# DIFFERENTIAL ITEM FUNCTIONING ON A COLLEGE FRESHMAN CHEMISTRY ACHIEVEMENT TEST USING THE CREDIBILITY INTERVAL APPROACH

by

AMINAH F. PERKINS

(Under the Direction of Karen Samuelsen)

## ABSTRACT

Assessments are used to measure test takers knowledge in a particular area. In this study, the Credibility Interval Approach using a Rasch model was used to identify differential item functioning (DIF) on an undergraduate chemistry achievement test. Results from this test can have impacts such as course placement and grade distribution. DIF, a psychometric characteristic that occurs when people from different groups who are matched on some latent trait have a significantly different probability of giving a certain response to a test item, is present for a portion of the items on this test. Item difficulties and 95% posterior credibility intervals of the items were analyzed to determine which items exhibited DIF. Items that were shown to have DIF were further explored to assess whether males or females were advantaged. In addition, it was determined whether DIF was consistent by group, test form, or concept grouping.

INDEX WORDS: Differential Item Functioning, Credibility Interval Approach, Chemistry, Achievement Test

# DIFFERENTIAL ITEM FUNCTIONING ON A COLLEGE FRESHMAN CHEMISTRY ACHIEVEMENT TEST USING THE CREDIBILITY INTERVAL

by

AMINAH PERKINS

B.S., Spelman College, 2004

A Thesis Submitted to the Graduate Faculty of The University of Georgia in Partial Fulfillment of the Requirements for the Degree

MASTERS OF ARTS

ATHENS, GEORGIA

2008

DIFFERENTIAL ITEM FUNCTIONING ON A COLLEGE FRESHMAN CHEMISTRY ACHIEVEMENT TEST USING THE CREDIBILITY INTERVAL

by

AMINAH F. PERKINS

| | |
|---|---|
| Major Professor: | Karen Samuelsen |
| Committee: | Deborah Bandalos<br>Allan Cohen |

Electronic Version Approved:

Maureen Grasso
Dean of the Graduate School
The University of Georgia
May 2008

# DEDICATION

I dedicate this document as I did my undergraduate thesis to my wonderful mother. She has been an example of strength and determination for me throughout my life. She served as my rock never faltering in her love or lacking in encouragement. She deserves all the credit for my accomplishments. Additionally, I dedicate this document to my loving grandmother whose memory I hold close to my heart. These women are my inspiration.

## ACKNOWLEDGEMENTS

First, I would like to extend gratitude to Dr. Charles Atwood for providing the data set. Also, thank you to Dr. Atwood and Ms. Kimberly Schurmeier for supplying background information about the achievement test and readily answering a multitude of questions. I would also like to thank my advisor and committee chair, Dr. Karen Samuelsen, for her guidance and mentoring through the entire thesis process. Furthermore, I wish to thank Drs. Deborah Bandalos and Allan S. Cohen for serving on my thesis committee and guiding me through this process.

.

# TABLE OF CONTENTS

# LIST OF TABLES

Page

## LIST OF FIGURES

# CHAPTER 1

# INTRODUCTION AND DIFFERENTIAL ITEM FUNCTIONING

## Introduction

The present study was designed to employ the Credibility Interval Approach to identify differential item functioning on a chemistry achievement test. Differential item functioning (DIF), a psychometric characteristic that occurs when people from different groups who are matched on some latent trait have a significantly different probability of giving a certain response to a test item, is present for a portion of the items on this test. Item difficulties and 95% posterior credibility intervals for the differences between item difficulty parameters for males and females were analyzed to determine which items exhibit DIF. Items that were shown to have DIF were further explored to assess whether males or females were advantaged. In addition, it was determined whether DIF was consistent by group, test form, or concept grouping.

The chemistry test was administered to 1,281 undergraduate students at a large public Southeastern university by means of a computer. The need for a variety of testing administrations made the use of multiple test forms ideal for security purposes. In turn, these forms, which will be described in more depth in subsequent sections, were administered to a small number of students. Traditional methods were not appropriate for identifying DIF due to the small sample size of each form. Forms were generated using a test bank of items. Test developers created the test so that items could appear on multiple forms. For example, only 32 examinees received form 4, while 305 examinees received item v001, which was a member of form 4. Traditional methods for assessing DIF are not robust enough to handle the complexity of these data. For that reason the Markov Chain Monte Carlo sampling was employed. This was done using the Credibility Interval Approach to assess DIF within a Bayesian framework (Samuelsen & Bradshaw, 2008).

This study is unique in two ways. First there has not been extensive research of differential item functioning in the area of chemistry in general nor in the chemistry department at the institution where the test was administered. The study fills an area of need in not only chemistry education but also psychometrics. Second, this study takes a new procedure, the Credibility Interval Approach proposed by Samuelsen and Bradshaw (2008) using a simulation, and applies it to data. The Credibility Interval Approach utilizes the WinBUGS (Spiegelhalter, Thomas, Lunn & Best, 2000) software to employ Markov Chain Monte Carlo sampling. The following information in this first chapter presents an exploration of DIF and a discussion of its importance.

## Differential Item Functioning

Tests are used across countless areas, including academic achievement and placement, job placement, and licensure. The results from these tests often times have non-ignorable impacts on the lives of many people. Students can receive grades or be placed in specific academic courses based on the results of these scores. For that reason it is important for test developers to be aware of many threats to validity when constructing assessments. Messick (1989) described validity as “an inductive summary of both the existing evidence for and the potential consequences of score interpretation and use” p. 13. The essence of his ideas was later adopted in the APA, AERA and NCME standards (1999). One such threat to validity is the presence of differential item functioning because this can result in issues of equity when interpreting the tests results. For example, assume that for a given test a female and a male receive the same score; therefore one might infer that they are of equal ability. Now assume that DIF analyses show that the same test has several items that disadvantage females. The previous inferences about the ability levels of the female and male would be incorrect because the DIF analysis suggests females have a significantly different probability of giving a certain response to a test item as compared to males. If decisions, e.g. course placement or job assignments, had been made based on the test scores of the female and male without performing a DIF analysis they might be inaccurate.

It is important to start with a discussion of the differences between item bias and differential item functioning, both considered threats to validity. Bias is the presence of some characteristic of an item that results in differential performance for individuals of the same ability level but from different groups (Holland & Thayer, 1988). Clauser and Mazor (1998) define DIF as the differing probabilities of success on an item between groups after they have been matched on ability. As such, a biased item will exhibit DIF but DIF alone is not an adequate indicator of item bias. Many now use the terms DIF and bias interchangeably but historically these two have been separate but related issues (Zumbo, 2007). Another important differentiation to note is between DIF and impact. Impact is the difference in developed abilities measured by tests and intact groups (Dorans & Holland, 1993). DIF differs from impact in that differences amongst groups are assessed after the groups have been matched on the ability of interest. Understanding the differences amongst the definitions of DIF, item bias, and impact is necessary prior to DIF research and appropriately interpreting DIF results.

An important step in determining whether there is DIF is matching the examinees on the ability of interest. Valid external measures would be the most optimal choice as a matching criteria but it is rare that such a measure would be available. Thus, the use of an internal criterion such as total test score is typically the best available measure. The use of an internal criteria means that items with DIF are included in the matching criterion. Clauser and Mazor (1998) note the limitations in the use of an internal criterion by stating that the use of total test score as the matching criteria requires that the measure of effect size for DIF items favoring the reference group must approximately offset that for items favoring the focal group. For example, if a test has a moderate item with DIF that favors males, then a moderate item with DIF favoring females might work to offset it so that the total test score, the internal criteria, is still relevant. Additionally, the validity of the test score or other internal measure must be considered as it is used in the analysis.

When differences between groups are present after matching on a particular ability then it is assumed that performance on that item depends on something other than ability that has not been taken into account (Clauser & Mazor, 1998). Ackerman (1992) refers to these as nuisance abilities defining them as skills used by an examinee to solve an item that are not intended to be assessed by the test. This leads to the existence of two or more dimensions, one being the ability that was intended to be measured and the other the nuisance abilities described by Ackerman (1992). When a test is measuring skills other than those that were intended the test lacks construct validity. Construct validity as defined by Messick (1989) is evaluated by "determining the degree to which certain explanatory concepts or constructs account for performance on the test" p. 16. Nuisance abilities are threats to construct validity in the form of what is referred to as construct irrelevant variance, meaning the assessment is too broad and contains extra variance associated with other constructs. Construct irrelevant difficulty and construct irrelevant easiness are two types of construct irrelevant variance. When there are skills necessary to answer the item that are outside the realm of the construct intended to be measured, which in turn make the item more difficult for select groups, construct irrelevant difficulty has occurred. Quite the opposite, construct irrelevant easiness results when clues within an item exist that allow select individuals to respond correctly to items in ways that exceed the construct being measured.

The first general step for a DIF analysis requires identifying which groups will be compared. The focal group is typically the group of interest. In most cases this group represents some minority or often disadvantaged group such as females or African-Americans. The reference group will then be the basis for comparison. After identifying the groups of interest a matching criterion must be determined as was discussed above. At this point, one of the many DIF statistical methods must be selected and subsequent analysis performed. Choosing which method to employ will be determined by the available data and the research question. There are several statistical procedures that can be used to assess DIF. The next chapter explores Item Response Theory based methods and the Credibility Interval Approach, a procedure for assessing DIF in complex models. As a final step in DIF analysis, results must be interpreted and decisions must be made.

DIF can occur in two different types as uniform and non-uniform. Uniform DIF occurs when there is not an interaction between ability level and group membership on an item. For example, a math test might show that females at all ability levels consistently score higher than their male counterparts. Non-uniform DIF occurs when an interaction between ability level and group membership exists. In this situation perhaps higher ability girls score more than higher ability boys, while lower ability girls might score less than lower ability boys.

One way of conceptualizing DIF is through the use of contingency tables as in the Mantel-Haenszel procedure. Originally introduced by Mantel and Haenszel (1959), the procedure later coined by their names as the Mantel-Haenszel (MH) approach was an original way of studying matched groups. Mantel and Haenszel (1959) used this procedure in the area of cancer research and Dorans and Holland (1985) adapted the unique procedure for the assessment of DIF. In the usual two group case, a reference and focal group are formed, then matched on the ability of interest, then compared for probabilities of success. When these probabilities are significantly different DIF is present. The MH procedure utilizes 2 x 2 contingency tables of the following form to study each item at each level of the matching criterion.

Table 1.
*Contingency Table of $j^{th}$ Matched Set of Members*

| | Item Score | | |
|---|---|---|---|
| Group | 1 | 0 | Total |
| Reference | $A_j$ | $B_j$ | $N_{rj}$ |
| Focal | $C_j$ | $D_j$ | $N_{Fj}$ |
| Total | $M_{1j}$ | $M_{0j}$ | $T_j$ |

$A_j$ corresponds to members of the reference group of the $j^{th}$ matched set who correctly answered the item. $N_{rj}$ represents all individuals in the reference group who are members of the $j^{th}$ matched set. Finally, $T_j$ corresponds to the total number of reference group and focal group members in the $j^{th}$ matched set. Definitions for the other entries in the table are similar to those just outlined.

A limitation to the MH approach is that the procedure is intended to measure uniform DIF and is not sensitive to non-uniform DIF meaning it does not allow for the testing of an interaction between the ability level and group membership (Rogers & Swaminathan, 1993). This shortcoming is shared by the Rasch model that will be described in the next chapter.

# CHAPTER 2

# STATISTICAL PROCEDURES

## Item Response Theory

Item response theory (IRT) is a way to examine the relationship between individuals' responses to test items and the construct measured by a test (Thissen, Steinberg, & Wainer, 1993). In unidimensional IRT, each examinee has an unobservable latent trait such as ability, typically represented by the Greek letter theta, *$\theta$*. At each ability level, there is a probability, *$P(\theta)$,* estimated that the examinee will obtain a correct answer to the item (Baker, 2001). If the IRT model fits the data, at higher ability levels *$P(\theta)$* will be subsequently larger. A graphical representation of the relationship between ability level and *$P(\theta)$* can be achieved through the use of an item characteristic curve (ICC). An ICC is a monotonically increasing S-shaped curve whose slope changes as a function of the ability level *$\theta$* and reaches a maximum value at the point where ability is equal to the difficulty of the item (Baker, 2001). An ICC exists for every item on a particular test. Figure 1 displays a typical ICC.

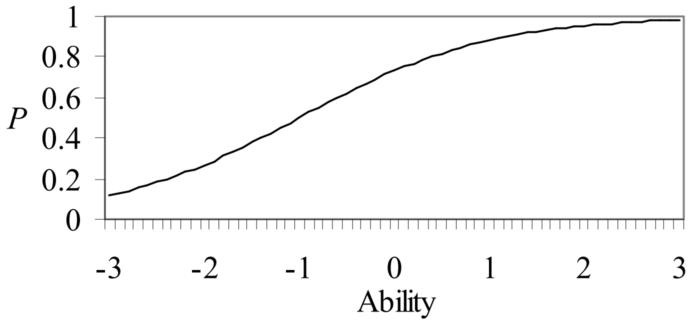

*Figure 1.* A typical item characteristic curve

There are three item parameters that determine the shape of the ICC, discrimination, difficulty, and pseudo-guessing. Item discrimination, denoted by the letter *a*, refers to how well an item can differentiate between examinees of varying ability levels. This parameter accounts for the steepness or flatness of the middle section of the ICC as seen in Figure 2. Steep ICC's represent items that discriminate well while flatter ICC's correspond to items that do not discriminate well. If the curve is flat, the probability of a correct response at a higher ability level and a lower ability level will be very similar. In Figure 2, the ICC modeled with a = .5 does not discriminate as well as the ICC modeled with a = 2.5. The typical range for the discrimination parameter is -2.80 < *a* < +2.80 (Baker, 2001).

