W. P., Rocklin, T. R., & Wang, T. (1994). Individual differences and test administration procedures: a comparison of fixed-item, computerized-adaptive, and self-adapted testing. Applied measurement in education, 7(1), 53-79.

Vispoel, W. P., Rocklin, T. R., Wang, T. & Bleiler, T. (1999). Can examinees use a review option to obtain positively biased estimates on a computerized adaptive test? Journal of educational measurement, 36 (2), 141-157.

Waddell, D. L. & Blankenship, J. C. (1994). Answer changing: A meta-analysis of the prevalence and patterns. The journal of continuing education in nursing, 25, 155-158.

Wagner, D., Cook, G., & Friedman, S. (1998). Staying with their first impulse? The relationship between impulsivity/reflectivity, field dependence/field independence and answer changes on a multiple-choice exam in a fifth-grade sample. Journal of research and development in education, 31 (3), 166-175.

Wainer, H. (1992). Some practical considerations when converting a linearly administered test to an adaptive format. (Research Report 92-21). Princeton, NJ: Educational Testing Service.

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

Wainer, H., Dorans, N. L., Green, B. F., Mislevy, R. J., Steinberg, L. & Thissen, D. (2000). Future challenges. In H. Wainer (Ed.), Computerized adaptive testing: A primer (2nd ed., pp.231-270). Mahwah, NJ: Laurence Erlbaum Associates Publishers.

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

Wise, S. L. (1996). A critical analysis of the arguments for and against item review in computerized adaptive testing. Paper presented at the annual meeting of the National Council on Measurement in Education, New York. (ERIC document reproduction service No. ED 400 267).

Wise, S. L. (1997a). Examinee issues in CAT. Paper presented at the annual meeting of the National Council on Measurement in Education, Chicago, IL. (ERIC document reproduction service No. ED 408 329).

Wise, S. L. (1997b). Overview of practical issues in a CAT program. Paper presented at the annual meeting of the National Council on Measurement in Education, Chicago, IL. (ERIC document reproduction service No. ED 408 330).

Wise, S. L., Freeman, S. A., Finney, S.J ., Enders, C. K., & Severance, D. D. (1999). Examinee judgments of changes in item difficulty: Implications for item reviewing in computerized adaptive testing. Applied measurement in education, 12(2), 185-198.

Wise, S. L., Roos, L. R., Plake, B. S., & Nebelsick-Gullett, L. J. (1994). The relationship between examinee anxiety and preference for self-adapted testing. Applied measurement in education, 7(1), 81-91.