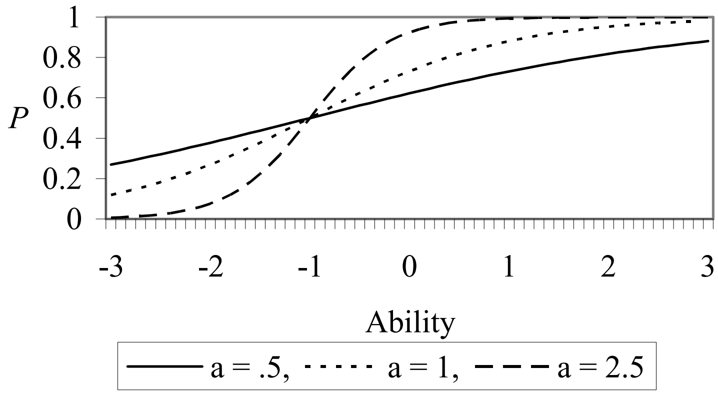

*Figure 2*. ICC's with differing discrimination values and all other parameters held constant

The difficulty parameter denoted by the letter *b*, is often referred to as the location parameter because it describes where the ICC inflects along the ability scale as depicted in Figure 3 (Baker, 2001). In the absence of guessing, the difficulty parameter is defined as the point on the ability scale where *P(θ)* is equal to .5 which is illustrated with the extrapolated lines shown in Figure 3 (Baker, 2001),. At this point the ICC changes from increasing to decreasing. Typically the range for the difficulty parameter is -3 < *b* < +3 (Baker, 2001).

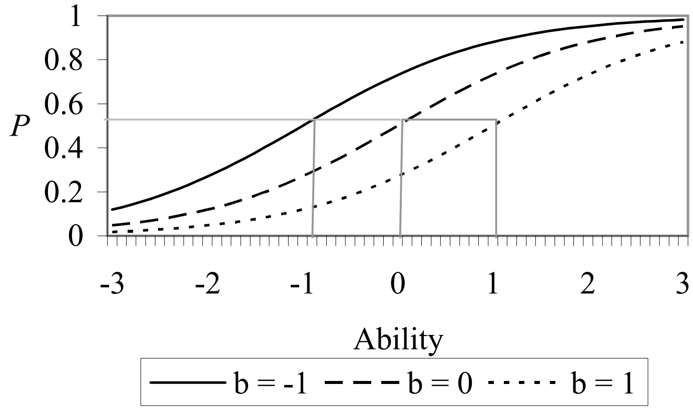

*Figure 3.* ICC's with differing difficulty values and all other parameters held constant

The pseudo-guessing parameter, denoted by the letter *c*, refers to the probability of obtaining a correct response on an item by either randomly guessing or guessing from a limited number of options after some distractors have been judged incorrect. This parameter corresponds to the lower asymptote of the ICC. Figure 4 demonstrates the change in an ICC as the guessing parameter increases.

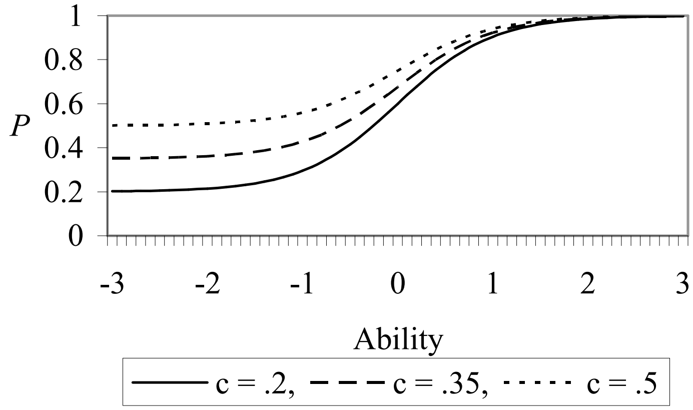

*Figure 4*. ICC's with differing pseudo-guessing values and all other parameters held constant

In IRT each individual item on a test is assumed to measure or is at least modeled so it measures an underlying latent trait. This allows for the amount of information to be computed at any ability level based on a given item (Baker, 2001). Item information is maximized at the point where ability level and item difficulty are equivalent. A decrease is seen in item information as the ability level travels further from the difficulty parameter.

For the purpose of measuring if an item exhibits DIF, the ICC's for the reference and focal groups must be compared. If the two ICC's differ, then the item is said to exhibit DIF (Thissen, Steinberg, & Wainer, 1993). When the ICC's are identical, the item parameters are not different and thus, the item does not exhibit DIF. Figure 5 presents two ICC's, one for a group of females and the other for a group of males. This item can be said to have uniform DIF where the females at all ability levels consistently score higher than the males. On the contrary, Figure 6 displays an item that exhibits non uniform DIF as defined in previous sections.

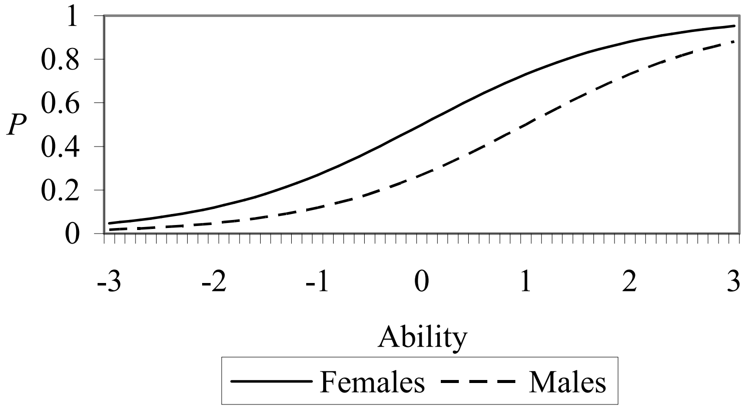

*Figure 5*. Illustration of uniform DIF with females advantaged

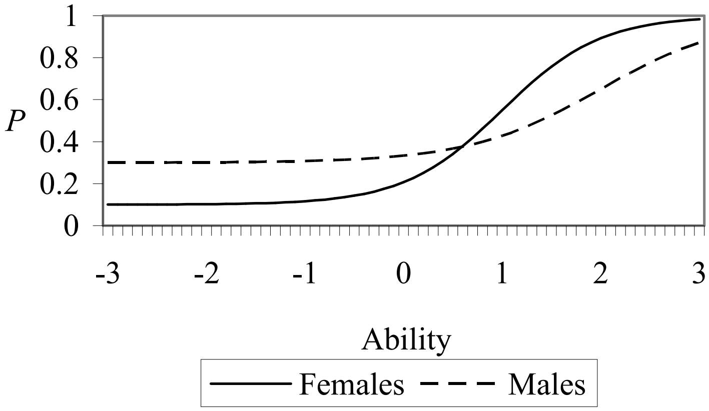

*Figure 6*. Illustration of non-uniform DIF

There are three common models, the one-, two-, and three- parameter logistic models, represented by mathematical equations that are typically used to signify the relation of the probability of a correct response to ability in IRT. One or more of the item parameters discussed are utilized in each model to generate an ICC.

The Rasch or one-parameter logistic model employs only the difficulty parameter, *b* as can be seen in equation 1. Under this model the probability of answering an item correctly, when *θ* is equal to *b,* is always .5. It is at this point that item information is maximized.

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

The two-parameter logistic (2PL) model considers the difficulty parameter, *a,* in addition to the aforementioned discrimination parameter. As in the Rasch model, *P(θ)* is .5 when *θ* is equal to *b*. Additionally, if the *a* parameter is restricted to be one for all items on the test, then a 2PL model can be thought of as a Rasch model.

$$P(\theta) = \frac{1}{1 + e^{-L}} = \frac{1}{1 + e^{-a(\theta - b)}} \tag{2}$$

In equation 2, *e* indicates the natural log and *L* is equal to the logistic deviate which is *a(θ-b)*. In the Rasch and 2PL models, the lower limit of the ICC is zero.

The three-parameter logistic (3PL) model incorporates the pseudo-guessing parameter, *c,* in addition to the difficulty, *a,* and the discrimination, *b,* parameters.

$$P(\theta) = c + (1 - c)\frac{1}{1 + e^{-a(\theta - b)}} \tag{3}$$

As stated before, *c* corresponds to the lower asymptote of the ICC. In the Rasch and 2PL models this parameter was zero.

In terms of assessing DIF, items for the reference and focal groups can differ on one parameter such as difficulty in the simplest case or on all the parameters (difficulty, discrimination, guessing). IRT allows for the reference and focal groups to be placed on the same scale. This is advantageous because it permits the between-group differences in the item parameters for the specific model to be estimated. In addition, IRT allows for the estimation of *θ,* which will not be test specific and one would expect to arrive at a similar estimation for different items. Major limitations to IRT methods include the need for large samples as well as the need for the data to meet the unidimensionality assumptions of the models (Clauser & Mazor, 1998).

## Credibility Interval Approach

There are many instances where data can become very complex and conventional methods for assessing DIF might not prove robust enough to yield useful results. Complex data can be obtained, as in the case of this study, from the use of multiple test forms. In this way, there exists a test bank of questions that are systematically allocated to multiple test forms. This process allows for some items to be present on multiple test forms. Although the data set itself might be large, the number of respondents for each test form, depending on the number of forms, could be relatively small. When data are presented in a matrix format of respondents by items there can be a significant amount of missing data present. In these situations, as well as in the case of computer adaptive testing, using the IRT or MH approaches presented earlier could be problematic. Samuelsen and Bradshaw (2008) presented one method to approaching complex models using a credibility interval approach to detect DIF within a Bayesian framework. Bayesian estimation uses Bayes' theorem, presented in equation 4, to estimate the likelihood of an unknown probability density function.

$$p(\theta|y) = \left( \frac{p(y|\theta)p(\theta)}{\int p(y|\theta)p(\theta)d\theta} \right) \qquad (4)$$

In equation 4, $p(y|\theta)$ denotes the likelihood and $p(\theta)$ the prior density for the vector of $k$ model parameters $\theta$.

Markov Chain Monte Carlo (MCMC) methods can be incorporated when a solution is unattainable using mathematical solutions. MCMC procedures simulate random samples from a theoretical distribution and then use those samples in making inferences about the features of that theoretical distribution. That is given a set of random draws $\theta^{(1)}$, $\theta^{(2)}$,…. , $\theta^{(G)}$ from the posterior distribution, virtually all summaries of interest can be estimated from the posterior distribution. Kass, Carlin, Gelman, and Neal (1998) state that MCMC methods have been doing well because they allow for simulations to be drawn from a wide range of distributions. This can be conceptualized through the understanding that the sample mean is an unbiased estimator of the population mean. Suppose that a random sample, $X_1$, $X_2$,…, $X_n$, is collected from a population. An estimate of the population mean, a sample mean, can be calculated. If this sampling were repeated a second time a different set of observations would be obtained and thus a sample mean different from the first sample would be calculated. If these estimates were averaged over many repetitions of the sampling experiment then the assumption is that it would approximate the true population mean. This study is using this sampling method.

The first step in a MCMC approach that utilizes the WinBUGS software (Spiegelhalter, Thomas, Lunn & Best, 2000), such as the Credibility Interval Approach, is the burn-in process. In this process, the number of iterations needed for convergence is determined and those iterations are discarded resulting in only draws from the posterior distribution being used with all further samples being thought of as coming from the stationary distribution. Convergence can be determined in many ways, including time series plots, plots of the autocorrelation function, and Brooks-Gelman-Rubin (BGR) diagnostic plots (Samuelsen & Bradshaw, 2008). In order to obtain confidence in the inferences made about the posterior distributions, density plots are examined for smoothness to determine if a sufficient number of iterations have been run.

In this approach "DIF is defined as the difference between item difficulties across the reference and focal groups" (Samuelsen & Bradshaw, 2008, p. 5). Differences in the posterior distributions are assessed using the WinBUGS software (Spiegelhalter, Thomas, Lunn & Best, 2000). Items are identified as functioning differently when zero is not included in the 95% credibility interval of the differences between item difficulties. Currently this method has only been tested using a Rasch model. As such, item difficulty is the only parameter that would be estimated.

# CHAPTER 3

# PROCEDURE

## Instrumentation

The current study used data from a chemistry achievement test from a large public Southeastern university administered during the Fall semester of 2006. This test was required of all students, who were mainly freshman (<1% upperclassmen), enrolled in a particular course section and is the second of three general examinations administered to students via computer. The test is given at the commencement of the semester and taken by examinees in a university computer lab where they were given 80 minutes to complete the exam. During that time examinees were allowed three tries to answer each question. Try one was submitted during the exam and immediately graded and the questions answered incorrectly were given back to examinees for a second try. This procedure was repeated three times. If a question was answered correctly on try one the examinee received 100% credit, try 2 - 50% credit and try 3 - 25% credit for each question. These items consist of number entry, multiple choice, multiple answer and formula entry question types. In this study only results from the first try were analyzed.

## Sample

The data for the Fall 2006 university chemistry achievement test consists of responses from the 1,281 students who took the test. A total of 43 respondents did not specify their gender and were consequently omitted from further analysis resulting in a sample size of 1,238. Females represented 59.5% of the examinees and males 40.5%.

There were a total of 110 items in the overall test bank. Items were clustered into 25 groupings based on chemistry concepts. As an example, gas laws can be classified as a chemistry concept. This specific concept group could contain as many as three items that focus on some aspect of the following equation, PV=nRT. One question might ask examinees to calculate Pressure (P); others might require calculating Volume (V) or Temperature (T). An item from each of the 25 concept groups was taken to create 37 test forms each with 25 items. Each item was placed on multiple test forms. The assignment of items to forms was performed by the test developers. This is illustrated in table 2 that shows the first 10 items and 15 of the 37 forms. Notice for example, that item v001 can be found on forms 4, 8, and 12. On average an item appears on 8 different forms. There is a mode of 37 respondents who completed each form with a range of 8 to 42. Each test item was administered to 281 respondents on average. Table 3 displays the concept grouping and question type for each item. For example, items v001, v002, v003, and v004 are all members of concept grouping 1 and are numerical questions.

Table 2.

*Depiction of individual items on multiple test forms.*

| | | | | | Item | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| **Form** | **v001** | **v002** | **v003** | **v004** | **v005** | **v006** | **v007** | **v008** | **v009** | **v010** |
| 1 | | X | | | | X | | | X | |
| 2 | | | X | | | | X | | | X |
| 3 | | | | X | X | | | | | |
| 4 | X | | | | | X | | X | | |
| 5 | | X | | | | | X | | X | |
| 6 | | | X | | X | | | | | X |
| 7 | | | | X | | X | | | | |
| 8 | X | | | | | | X | X | | |
| 9 | | X | | | X | | | | X | |
| 10 | | | X | | | X | | | | X |
| 11 | | | | X | | | X | | | |
| 12 | X | | | | X | | | X | | |
| 13 | | X | | | | X | | | X | |
| 14 | | | X | | | | X | | | X |
| 15 | | | | X | X | | | | | |

*Note*. “X” implies item was present on the given form.

Table 3.

*Item Groupings*

| **Concept** | **Items** | **Question Type** |
|---|---|---|
| 1 | v001, v002, v003, v004 | numerical answer |
| 2 | v005, v006, v007 | multiple choice |
| 3 | v008, v009, v010, v011 | numerical answer |
| 4 | v012, v013, v014, v015 | multiple choice |
| 5 | v016, v017, v018, v019, v020 | multiple choice |
| 6 | v021, v022, v023, v024, v025 | multiple answer |
| 7 | v026, v027, v028, v029 | multiple choice |
| 8 | v030, v031, v032, v033 | numerical answer |
| 9 | v034, v035, v036, v037, v038, v039 | multiple choice |
| 10 | v040, v041, v042, v043, v044 | numerical answer |
| 11 | v045, v046, v047, v048 | multiple choice |
| 12 | v049, v050, v051, v052, v053 | numerical answer |
| 13 | v054, v055, v056, v057, v058 | multiple choice |
| 14 | v059, v060, v061, v062, v063 | multiple answer |
| 15 | v064, v065, v066, v067, v068 | multiple choice |
| 16 | v069, v070, v071, v072, v073 | numerical answer |
| 17 | v074, v075, v076, v077, v078, v079 | multiple choice |
| 18 | v080, v081, v082, v083 | numerical answer |
| 19 | v084, v085 | multiple choice |
| 20 | v086, v087, v088, v089 | numerical answer |
| 21 | v090, v091, v092, v093, v094, v095 | multiple choice |
| 22 | v096, v097, v098, v099 | numerical answer |
| 23 | v100 | multiple choice |
| 24 | v101, v102, v103, v104, v105 | text entry |
| 25 | v106, v107, v108, v109, v110 | numerical answer |

## Method

Although the overall data set used for this study is relatively large, the average number of respondents for each test form was only 33. This small sample size makes most traditional approaches for detecting DIF a poor choice. Presenting this data in a matrix format of respondents by items leads to a large amount of missing data since every respondent does not receive every item. On average each item was administered to 281 students, or 22.7% of the students. Refer to table 4 to see a representation of this situation. Table 4 presents a snapshot of the data in the form of a matrix of item responses. There are 74 instances of missing data represented by "NA" and only 26 pieces of useful information.

Table 4.

*Matrix of respondents by items*

| | Items | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| | **v001** | **v002** | **v003** | **v004** | **v005** | **v006** | **v007** | **v008** | **v009** | **v010** |
| 10003 | NA | 0 | NA | NA | NA | 0 | NA | NA | 1 | NA |
| 10004 | NA | 0 | NA | NA | NA | 0 | NA | NA | 1 | NA |
| 10006 | NA | NA | 0 | NA | NA | NA | 1 | NA | NA | 0 |
| 10012 | NA | NA | 0 | NA | NA | NA | 0 | NA | NA | 0 |
| 10013 | NA | NA | 0 | NA | NA | NA | 0 | NA | NA | 1 |
| 10014 | NA | NA | 0 | NA | NA | NA | 0 | NA | NA | 0 |
| 10016 | NA | NA | NA | 0 | 0 | NA | NA | NA | NA | NA |
| 10019 | NA | NA | NA | 0 | 0 | NA | NA | NA | NA | NA |
| 10020 | NA | NA | NA | 0 | 0 | NA | NA | NA | NA | NA |
| 10022 | NA | NA | NA | 1 | 0 | NA | NA | NA | NA | NA |

*Note.* "NA" means the respondent did not receive the item. "0" means the respondent answered the item incorrectly. "1" corresponds to an item being correctly answered by a respondent.

The complexity of these data requires the use of a statistical technique that will adequately assess DIF. The method that was utilized is implemented using the WinBUGS software for Bayesian analysis of these data using MCMC techniques. The DIF test for this approach uses the Credibility Interval Approach described earlier. The complete matrix of all respondents in the data set by every item with corresponding 1's, 0's, and NA's, as in Table 4 above, constitutes the raw data file that was inputted into WinBUGS using the Credibility Interval Approach. This method was presented by Samuelsen and Bradshaw (2008) in their research with data simulated using the Rasch model. Samuelsen and Bradshaw note two reasons for using the Rasch model. The first is that their research is the first to use the Credibility Interval Approach and thus it seems logical to start with the simplest of models. Secondly, they chose the Rasch model because it "is ubiquitous in state testing" (Samuelsen & Bradshaw, 2008, p. 6). The present study will do the same as there has been no research done supporting the use of any other models.

The WinBUGS code used for this study was adapted from the one used by Samuelsen and Bradshaw (2008) and can be found in Appendix A of this document. The parameter estimation of the two gender groups was done concurrently ensuring the items were on the same scale. Within the WinBUGS code there exists a trigger mechanism for the Rasch model that activates group membership. The vector for group membership, x[i,111], is dichotomously coded to correspond to the reference and focal groups. In the coded Rasch model, when a cell from this vector represents a female it will then incorporate that females response data to contribute information to determine her ability and item difficulties for each item she received. The process is then repeated for all members of each gender group. A mean item difficulty for each item is calculated for each gender group. Then the gender group differences between item difficulties are calculated for each item by subtracting the mean item difficulty for the focal group from the mean item difficulty of the reference group. Also a mean ability is calculated for each gender group. In order to monitor the differences in the posterior distributions of the item difficulties of reference and focal groups, prior distributions must be set and used. The following distributions are recommended by prior research (Samuelsen & Bradshaw, 2008; Bolt, Cohen, & Wollack, 2002; Wollack, Cohen, & Wells, 2003) to ensure convergence.

- Item difficulties (i) within groups (g): $b[i,g] \sim \text{Normal}(0,1)$
- Ability distributions within groups: $\theta[n,g] \sim \text{Normal}(\mu[g],1)$

- Means of the ability distributions within groups: μ[g] ~ Normal(0,1)
- Item responses for examinees on items: x[n,i] ~ Bernoulli(P[n,i])

Constraints were placed on the model based on the item difficulties within groups summing to zero. This results in item parameters from both groups being modeled on the same scale. In this way if DIF does not exist then they are the same within estimation error, and if DIF does in fact exist, then the DIF averages to zero since the item parameters are being centered around zero in both groups. The process for making the item parameters center around zero begins with estimating the item difficulties for the first J-1 items, where J is the total number of items. Then the item difficulty for the $J^{th}$ item was defined as the negative sum of the other items within a given class.

# CHAPTER 4

# RESULTS

The present study uses three chains to estimate the posterior distributions for the data. Since the sampling results come from a Markov Chain they will be correlated to some small extent even after the burn-in process. Multiple chains are one way to further ensure we have uncorrelated samples providing more confidence in the posterior credibility intervals that are estimated (Kass, et.al., 1998). Another way to ensure this is through examining autocorrelations which will be described next. The initial step for running MCMC using WinBUGS is to determine the burn-in, the number of iterations needed to reach convergence. These are then discarded and posterior estimates are based on subsequent iterations. The present study uses two methods for assessing this convergence, the Brooks-Gelman-Rubin (BGR) diagnostic plots and the autocorrelation function plots. Following these tests a burn-in of 4,000 was determined to be appropriate for this model. The BGR diagnostic is based on the ratio of between to within chain variances. For the BGR graphs, shown in Figures 7 and 8, the width of the central 80% interval of the pooled runs is green, the average width of the 80% intervals within the individual runs is blue, and their ratio *R* (= pooled / within) is red. The pooled and within interval widths are normalized to have an overall maximum of one for plotting purposes (Spiegelhalter, Thomas, Best, & Lunn, 2003). Figures 7 and 8 are examples of the BGR graphs for this study. Figure 7 provides a view of the data after only 500 iterations where the chains have not yet converged. BGR graphs of the central 80% interval after a sufficient burn-in of 4,000 are presented in Figure

8. Convergence using the BGR graphs is confirmed by observing that *R* or the red line goes to one and the blue and green lines, the pooled and individual runs, converge to stability.

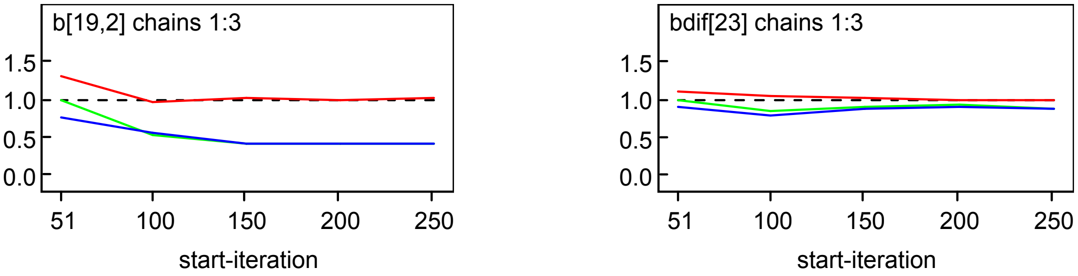

*Figure 7*. Representative BGR graphs of central 80% interval after 500 iterations

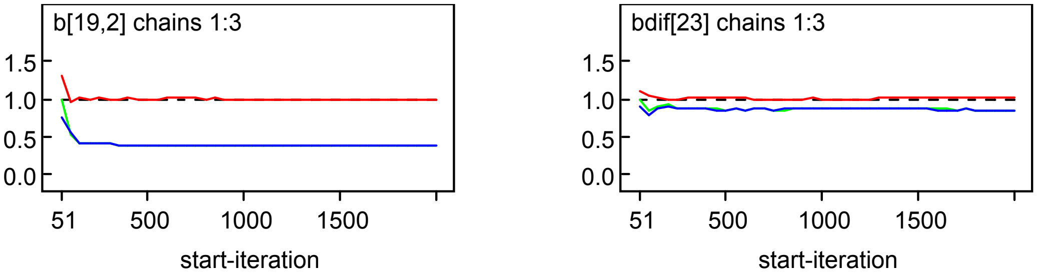

*Figure 8*. Representative BGR graphs of central 80% interval after discarding 4,000 burn-in

The second piece of information, the autocorrelation function plots, describes the correlation between the processes at different points in time. WinBUGS plots the autocorrelation function out to a lag-50. Graphically this looks like a histogram that is peaked at the beginning and almost a flat line at the end. This implies that the samples are uncorrelated. Figure 9 provides a view of the autocorrelation graphs after 500 iterations. Compare this to Figure 10 which presents the autocorrelation graphs after an adequate 4,000 burn-in.

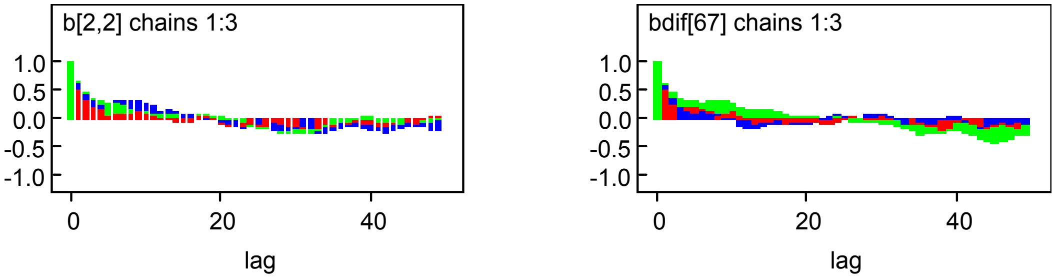

*Figure 9*. Representative autocorrelation graphs after 500 iterations

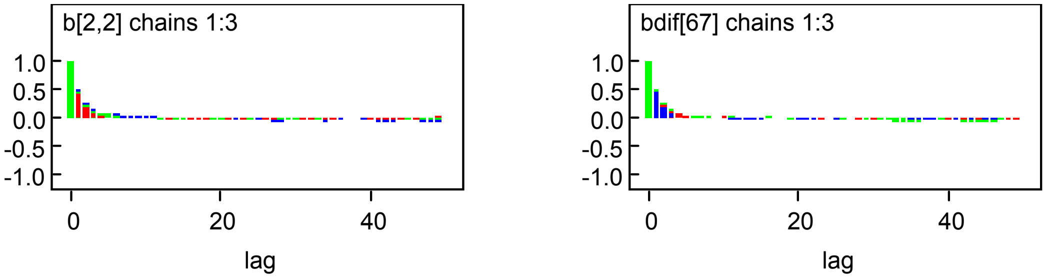

*Figure 10*. Representative autocorrelation graphs after 4,000 burn-in

Next, an additional number of iterations must be run in order to obtain posterior distributions from which inferences can be made. This assessment was made by examining kernel density plots for complete smoothness. Smoothness infers that all three chains have combined to form a posterior distribution of the sample. It was determined that 11,000 additional iterations were necessary for this to happen. That is 33,000 iterations in total for the three chains. Figure 11 provides four of the kernel density plots obtained from this model. Additionally, history plots were examined as an added means to assess that stationarity was obtained. Graphically this is determined by the extension of each colored line corresponding to each chain fully extending the length of the graph, demonstrating that the chains are fully mixed. Figure 12 presents representatives of these graphs.

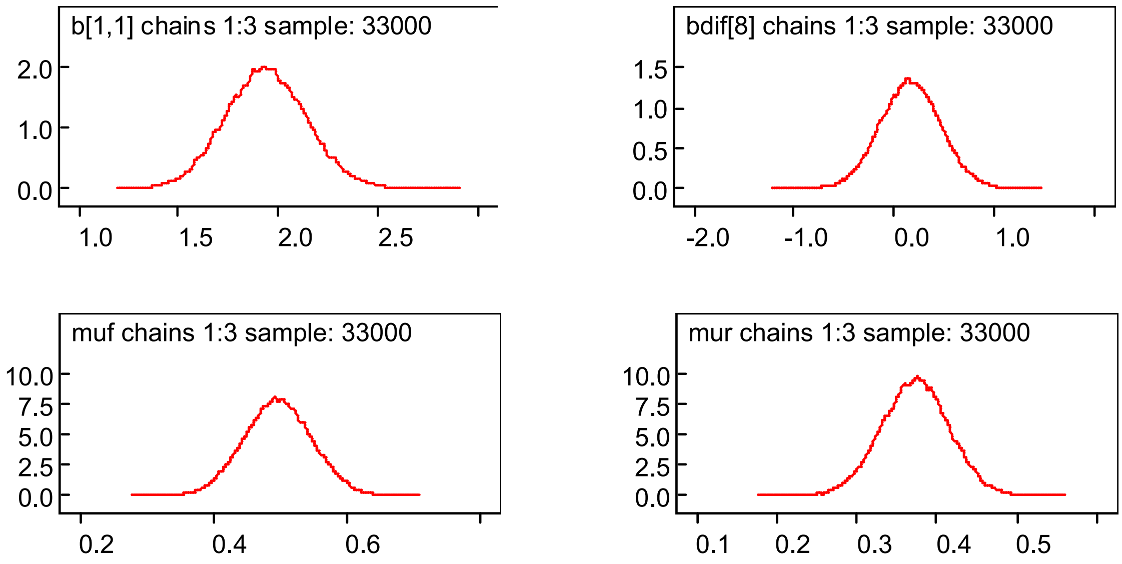

*Figure 11*. Representative kernel density plots.

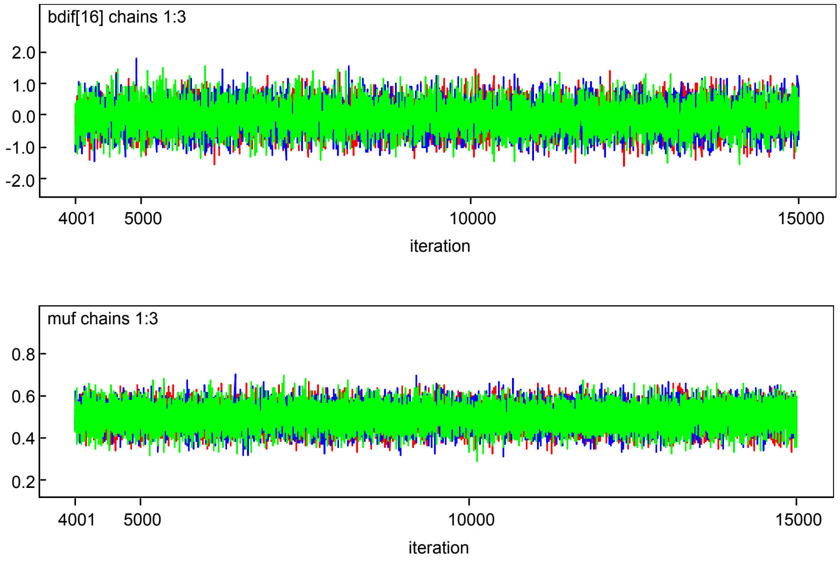

*Figure 12*. Representatives of history plots.

As suggested by Samuelsen and Bradshaw (2008), “Those items having 95% posterior credibility intervals of the differences between item difficulties not containing zero are identified as functioning differentially" (p. 2). This was determined to be 10% of the items which is in line with previous research (Rogers & Swaminathan, 1993; Zenisky, Hambelton, & Robin, 2003). The Credibility Interval Approach is modeled to generate *b's,* item difficulty parameters, for each item by gender grouping. Item difficulties for every item on the test are provided in Appendix B. As was previously noted, *bdif*, the gender group differences between item difficulties, are calculated for each item by subtracting the mean item difficulty for the focal group from the mean item difficulty of the reference group. To assess DIF the 95% posterior credibility intervals for the items are observed to see if they do not include zero. This interval corresponds to the 2.50% and 97.50% columns in Table 5. All items in Table 5 are differentially functioning, therefore the corresponding 95% credibility intervals do not include zero. Statistics for every item can be found in Appendix B, where differentially functioning items are in bold print. Table 5 also shows which gender group was advantaged for each differentially functioning item. Advantage is determined by establishing which gender group had the largest difficulty parameter for the observed item. For example, males are advantaged for item 9 in Table 5 because their mean item difficulty was smaller than the females mean item difficulty. In other words, the item was easier for males than for females.

Table 5.

*Items identified as having DIF*

| Node | mean | Sd | MC error | 2.50% | median | 97.50% | Advantaged |
|---|---|---|---|---|---|---|---|
| bdif[9] | 0.7126 | 0.2873 | 0.002209 | 0.1541 | 0.7135 | 1.279 | Male |
| bdif[41] | -0.5863 | 0.2944 | 0.002215 | -1.166 | -0.5866 | -0.007427 | female |
| bdif[54] | 0.7494 | 0.2995 | 0.002518 | 0.1637 | 0.7462 | 1.341 | Male |
| bdif[57] | 0.754 | 0.3069 | 0.002514 | 0.1536 | 0.7523 | 1.355 | Male |
| bdif[62] | -0.7437 | 0.3232 | 0.00282 | -1.378 | -0.7387 | -0.1149 | female |
| bdif[73] | -0.5716 | 0.2835 | 0.002202 | -1.127 | -0.5732 | -0.02043 | female |
| bdif[84] | 0.3759 | 0.1855 | 0.001258 | 0.01505 | 0.3765 | 0.7402 | Male |
| bdif[85] | 0.4545 | 0.1832 | 0.001224 | 0.09453 | 0.4559 | 0.8124 | Male |
| bdif[88] | 0.7473 | 0.2726 | 0.002029 | 0.2082 | 0.7472 | 1.28 | Male |
| bdif[98] | -0.5432 | 0.2682 | 0.002039 | -1.071 | -0.5425 | -0.01754 | female |
| bdif[106] | 0.6864 | 0.3284 | 0.002729 | 0.04291 | 0.6864 | 1.337 | Male |

*Note.* Node is the item of interest. Mean is the mean of the differences between the difficulties of the gender groups. MC error is an estimate of the Monte Carlo standard error of the mean, $\sigma/N^{1/2}$. SD is the sample standard deviation. 2.50%, median and 97.50% are the quantiles for the node.

DIF ranges from 0.3759 to 0.754 for the items. Of the 11 items with DIF four advantage females and seven advantage males showing DIF on this test was not consistent for males or females. Table 6 summarizes descriptive information about the items with DIF including the corresponding concept grouping, the item type, the forms where they reside, as well as the total number of respondents who were presented with the item.

Table 6.

*Item characteristics of differentially functioning items*

| **Item** | **Concept** | **Item Type** | **Forms** | **Respondents** |
|---|---|---|---|---|
| v009 | 3 | numerical answer | 1, 5, 9, 13, 17, 21, 25, 29, 33, 37 | 329 |
| v041 | 10 | numerical answer | 1, 6, 11, 16, 21, 26, 31, 36 | 256 |
| v054 | 13 | multiple choice | 5, 10, 15, 20, 25, 30, 35 | 253 |
| v057 | 13 | multiple choice | 3, 8, 13, 18, 23, 28, 33 | 225 |
| v062 | 14 | multiple answer | 3, 8, 13, 23, 28, 33 | 189 |
| v073 | 16 | numerical answer | 4, 9, 14, 19, 24, 29, 34 | 246 |
| v084 | 19 | multiple choice | 2, 4, 6, 8, 10, 12, 14, 16, 18, 20, 22, 24, 26, 28, 30, 32, 34, 36 | 610 |
| v085 | 19 | multiple choice | 1, 3, 5, 7, 9, 11, 13, 15, 17, 19, 21, 23, 25, 27, 29, 31, 33, 35, 37 | 628 |
| v088 | 20 | numerical answer | 2, 6, 10, 14, 18, 22, 26, 30, 34 | 304 |
| v098 | 22 | numerical answer | 2, 6, 10, 14, 18, 22, 26, 30, 34 | 304 |
| v106 | 25 | numerical answer | 5, 10, 15, 20, 25, 30, 35 | 253 |

The four items, v041, v062, v073, and v098 that advantaged females came from four different concept groupings, 10, 14, 16, and 22. These four items were of two item types, numerical answer and multiple answer. There do not appear to be any commonalities between these items. A total of seven items, v009, v054, v057, v084, v085, v088, and v006 with DIF were found to advantage males. The item types varied just as with the items that advantaged females. However, there were commonalties found within concept groupings. Items v054 and v057 are both part of concept grouping 13. Items v084 and v085 are both part of concept grouping 19. Concept group 13 consisted of mixed mass-volume problems while concept group 19 contained a set of questions on sequential reactions. Concept group 13 contains a total of five items meaning 40% of the items in this concept group have DIF. Concept group 19 has two items both of which exhibit DIF. Thus, DIF not only appears more frequently within certain concept groupings, the group that is consistently advantaged for those items that present DIF is males. Now let's take a moment and look at a parameter estimates for a concept group that did not contain any differentially functioning items.

Table 7.

*Concept Grouping 1*

| Item | Parameter Estimate |
|---|---|
| v001 | 1.993 |
| v002 | 1.819 |
| v003 | 1.989 |
| v004 | 1.765 |

Concept grouping 1 contains four items, shown in Table 7 along with corresponding parameter estimates. Notice that the parameter estimates are all very similar within the concept. Items within a concept are closely related and seek to measure the same idea. Thus, if DIF is not presented, the difficulty of the items within a given concept would be expected to be quite alike, as is seen for concept group 1. When DIF occurs for one of these items, however, this is no longer the case.

All 37 test forms contain between one and five differentially functioning items, meaning between 4% and 20% of the items on each form were differentially functioning. Items v054 and v106 appeared together on the following seven test forms; 5, 10, 15, 20, 25, 30, and 35. These test forms were found to have 4, 5, 3, 3, 4, 5, and 3 differentially functioning items respectively. Items v054 and v106 were found to advantage males. Also items v088 and v098 appear together on these 9 forms; 2, 6, 10, 14, 18, 22, 26, 30, and 34. These test forms were found to have 3, 4, 5, 4, 5, 3, 4, 5, and 4 differentially functioning items respectively. Recall that item v088 was shown to advantage males and item v098 to advantage females.

Item 84 has the smallest magnitude of DIF at 0.3759 compared to the other items with DIF. Item 57 presented with the largest amount of DIF. In their simulation, Samuelsen and Bradshaw (2008) demonstrated when the magnitude of DIF is 0.80, the presence of DIF will be more accurately detected. However, under the conditions modeled by Samuelsen and Bradshaw (2008), when the magnitude of DIF drops to 0.40 sufficient power cannot be reached. As the data under investigation in this study were modeled in a similar manner, the Samuelsen and Bradshaw (2008) results should hold. An important caveat to note is that the Samuelsen and Bradshaw (2008) study did not include missing data and they had more respondents for every item, 500 compared to 281 in the present study. None of the items presenting with DIF in this study exceed the magnitude of 0.80 and only one of the items, item 84, falls below 0.40. It is important to note that, although a parameter estimate might be statistically significant, it might not necessarily be meaningful, as there are what one might refer to as levels of DIF. In the case of Educational Testing Service (ETS), the magnitude of DIF is used to classify items into categories for the purpose of choosing the items to retain for use on operational tests (Zieky, 1993).

One would expect the ability distributions of the males and females to be the same. Both groups were provided with the same instructional materials, presented with the same assessment tool, and each received high school preparation. A t-test was conducted to determine if the mean ability levels of the males and females were statistically significantly different. This test was performed using the posterior means and standard deviations of the gender groups seen in Table 8. It was found that that the ability levels of the gender groups were statistically significantly different. There are factors that could contribute to this difference such as motivation or DIF. Differences in the means could be attributed to the fact that more items were found to favor the males. Further research would need to be done in order to assess further reasons for this difference.

Table 8.

*Ability Distributions*

| **Node** | **mean** | **sd** | **MC error** | **2.50%** | **median** | **97.50%** |
|---|---|---|---|---|---|---|
| Males | 0.4971 | 0.05074 | 3.77E-04 | 0.3984 | 0.497 | 0.5963 |
| Females | 0.373 | 0.04188 | 2.87E-04 | 0.2904 | 0.3734 | 0.454 |

*Note.* Node is the item of interest. Mean is the mean of the differences between the difficulties of the gender groups. MC error is an estimate of the Monte Carlo standard error of the mean, $\sigma/N^{1/2}$. SD is the sample standard deviation.  2.50%, median and 97.50% are the quantiles for the node.

# CHAPTER 5

# SUMMARY AND DISCUSSION

## Summary

The present study was designed to identify DIF on an undergraduate chemistry achievement test. Assessments are generally used to measure test takers knowledge in a particular area. This assessment was designed to evaluate chemistry knowledge of students being taught in a particular undergraduate course at the university. For security purposes, the test developers made use of multiple test forms. Although this data set contained a large number of respondents, 1,281, when analyzed by test form there were on average only 33 respondents for each test form. The complexity of these data required the use of a statistical procedure that would not be adversely affected by the small sample sizes. The Credibility Interval Approach, a more recent method for identifying DIF, was determined to be a potentially useful choice for this study. When analyzing the data in a matrix format the amount of missing data are very large, making some traditional methods of assessing DIF inappropriate for this study. MCMC techniques were useful because of the large amount of missingness in the data. The Credibility Interval Approach allowed for DIF to be analyzed given results from the Bayesian estimation algorithm used in this study.

The ability distributions for the males and females in this study were found to be different. One would expect that since both groups were provided with the same instructional materials and presented with the same assessment tool their ability distributions would not be different. Determining all of the reasons for this difference requires additional research outside of the present study. One reason for the difference in ability levels is the presence of DIF. Differences in the means could be attributed to the fact that more items were found to favor the males. Females were advantaged for 36% of the items with DIF and males for 64% of the items with DIF.

Items in two of the concept groupings, 13 and 19, exhibited DIF more frequently than others suggesting there may be some underlying reason the items in these concept groups were performing differently than the rest. Concept group 13 consisted of mixed mass-volume problems while concept group 19 contained a set of questions on sequential reactions. More concerning was that every item in concept group 19 exhibited DIF. Interestingly, males were found to be advantaged in both concept groupings 13 and 19.

As stated, concept group 13 consisted of items that asked examinees to determine mass, volume, and number of moles (molecular weight in grams) of several substances. These types of problems are mathematical and require an understanding of what the question is requesting. One such question in this concept group could be to determine the number of moles of the unknown, which would require among other things, students manipulating the following equation: $\frac{moles\ given}{coefficient\ of\ given} = \frac{moles\ unknown}{coefficient\ of\ unknown}$. DIF in this concept, as well as in concept 19, could be present due to a common misconception that might be shared by a group of students. Perhaps this misconception lies in the balancing of the equations that must take place or in the identification of the unknowns. Another thought might be that the wording of the item could be causing some confusion to students. This would have to be further explored by the test developers in order to better understand why DIF was more prevalent in a specific concept.

Consistencies with test forms were also present. Differentially functioning items v054 and v106 appeared together on these 7 forms; 5, 10, 15, 20, 25, 30, and 35, meaning every examinee that received item v054 also received item v106. Males were advantaged for each of these items. Consistencies were also found with items v088 and v098, which appeared together on these 9 forms; 2, 6, 10, 14, 18, 22, 26, 30, and 34. In this case males were advantaged for item v088 and females for item v098.

DIF presented on this test serves as a threat to the overall validity of the assessment. The fact that DIF was present for a percentage of items on this test signifies that performance on those items was probably dependent upon something other than solely the ability that was intended to be measured. This second dimension is what Ackerman (1992) refers to as nuisance abilities. What encompasses the nuisance abilities is unknown but the mere presence of such nuisance abilities infers a lack of construct validity. This evidence suggests that the assessment contains superfluous variance that is associated with other constructs. Reasons for this occurrence are outside of the realm of this study but the test developers and instructors in the chemistry department at the institution where the test was administered may have more insight into this result.

## Discussion

This study fills two gaps in the present research. First it applies the method of DIF analysis in chemistry education, an area where it has not routinely been employed. The creators of this particular achievement test can utilize the results provided here to assess the content and wording of the items said to have DIF. These items can be further analyzed by test developers for content and wording. In particular, consistencies by concept grouping and commonalities with test forms can provide more information for developers. Since DIF was found to be more prevalent in two concept areas, these results could be utilized as a teaching tool. Currently, the items with DIF are not measuring solely the intended construct. If the items from common concept groupings are thought to be well written and closely related to the construct then the gap may lie in the classroom. It was found that some differentially functioning items reside on the same group of test forms and thus were administered to the same examinees. This commonality could be further analyzed by looking for similarities between the examinees who were presented with the common forms and determining if there were any previously unrecognized differences in those examinees testing environments.

Secondly, this research uses a fairly new technique as a means to assess DIF. Previous research by Samuelsen and Bradshaw (2008) used simulated data to lend soundness to the use of this approach. This study takes what they have presented and puts it to the test. Applying this technique to real data for the first time lends strength to the credibility interval approach. The Rasch model was utilized to perform this analysis. If other models such as the 2PL and the 3PL are found to be viable the results from these models would be interesting to compare. Using other models will allow the estimation of the discrimination and pseudo-guessing parameters. This would allow for inferences to be made about how the items differentiate between examinees of varying ability levels and provide the probability of obtaining an answer by incorporating a guessing method.

As stated previously, Samuelsen and Bradshaw (2008) give evidence through a simulation study that the Credibility Interval Approach can recover DIF in situations where 10% and 30% of items are simulated to function differentially, test lengths are 20 or 40 items long, there are 500 respondents for each item, and groups are matched or unmatched on ability distributions. In their study, it was found that the power to detect DIF using this method was impacted by the magnitude of the differential function, as well as similarities or differences in the mean abilities of the reference and focal groups. However, there has yet to be an analysis that examines the detection of DIF using the Credibility Interval Approach when there is missing data. Despite this limitation, this method was employed in the present study based on its proven strength with detecting DIF in non-missing data situations. The fact that the present study resulted in DIF with a magnitude on average above 0.60 it can be inferred that the results are not unreasonable when compared to those found in the Samuelsen and Bradshaw (2008) study. Having majority DIF above the 0.40 level set forth by Samuelsen and Bradshaw (2008) combined with similar modeling conditions suggests that power might not be decreased in this study. If this is in fact the case, then the presence of DIF is more accurately detected as in the Samuelsen and Bradshaw (2008) study. In addition, the Credibility Interval Approach is a straightforward and quick way to detect DIF given the complexities of the data.

Based on this information, generalizations of this study should not be made until further analysis has been conducted asserting that the Credibility Interval Approach produces sufficient results when there is a large amount of missing data. Further work needs to be done to determine how missing data impacts the ability to detect DIF using this method, specifically, in situations where the data is missing intentionally, as in this study, as well as missing at random.

# REFERENCES

Ackerman, T. A. (1992). A didactic explanation of item bias, item impact, and item validity from a multidimensional perspective. *Journal of Educational Measurement*, *29(1),* 67-91.

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

Baker, F. (2001). The basics of item response theory. College Park, Maryland: ERIC Clearinghouse on Assessment and Evaluation, University of Maryland,

Bolt, D. M., Cohen, A. S., & Wollack, J. A. (2002). Item parameter estimation under conditions of test speededness: Application of a mixture Rasch model with ordinal constraints. *Journal of Educational Measurement, 39*, 331-348.

Clauser, B. E., & Mazor, K. M. (1998). Using Statistical Procedures to Identify Differentially Functioning Test Items. Instructional Module for the National Council on Measurement in Education, Spring 1998.

Clauser, B., Mazor, K., & Hambleton, R. K. (1993). The effects of purification of the matching criterion on the identification of DIF using the Mantel-Haenszel procedure. *Applied Measurement in Education, 6(4)*, 269-279.

Dorans, N. J., & Holland, P. W. (1993). DIF Detection and Description: Mantel-Haenszel and Standardization. In P.W. Holland & H. Wainer (Eds.) *Differential Item Functioning*. Hillsdale, N.J.: Lawrence Erlbaum.

Hambleton, R. K., & Rogers, H. J. (1989). Detecting potentially biased test item: Comparison of IRT area and Mantel-Haenszel methods. *Applied Measurement in Education, 2,* 313-334.

Holland, P. and Thayer, D. (1988) Differential item performance and the mantel haenszel procedure. In Wainer, H. and Braun, H., *Test validity* (105 – 128). Hillsdale, New Jersey: Lawrence Erlbaum Associates, Inc.

Kass, R. E., Carlin, B. P., Gelman, A., & Neal R. M. (1998). Markov chain monte carlo in practice: A roundtable discussion. *The American Statistician, 52(2),* 93-100.

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

Mazor, K. M., Clauser, B. E., & Hambleton, R. K. (1992). The effect of sample size on the functioning of the mantel-haenszel statistic. *Educational and Psychological Measurement, 52*, 443-451.

Rogers, H. J., & Swaminathan, H. (1993). A comparison of logistic regression and mantel-haenszel procedures for detecting differential item functioning. *Applied Psychological Measurement, 13(2)*, 105-116.

Samuelsen, K. & Bradshaw, L. (2008, March). The credibility interval method for the detection of DIF within a bayesian framework. Paper presented at the National Council on Measurement in Education annual conference, New York, NY.

Spiegelhalter, D., Thomas, A., Lunn, D. J. & Best, N. (2000). WinBUGS (Version 1.4.3) [computer program].

Spiegelhalter, D., Thomas, A., Best, N., & Lunn, D. (2003). *WinBUGS User's manual, version 1.4*. Last accessed on April 14, 2008 from the web-site: http://www.mrcbsu.cam.ac.uk/bugs.

Swaminathan, H., & Rogers, H. J. (1990). Detecting Differential Item Functioning Using

Logistic Regression Procedures. *Journal of Educational Measurement, 27*, 361-370.

Thissen, D., Steinberg, L., & Wainer, H. (1988). Use of item response theory in the study of group differences in trace lines. In H. Wainer & H. Braun (Eds.), *Test Validity*. Hillsdale, NJ: Erlbaum, pg. 147-169.

Thissen, D., Steinberg, L., & Wainer, H. (1993). Detection of Differential Item Functioning Using the Parameters of Item Response Models. In P.W. Holland & H. Wainer (Eds.) *Differential Item Functioning*. Hillsdale, NJ: Erlbaum, pg. 67-113.

Wainer, H. (1993). Model-Based Standardized Measurement of an Item's Differential Impact. In P.W. Holland & H. Wainer (Eds.) *Differential Item Functioning*. Hillsdale, N.J.: Lawrence Erlbaum.

Wollack, J. A., Cohen, A. S., & Wells, C. S. (2003). A method for maintaining scale stability in the presence of test speededness. *Journal of Educational Measurement, 40(4),* 307-330.

Zenisky, A. L., Hambelton, R. K., & Robin, F. (2003). Detection of differential item functioning in large-scale state assessments: A study evaluating a two-stage approach. *Educational and Psychological Measurement, 63(1),* 51-64.

Zieky, M. (1993). Practical Questions in the Use of DIF Statistics in Test Development. In P.W. Holland & H. Wainer (Eds.) *Differential Item Functioning*. Hillsdale, N.J.: Lawrence Erlbaum.

Zumbo, B. D. (2007). Three generations of dif analyses: Considering where it has been, where it is now, and where it is going. *Language Assessment Quarterly, 4(2),* 223-233.

# APPENDICES

## A. WinBUGS Code

```
model credibility interval approach
{
mur ~ dnorm(0,1);                # Mean ability for reference group
muf ~ dnorm(0,1);                # Mean ability for focal group
for (i in 1:N){
thetar[i] ~ dnorm(mur,1);        #Theta for the reference group is set to distribute normally with a
mean equal to the reference group mean and a variance of 1
thetaf[i] ~ dnorm(muf,1);        #Theta for the focal group is set to distribute normally with a mean
equal to the focal group mean and a variance of 1
}
# x[i,111] is a dichotomous variable referring to the grouping variable (1=reference, 0=focal)

#A loop that generates the item difficulties for each group
for (j in 1:J-1){                #J= total number of items
for (k in 1:2){                  #k = number of groups
b[j,k] ~ dnorm(0,1);             # Prior for the item parameter in the latent classes
}
bdif[j]<-b[j,1]-b[j,2];          #Calculation for the difference between item difficulties of the
reference and focal groups
}
b[J,1] <- -1*sum(b[1:(J-1),1]);         # Item difficulties for reference group sum to zero
b[J,2] <- -1*sum(b[1:(J-1),2]);         # Item difficulties for reference group sum to zero
bdif[J] <- b[J,1]-b[J,2];

# Rasch model which triggers manifest group membership
for (i in 1:N){
for (j in 1:J){
#Calculation of numerator in Rasch model.
numer[i,j] <- exp(x[i,111]*(thetar[i]-b[j,1])+(1-x[i,111])*(thetaf[i]-b[j,2]));
#Calculation of denominator in Rasch model
denom[i,j] <- 1+ exp(x[i,111]*(thetar[i]-b[j,1])+(1-x[i,111])*(thetaf[i]-b[j,2]));
#Item responses for examinees are set to distribute as a Bernoulli distribution
p[i,j] <- numer[i,j]/denom[i,j];
x[i,j] ~ dbern(p[i,j]);
}}
}
```

### B. WinBUGS Output

| node | mean | sd | MC error | 2.50% | median | 97.50% |
|---|---|---|---|---|---|---|
| b[1,1] | 1.94 | 0.2038 | 0.001655 | 1.549 | 1.937 | 2.349 |
| b[1,2] | 2.046 | 0.2241 | 0.001636 | 1.618 | 2.042 | 2.499 |
| b[2,1] | 1.626 | 0.1837 | 0.001482 | 1.269 | 1.623 | 1.992 |
| b[2,2] | 2.012 | 0.2177 | 0.001661 | 1.594 | 2.009 | 2.445 |
| b[3,1] | 1.95 | 0.1812 | 0.001318 | 1.6 | 1.948 | 2.309 |
| b[3,2] | 2.028 | 0.2566 | 0.002244 | 1.537 | 2.025 | 2.544 |
| b[4,1] | 1.99 | 0.1966 | 0.00153 | 1.61 | 1.987 | 2.382 |
| b[4,2] | 1.54 | 0.2192 | 0.001684 | 1.118 | 1.537 | 1.978 |
| b[5,1] | 3.171 | 0.2348 | 0.001983 | 2.729 | 3.165 | 3.653 |
| b[5,2] | 2.969 | 0.2708 | 0.002281 | 2.459 | 2.961 | 3.527 |
| b[6,1] | 3.501 | 0.2608 | 0.002405 | 3.012 | 3.496 | 4.029 |
| b[6,2] | 3.111 | 0.2598 | 0.002273 | 2.616 | 3.104 | 3.636 |
| b[7,1] | 2.898 | 0.2276 | 0.00184 | 2.469 | 2.893 | 3.356 |
| b[7,2] | 2.33 | 0.215 | 0.001555 | 1.918 | 2.327 | 2.768 |
| b[8,1] | 2.023 | 0.2086 | 0.001572 | 1.625 | 2.018 | 2.44 |
| b[8,2] | 1.86 | 0.2165 | 0.001735 | 1.44 | 1.857 | 2.292 |
| **b[9,1]** | **2.192** | **0.2112** | **0.001696** | **1.787** | **2.188** | **2.622** |
| **b[9,2]** | **1.479** | **0.1943** | **0.001492** | **1.101** | **1.48** | **1.862** |
| b[10,1] | 1.984 | 0.1831 | 0.001453 | 1.632 | 1.981 | 2.35 |
| b[10,2] | 2.026 | 0.2551 | 0.002127 | 1.538 | 2.021 | 2.537 |
| b[11,1] | 2.069 | 0.2024 | 0.001531 | 1.683 | 2.065 | 2.478 |
| b[11,2] | 2.032 | 0.237 | 0.001825 | 1.576 | 2.028 | 2.505 |
| b[12,1] | -0.7325 | 0.1839 | 0.001392 | -1.096 | -0.7308 | -0.3746 |
| b[12,2] | -0.5924 | 0.205 | 0.001489 | -1.004 | -0.591 | -0.1974 |
| b[13,1] | -0.6416 | 0.1745 | 0.001341 | -0.9873 | -0.64 | -0.3031 |
| b[13,2] | -0.9027 | 0.2042 | 0.001493 | -1.311 | -0.9009 | -0.5096 |
| b[14,1] | -0.4179 | 0.1637 | 0.001235 | -0.7425 | -0.4185 | -0.1005 |
| b[14,2] | -0.4445 | 0.2248 | 0.001843 | -0.8863 | -0.4452 | -0.00439 |
| b[15,1] | -0.8548 | 0.1836 | 0.001369 | -1.219 | -0.852 | -0.4999 |
| b[15,2] | -0.9183 | 0.2321 | 0.001896 | -1.381 | -0.9143 | -0.4707 |

*Note.* Items displaying DIF are in bold font. Node is the item of interest. Mean is the mean of the differences between the difficulties of the gender groups. MC error is an estimate of the Monte Carlo standard error of the mean, $\sigma/N^{1/2}$. SD is the sample standard deviation. 2.50%, median and 97.50% are the quantiles for the node.

| node | mean | sd | MC error | 2.50% | median | 97.50% |
|---|---|---|---|---|---|---|
| b[16,1] | -1.828 | 0.2424 | 0.002201 | -2.32 | -1.822 | -1.372 |
| b[16,2] | -1.753 | 0.3287 | 0.003264 | -2.423 | -1.742 | -1.131 |
| b[17,1] | -2.544 | 0.324 | 0.003476 | -3.208 | -2.536 | -1.938 |
| b[17,2] | -2.445 | 0.3535 | 0.003437 | -3.177 | -2.431 | -1.794 |
| b[18,1] | -2.404 | 0.2923 | 0.002841 | -3.004 | -2.395 | -1.862 |
| b[18,2] | -2.266 | 0.3599 | 0.003667 | -3.013 | -2.251 | -1.604 |
| b[19,1] | -2.975 | 0.3816 | 0.004517 | -3.77 | -2.958 | -2.28 |
| b[19,2] | -2.77 | 0.409 | 0.004262 | -3.626 | -2.748 | -2.027 |
| b[20,1] | -2.626 | 0.3285 | 0.003634 | -3.306 | -2.612 | -2.018 |
| b[20,2] | -2.421 | 0.3721 | 0.004015 | -3.2 | -2.404 | -1.741 |
| b[21,1] | -1.715 | 0.2361 | 0.002255 | -2.184 | -1.711 | -1.266 |
| b[21,2] | -1.37 | 0.2981 | 0.002863 | -1.978 | -1.362 | -0.8038 |
| b[22,1] | -1.006 | 0.2083 | 0.001672 | -1.416 | -1.004 | -0.6068 |
| b[22,2] | -1.12 | 0.2486 | 0.002064 | -1.621 | -1.116 | -0.6422 |
| b[23,1] | -0.3831 | 0.1815 | 0.001357 | -0.7375 | -0.3818 | -0.03132 |
| b[23,2] | -0.2279 | 0.2313 | 0.001893 | -0.6849 | -0.2273 | 0.2235 |
| b[24,1] | -0.2347 | 0.2031 | 0.001678 | -0.6333 | -0.2337 | 0.1617 |
| b[24,2] | -0.0536 | 0.216 | 0.001737 | -0.4804 | -0.0534 | 0.3639 |
| b[25,1] | -1.277 | 0.2222 | 0.001908 | -1.721 | -1.273 | -0.8486 |
| b[25,2] | -1.588 | 0.2918 | 0.002646 | -2.174 | -1.583 | -1.038 |
| b[26,1] | 0.03763 | 0.17 | 0.001325 | -0.2962 | 0.03802 | 0.3677 |
| b[26,2] | -0.3204 | 0.197 | 0.001415 | -0.709 | -0.3181 | 0.06635 |
| b[27,1] | -0.3032 | 0.1677 | 0.001268 | -0.6321 | -0.3011 | 0.02355 |
| b[27,2] | -0.0988 | 0.1856 | 0.00133 | -0.4645 | -0.0977 | 0.2621 |
| b[28,1] | 0.3357 | 0.1554 | 0.001169 | 0.02687 | 0.3365 | 0.6367 |
| b[28,2] | 0.5617 | 0.2156 | 0.001634 | 0.1393 | 0.5611 | 0.9885 |
| b[29,1] | 0.6201 | 0.1655 | 0.001316 | 0.2986 | 0.6198 | 0.9484 |
| b[29,2] | 0.8134 | 0.2037 | 0.001542 | 0.4191 | 0.8142 | 1.216 |
| b[30,1] | -1.15 | 0.1981 | 0.00154 | -1.548 | -1.147 | -0.7706 |
| b[30,2] | -1.032 | 0.2205 | 0.001734 | -1.474 | -1.03 | -0.6075 |
| b[31,1] | -0.6717 | 0.1758 | 0.001322 | -1.019 | -0.6705 | -0.3321 |
| b[31,2] | -0.7004 | 0.1996 | 0.001534 | -1.101 | -0.6981 | -0.318 |
| b[32,1] | -1.814 | 0.225 | 0.001969 | -2.27 | -1.809 | -1.386 |

*Note.* Items displaying DIF are in bold font. Node is the item of interest. Mean is the mean of the differences between the difficulties of the gender groups. MC error is an estimate of the Monte Carlo standard error of the mean, $\sigma/N^{1/2}$. SD is the sample standard deviation. 2.50%, median and 97.50% are the quantiles for the node.

| node | mean | sd | MC error | 2.50% | median | 97.50% |
|---|---|---|---|---|---|---|
| b[32,2] | -1.75 | 0.2968 | 0.002772 | -2.357 | -1.742 | -1.195 |
| b[33,1] | -1.39 | 0.2037 | 0.001687 | -1.8 | -1.386 | -1.001 |
| b[33,2] | -1.783 | 0.2941 | 0.002681 | -2.379 | -1.775 | -1.227 |
| b[34,1] | 0.7768 | 0.1914 | 0.001496 | 0.4038 | 0.7752 | 1.153 |
| b[34,2] | 0.8674 | 0.2429 | 0.002016 | 0.3987 | 0.8668 | 1.351 |
| b[35,1] | 1.24 | 0.208 | 0.001659 | 0.834 | 1.238 | 1.65 |
| b[35,2] | 1.541 | 0.2357 | 0.001866 | 1.085 | 1.538 | 2.012 |
| b[36,1] | -0.6131 | 0.2271 | 0.001993 | -1.062 | -0.6094 | -0.1761 |
| b[36,2] | -0.1158 | 0.2529 | 0.002198 | -0.6174 | -0.1146 | 0.3797 |
| b[37,1] | 0.769 | 0.2038 | 0.001611 | 0.3718 | 0.7686 | 1.169 |
| b[37,2] | 0.7057 | 0.245 | 0.00204 | 0.228 | 0.7037 | 1.196 |
| b[38,1] | 1.033 | 0.2007 | 0.00158 | 0.6426 | 1.032 | 1.428 |
| b[38,2] | 1.561 | 0.2642 | 0.002341 | 1.054 | 1.557 | 2.087 |
| b[39,1] | 0.9627 | 0.2073 | 0.001675 | 0.5617 | 0.9636 | 1.374 |
| b[39,2] | 1.318 | 0.2451 | 0.002006 | 0.8422 | 1.316 | 1.801 |
| b[40,1] | -0.8409 | 0.195 | 0.001531 | -1.225 | -0.8393 | -0.4642 |
| b[40,2] | -0.8426 | 0.2604 | 0.002176 | -1.36 | -0.8409 | -0.3387 |
| **b[41,1]** | **-0.6915** | **0.1986** | **0.001517** | **-1.087** | **-0.6892** | **-0.3086** |
| **b[41,2]** | **-0.1052** | **0.2169** | **0.001668** | **-0.5322** | **-0.1038** | **0.318** |
| b[42,1] | -0.6438 | 0.1871 | 0.001389 | -1.014 | -0.6421 | -0.2803 |
| b[42,2] | -0.6624 | 0.2433 | 0.001957 | -1.15 | -0.66 | -0.1922 |
| b[43,1] | -1.235 | 0.23 | 0.001986 | -1.696 | -1.232 | -0.7901 |
| b[43,2] | -1.131 | 0.2527 | 0.002221 | -1.639 | -1.128 | -0.6475 |
| b[44,1] | -1.096 | 0.2154 | 0.001723 | -1.53 | -1.093 | -0.6835 |
| b[44,2] | -0.9001 | 0.2459 | 0.002022 | -1.393 | -0.8982 | -0.4243 |
| b[45,1] | -2.432 | 0.2808 | 0.00261 | -3.001 | -2.424 | -1.903 |
| b[45,2] | -2 | 0.2838 | 0.002347 | -2.581 | -1.992 | -1.464 |
| b[46,1] | -0.884 | 0.1797 | 0.001348 | -1.241 | -0.882 | -0.5344 |
| b[46,2] | -1.21 | 0.2192 | 0.001623 | -1.65 | -1.207 | -0.7906 |
| b[47,1] | -1.092 | 0.186 | 0.001428 | -1.461 | -1.09 | -0.7303 |
| b[47,2] | -1.226 | 0.2573 | 0.002121 | -1.746 | -1.22 | -0.7382 |
| b[48,1] | -2.191 | 0.2541 | 0.002394 | -2.698 | -2.185 | -1.706 |
| b[48,2] | -1.62 | 0.2772 | 0.002317 | -2.18 | -1.613 | -1.092 |

*Note.* Items displaying DIF are in bold font. Node is the item of interest. Mean is the mean of the differences between the difficulties of the gender groups. MC error is an estimate of the Monte Carlo standard error of the mean, $\sigma/N^{1/2}$. SD is the sample standard deviation. 2.50%, median and 97.50% are the quantiles for the node.

| node | mean | sd | MC error | 2.50% | median | 97.50% |
|---|---|---|---|---|---|---|
| b[49,1] | 1.597 | 0.1912 | 0.001559 | 1.23 | 1.595 | 1.978 |
| b[49,2] | 1.321 | 0.2388 | 0.001948 | 0.8604 | 1.32 | 1.796 |
| b[50,1] | 1.499 | 0.1958 | 0.001496 | 1.123 | 1.498 | 1.886 |
| b[50,2] | 1.989 | 0.2559 | 0.002066 | 1.496 | 1.987 | 2.498 |
| b[51,1] | 1.672 | 0.1947 | 0.001623 | 1.296 | 1.669 | 2.059 |
| b[51,2] | 1.247 | 0.2329 | 0.00196 | 0.7983 | 1.244 | 1.714 |
| b[52,1] | 1.345 | 0.2218 | 0.001949 | 0.9192 | 1.343 | 1.783 |
| b[52,2] | 1.697 | 0.235 | 0.001929 | 1.246 | 1.694 | 2.17 |
| b[53,1] | 1.568 | 0.2053 | 0.001694 | 1.173 | 1.563 | 1.979 |
| b[53,2] | 1.365 | 0.2246 | 0.001675 | 0.9257 | 1.363 | 1.814 |
| **b[54,1]** | **1.595** | **0.192** | **0.001502** | **1.221** | **1.595** | **1.972** |
| **b[54,2]** | **0.8455** | **0.2303** | **0.001834** | **0.3963** | **0.8456** | **1.301** |
| b[55,1] | 1.319 | 0.1878 | 0.00147 | 0.9554 | 1.318 | 1.688 |
| b[55,2] | 0.9285 | 0.2194 | 0.001665 | 0.5061 | 0.927 | 1.365 |
| b[56,1] | 1.861 | 0.2011 | 0.00164 | 1.474 | 1.857 | 2.265 |
| b[56,2] | 1.576 | 0.2412 | 0.002057 | 1.112 | 1.574 | 2.051 |
| **b[57,1]** | **1.346** | **0.2231** | **0.001932** | **0.9128** | **1.345** | **1.791** |
| **b[57,2]** | **0.5924** | **0.2101** | **0.001622** | **0.1794** | **0.5936** | **1.008** |
| b[58,1] | 1.65 | 0.2085 | 0.001596 | 1.251 | 1.647 | 2.066 |
| b[58,2] | 1.902 | 0.2494 | 0.002067 | 1.428 | 1.897 | 2.402 |
| b[59,1] | 0.7657 | 0.1769 | 0.001267 | 0.4216 | 0.7653 | 1.116 |
| b[59,2] | 0.8445 | 0.2308 | 0.001844 | 0.3909 | 0.8428 | 1.299 |
| b[60,1] | 1.284 | 0.1885 | 0.001425 | 0.9178 | 1.281 | 1.66 |
| b[60,2] | 1.415 | 0.2318 | 0.001783 | 0.9654 | 1.413 | 1.876 |
| b[61,1] | 0.7675 | 0.1751 | 0.001359 | 0.4205 | 0.7683 | 1.11 |
| b[61,2] | 0.9435 | 0.2267 | 0.001709 | 0.5012 | 0.9419 | 1.391 |
| **b[62,1]** | **1.118** | **0.2132** | **0.001801** | **0.7062** | **1.116** | **1.541** |
| **b[62,2]** | **1.861** | **0.243** | **0.00207** | **1.395** | **1.858** | **2.349** |
| b[63,1] | 1.487 | 0.202 | 0.001705 | 1.093 | 1.486 | 1.885 |
| b[63,2] | 1.672 | 0.2362 | 0.002015 | 1.215 | 1.669 | 2.145 |
| b[64,1] | 0.3556 | 0.1752 | 0.001331 | 0.01244 | 0.356 | 0.7003 |
| b[64,2] | 0.1823 | 0.2311 | 0.001852 | -0.2714 | 0.1836 | 0.6328 |

*Note.* Items displaying DIF are in bold font. Node is the item of interest. Mean is the mean of the differences between the difficulties of the gender groups. MC error is an estimate of the Monte Carlo standard error of the mean, $\sigma/N^{1/2}$. SD is the sample standard deviation. 2.50%, median and 97.50% are the quantiles for the node.

| node | mean | sd | MC error | 2.50% | median | 97.50% |
|---|---|---|---|---|---|---|
| b[65,1] | 0.5152 | 0.18 | 0.001386 | 0.1618 | 0.5153 | 0.8649 |
| b[65,2] | 0.07906 | 0.2153 | 0.001563 | -0.3469 | 0.079 | 0.4969 |
| b[66,1] | 0.4788 | 0.1752 | 0.001329 | 0.135 | 0.48 | 0.8218 |
| b[66,2] | 0.0721 | 0.2254 | 0.001787 | -0.3736 | 0.0738 | 0.5155 |
| b[67,1] | 0.1518 | 0.1991 | 0.001624 | -0.2367 | 0.1512 | 0.5419 |
| b[67,2] | 0.1685 | 0.2123 | 0.001648 | -0.2491 | 0.1688 | 0.5859 |
| b[68,1] | 0.06356 | 0.1859 | 0.001416 | -0.307 | 0.06418 | 0.4263 |
| b[68,2] | 0.05128 | 0.2161 | 0.001628 | -0.3771 | 0.05283 | 0.4724 |
| b[69,1] | 0.4409 | 0.1742 | 0.001333 | 0.1004 | 0.4409 | 0.7815 |
| b[69,2] | 0.5912 | 0.2289 | 0.001676 | 0.1391 | 0.5915 | 1.043 |
| b[70,1] | 0.4838 | 0.1794 | 0.001353 | 0.134 | 0.4833 | 0.8353 |
| b[70,2] | 0.2983 | 0.2129 | 0.001767 | -0.1194 | 0.2977 | 0.7177 |
| b[71,1] | 0.477 | 0.1734 | 0.00129 | 0.1393 | 0.4769 | 0.8168 |
| b[71,2] | 0.9941 | 0.2292 | 0.001934 | 0.5482 | 0.9942 | 1.445 |
| b[72,1] | 0.2262 | 0.2004 | 0.001574 | -0.165 | 0.2269 | 0.6183 |
| b[72,2] | 0.2964 | 0.2107 | 0.001504 | -0.1173 | 0.2967 | 0.7102 |
| **b[73,1]** | **0.101** | **0.1883** | **0.00144** | **-0.2693** | **0.1005** | **0.4723** |
| **b[73,2]** | **0.6726** | **0.2136** | **0.001676** | **0.2581** | **0.6722** | **1.097** |
| b[74,1] | -1.052 | 0.2178 | 0.001851 | -1.491 | -1.047 | -0.6368 |
| b[74,2] | -0.9843 | 0.2716 | 0.002448 | -1.529 | -0.9785 | -0.4666 |
| b[75,1] | -1.657 | 0.2536 | 0.002244 | -2.174 | -1.652 | -1.181 |
| b[75,2] | -1.741 | 0.2943 | 0.002713 | -2.342 | -1.732 | -1.185 |
| b[76,1] | -0.1355 | 0.2117 | 0.001651 | -0.552 | -0.1346 | 0.2751 |
| b[76,2] | -0.3131 | 0.2552 | 0.002145 | -0.8261 | -0.3108 | 0.1797 |
| b[77,1] | -1.431 | 0.2535 | 0.002209 | -1.943 | -1.425 | -0.9456 |
| b[77,2] | -1.296 | 0.3107 | 0.003045 | -1.93 | -1.29 | -0.7028 |
| b[78,1] | -1.447 | 0.2472 | 0.002143 | -1.944 | -1.442 | -0.976 |
| b[78,2] | -1.744 | 0.3286 | 0.003291 | -2.418 | -1.734 | -1.129 |
| b[79,1] | -1.221 | 0.235 | 0.002072 | -1.691 | -1.217 | -0.765 |
| b[79,2] | -0.768 | 0.2651 | 0.002141 | -1.297 | -0.7633 | -0.2577 |
| b[80,1] | -0.6046 | 0.179 | 0.001447 | -0.9597 | -0.6035 | -0.258 |
| b[80,2] | -0.5147 | 0.2024 | 0.001511 | -0.9187 | -0.513 | -0.1239 |

*Note.* Items displaying DIF are in bold font. Node is the item of interest. Mean is the mean of the differences between the difficulties of the gender groups. MC error is an estimate of the Monte Carlo standard error of the mean, $\sigma/N^{1/2}$. SD is the sample standard deviation. 2.50%, median and 97.50% are the quantiles for the node.

| node | mean | sd | MC error | 2.50% | median | 97.50% |
|---|---|---|---|---|---|---|
| b[81,1] | -0.9173 | 0.1816 | 0.001443 | -1.28 | -0.9142 | -0.5675 |
| b[81,2] | -0.4772 | 0.1948 | 0.001417 | -0.8657 | -0.4744 | -0.09991 |
| b[82,1] | -0.5812 | 0.1691 | 0.00126 | -0.9138 | -0.5803 | -0.2524 |
| b[82,2] | -1.102 | 0.2544 | 0.001966 | -1.61 | -1.098 | -0.6193 |
| b[83,1] | -1.901 | 0.233 | 0.001973 | -2.375 | -1.896 | -1.457 |
| b[83,2] | -1.865 | 0.2998 | 0.002836 | -2.469 | -1.858 | -1.304 |
| **b[84,1]** | **0.4314** | **0.1153** | **8.03E-04** | **0.2061** | **0.432** | **0.6575** |
| **b[84,2]** | **0.05558** | **0.1451** | **9.49E-04** | **-0.2266** | **0.05544** | **0.3412** |
| **b[85,1]** | **0.4402** | **0.1179** | **8.55E-04** | **0.2101** | **0.4395** | **0.6729** |
| **b[85,2]** | **-0.0143** | **0.14** | **9.67E-04** | **-0.29** | **-0.0136** | **0.2565** |
| b[86,1] | 0.6465 | 0.1704 | 0.001244 | 0.3122 | 0.6466 | 0.982 |
| b[86,2] | 0.2756 | 0.1882 | 0.001366 | -0.0945 | 0.2745 | 0.6364 |
| b[87,1] | 1.247 | 0.1755 | 0.001259 | 0.9054 | 1.248 | 1.592 |
| b[87,2] | 0.8921 | 0.1841 | 0.001367 | 0.534 | 0.8911 | 1.255 |
| **b[88,1]** | **0.5904** | **0.1564** | **0.001146** | **0.2828** | **0.5914** | **0.8981** |
| **b[88,2]** | **-0.1569** | **0.22** | **0.001682** | **-0.5917** | **-0.1575** | **0.2715** |
| b[89,1] | 0.8556 | 0.166 | 0.001219 | 0.5355 | 0.8551 | 1.182 |
| b[89,2] | 0.3754 | 0.2045 | 0.001497 | -0.0259 | 0.3754 | 0.7738 |
| b[90,1] | -1.142 | 0.2192 | 0.001809 | -1.579 | -1.138 | -0.7219 |
| b[90,2] | -0.9814 | 0.2727 | 0.002294 | -1.525 | -0.9779 | -0.4571 |
| b[91,1] | -1.31 | 0.2344 | 0.001881 | -1.779 | -1.307 | -0.8618 |
| b[91,2] | -1.74 | 0.2941 | 0.00292 | -2.338 | -1.733 | -1.184 |
| b[92,1] | -1.293 | 0.2575 | 0.002515 | -1.812 | -1.288 | -0.8046 |
| b[92,2] | -0.6436 | 0.2655 | 0.00216 | -1.175 | -0.6383 | -0.1371 |
| b[93,1] | -1.136 | 0.2386 | 0.002165 | -1.621 | -1.129 | -0.6861 |
| b[93,2] | -1.201 | 0.2986 | 0.002794 | -1.809 | -1.193 | -0.6413 |
| b[94,1] | -0.8674 | 0.2192 | 0.001658 | -1.306 | -0.8646 | -0.4482 |
| b[94,2] | -0.6673 | 0.2636 | 0.00218 | -1.19 | -0.6648 | -0.1617 |
| b[95,1] | -1.327 | 0.2394 | 0.002145 | -1.811 | -1.321 | -0.871 |
| b[95,2] | -1.304 | 0.2991 | 0.002732 | -1.905 | -1.299 | -0.7356 |
| b[96,1] | -0.01812 | 0.1708 | 0.001253 | -0.3532 | -0.0179 | 0.3167 |
| b[96,2] | -0.0349 | 0.1912 | 0.001414 | -0.4122 | -0.03298 | 0.3387 |

*Note.* Items displaying DIF are in bold font. Node is the item of interest. Mean is the mean of the differences between the difficulties of the gender groups. MC error is an estimate of the Monte Carlo standard error of the mean, $\sigma/N^{1/2}$. SD is the sample standard deviation. 2.50%, median and 97.50% are the quantiles for the node.

| node | mean | sd | MC error | 2.50% | median | 97.50% |
|---|---|---|---|---|---|---|
| b[97,1] | 0.5855 | 0.1641 | 0.001157 | 0.2655 | 0.5858 | 0.9091 |
| b[97,2] | 0.2259 | 0.1814 | 0.001322 | -0.1316 | 0.2269 | 0.5793 |
| **b[98,1]** | **-0.06801** | **0.1604** | **0.001165** | **-0.3862** | **-0.06719** | **0.2438** |
| **b[98,2]** | **0.4752** | **0.2142** | **0.001684** | **0.05573** | **0.4756** | **0.8941** |
| b[99,1] | 1.101 | 0.1711 | 0.00133 | 0.7706 | 1.1 | 1.441 |
| b[99,2] | 1.271 | 0.2115 | 0.001524 | 0.8651 | 1.271 | 1.69 |
| b[100,1] | 0.001585 | 0.08259 | 5.12E-04 | -0.1617 | 0.001932 | 0.1622 |
| b[100,2] | -0.01899 | 0.1011 | 6.72E-04 | -0.2181 | -0.01847 | 0.1788 |
| b[101,1] | -1.766 | 0.2387 | 0.002157 | -2.25 | -1.762 | -1.312 |
| b[101,2] | -1.372 | 0.2959 | 0.002872 | -1.966 | -1.366 | -0.8081 |
| b[102,1] | -0.6165 | 0.1933 | 0.001448 | -1.003 | -0.6145 | -0.2393 |
| b[102,2] | -0.1933 | 0.2179 | 0.001805 | -0.6236 | -0.1916 | 0.2294 |
| b[103,1] | -0.7837 | 0.192 | 0.001501 | -1.166 | -0.7802 | -0.4149 |
| b[103,2] | -0.5469 | 0.2392 | 0.001865 | -1.023 | -0.5455 | -0.08393 |
| b[104,1] | 0.9462 | 0.2099 | 0.001838 | 0.542 | 0.9443 | 1.363 |
| b[104,2] | 0.9818 | 0.2137 | 0.001694 | 0.5695 | 0.9811 | 1.405 |
| b[105,1] | -1.432 | 0.2309 | 0.002087 | -1.89 | -1.428 | -0.9855 |
| b[105,2] | -1.286 | 0.267 | 0.002377 | -1.826 | -1.28 | -0.7769 |
| **b[106,1]** | **-0.3658** | **0.1818** | **0.001451** | **-0.7268** | **-0.3651** | **-0.01289** |
| **b[106,2]** | **-1.052** | **0.2723** | **0.002445** | **-1.6** | **-1.047** | **-0.5291** |
| b[107,1] | -0.4719 | 0.189 | 0.00146 | -0.8459 | -0.4703 | -0.1047 |
| b[107,2] | -0.7815 | 0.2341 | 0.001968 | -1.247 | -0.7777 | -0.3366 |
| b[108,1] | -0.5128 | 0.1841 | 0.001448 | -0.8793 | -0.5104 | -0.1585 |
| b[108,2] | -0.2816 | 0.2306 | 0.001819 | -0.7347 | -0.2806 | 0.1706 |
| b[109,1] | -0.5128 | 0.2055 | 0.001611 | -0.919 | -0.5118 | -0.1123 |
| b[109,2] | -0.6758 | 0.2345 | 0.001893 | -1.148 | -0.6718 | -0.224 |
| b[110,1] | -0.875 | 0.2096 | 5.44E-04 | -1.292 | -0.8728 | -0.4704 |
| b[110,2] | -0.8331 | 0.2489 | 6.06E-04 | -1.331 | -0.8303 | -0.3555 |
| bdif[1] | -0.1063 | 0.3044 | 0.002378 | -0.7029 | -0.1058 | 0.4873 |
| bdif[2] | -0.3857 | 0.2849 | 0.002237 | -0.9531 | -0.3833 | 0.1707 |
| bdif[3] | -0.0784 | 0.3155 | 0.00268 | -0.7101 | -0.0752 | 0.5344 |
| bdif[4] | 0.4502 | 0.2953 | 0.002308 | -0.1308 | 0.4525 | 1.023 |
| bdif[5] | 0.2015 | 0.3611 | 0.003084 | -0.5101 | 0.2022 | 0.9115 |

*Note.* Items displaying DIF are in bold font. Node is the item of interest. Mean is the mean of the differences between the difficulties of the gender groups. MC error is an estimate of the Monte Carlo standard error of the mean, $\sigma/N^{1/2}$. SD is the sample standard deviation. 2.50%, median and 97.50% are the quantiles for the node.

| node | mean | sd | MC error | 2.50% | median | 97.50% |
|---|---|---|---|---|---|---|
| bdif[6] | 0.3906 | 0.3693 | 0.003244 | -0.3332 | 0.393 | 1.117 |
| bdif[7] | 0.5682 | 0.3133 | 0.002362 | -0.0379 | 0.5658 | 1.188 |
| bdif[8] | 0.1635 | 0.3012 | 0.002371 | -0.424 | 0.1623 | 0.7579 |
| **bdif[9]** | **0.7126** | **0.2873** | **0.002209** | **0.1541** | **0.7135** | **1.279** |
| bdif[10] | -0.0423 | 0.3134 | 0.002602 | -0.661 | -0.0409 | 0.5685 |
| bdif[11] | 0.03705 | 0.3129 | 0.002217 | -0.5764 | 0.03705 | 0.6492 |
| bdif[12] | -0.1401 | 0.276 | 0.002009 | -0.6764 | -0.1403 | 0.4018 |
| bdif[13] | 0.2612 | 0.2686 | 0.001946 | -0.2633 | 0.2599 | 0.7882 |
| bdif[14] | 0.0266 | 0.2786 | 0.002178 | -0.524 | 0.02718 | 0.5726 |
| bdif[15] | 0.06354 | 0.2965 | 0.002348 | -0.5186 | 0.06442 | 0.6519 |
| bdif[16] | -0.0752 | 0.41 | 0.003948 | -0.8688 | -0.0790 | 0.7392 |
| bdif[17] | -0.0993 | 0.4819 | 0.005114 | -1.042 | -0.1014 | 0.853 |
| bdif[18] | -0.1383 | 0.4662 | 0.004711 | -1.044 | -0.1387 | 0.7912 |
| bdif[19] | -0.2054 | 0.5598 | 0.006202 | -1.292 | -0.2094 | 0.904 |
| bdif[20] | -0.2054 | 0.4938 | 0.005328 | -1.165 | -0.2064 | 0.7772 |
| bdif[21] | -0.3455 | 0.3799 | 0.003733 | -1.08 | -0.3501 | 0.4009 |
| bdif[22] | 0.1141 | 0.3244 | 0.002667 | -0.5166 | 0.1137 | 0.7538 |
| bdif[23] | -0.1552 | 0.2938 | 0.002237 | -0.7299 | -0.1564 | 0.4245 |
| bdif[24] | -0.1811 | 0.2965 | 0.002417 | -0.7628 | -0.1814 | 0.4019 |
| bdif[25] | 0.3109 | 0.3687 | 0.00324 | -0.4001 | 0.3084 | 1.043 |
| bdif[26] | 0.358 | 0.261 | 0.001945 | -0.1534 | 0.357 | 0.8746 |
| bdif[27] | -0.2043 | 0.2494 | 0.001846 | -0.6882 | -0.205 | 0.2875 |
| bdif[28] | -0.226 | 0.266 | 0.001943 | -0.7513 | -0.2239 | 0.2936 |
| bdif[29] | -0.1934 | 0.2631 | 0.001974 | -0.715 | -0.1943 | 0.3202 |
| bdif[30] | -0.1177 | 0.2958 | 0.002259 | -0.7006 | -0.1191 | 0.4665 |
| bdif[31] | 0.02874 | 0.267 | 0.001992 | -0.4911 | 0.02658 | 0.5558 |
| bdif[32] | -0.06369 | 0.3736 | 0.003455 | -0.7777 | -0.06786 | 0.6871 |
| bdif[33] | 0.3931 | 0.3566 | 0.003124 | -0.2941 | 0.39 | 1.104 |
| bdif[34] | -0.09065 | 0.3097 | 0.002535 | -0.6996 | -0.09144 | 0.517 |
| bdif[35] | -0.3003 | 0.3119 | 0.002491 | -0.9131 | -0.2991 | 0.3076 |
| bdif[36] | -0.4973 | 0.3389 | 0.002997 | -1.165 | -0.497 | 0.1661 |
| bdif[37] | 0.06339 | 0.316 | 0.002565 | -0.5619 | 0.06513 | 0.6834 |

*Note.* Items displaying DIF are in bold font. Node is the item of interest. Mean is the mean of the differences between the difficulties of the gender groups. MC error is an estimate of the Monte Carlo standard error of the mean, σ/N$^{1/2}$. SD is the sample standard deviation. 2.50%, median and 97.50% are the quantiles for the node.

| node | mean | sd | MC error | 2.50% | median | 97.50% |
|---|---|---|---|---|---|---|
| bdif[38] | -0.5285 | 0.3333 | 0.002835 | -1.192 | -0.5245 | 0.1144 |
| bdif[39] | -0.3556 | 0.3213 | 0.00252 | -0.9881 | -0.356 | 0.2758 |
| bdif[40] | 0.001683 | 0.3252 | 0.002682 | -0.6357 | 0.002643 | 0.6367 |
| **bdif[41]** | **-0.5863** | **0.2944** | **0.002215** | **-1.166** | **-0.5866** | **-0.00742** |
| bdif[42] | 0.01859 | 0.3072 | 0.002475 | -0.5853 | 0.01636 | 0.6253 |
| bdif[43] | -0.1036 | 0.3419 | 0.003027 | -0.7745 | -0.1041 | 0.5686 |
| bdif[44] | -0.1959 | 0.3251 | 0.002584 | -0.8377 | -0.1982 | 0.4412 |
| bdif[45] | -0.4326 | 0.3984 | 0.003434 | -1.21 | -0.4324 | 0.3569 |
| bdif[46] | 0.3261 | 0.2836 | 0.002182 | -0.2254 | 0.3245 | 0.8826 |
| bdif[47] | 0.1341 | 0.3183 | 0.002534 | -0.4795 | 0.1283 | 0.7732 |
| bdif[48] | -0.5711 | 0.3798 | 0.003308 | -1.316 | -0.5748 | 0.1835 |
| bdif[49] | 0.2757 | 0.3056 | 0.002415 | -0.3264 | 0.2767 | 0.8717 |
| bdif[50] | -0.4892 | 0.3228 | 0.002685 | -1.127 | -0.488 | 0.1452 |
| bdif[51] | 0.4253 | 0.3039 | 0.002551 | -0.1755 | 0.4261 | 1.017 |
| bdif[52] | -0.3525 | 0.3241 | 0.002781 | -0.9912 | -0.3524 | 0.2807 |
| bdif[53] | 0.2028 | 0.3045 | 0.002466 | -0.3882 | 0.201 | 0.8114 |
| **bdif[54]** | **0.7494** | **0.2995** | **0.002518** | **0.1637** | **0.7462** | **1.341** |
| bdif[55] | 0.3906 | 0.2888 | 0.002218 | -0.1749 | 0.3918 | 0.9542 |
| bdif[56] | 0.2848 | 0.3123 | 0.002694 | -0.3314 | 0.2848 | 0.8934 |
| **bdif[57]** | **0.754** | **0.3069** | **0.002514** | **0.1536** | **0.7523** | **1.355** |
| bdif[58] | -0.2516 | 0.3245 | 0.002708 | -0.8914 | -0.2515 | 0.382 |
| bdif[59] | -0.0787 | 0.2895 | 0.002271 | -0.6417 | -0.0787 | 0.491 |
| bdif[60] | -0.1317 | 0.2992 | 0.002245 | -0.718 | -0.1325 | 0.4506 |
| bdif[61] | -0.1761 | 0.2855 | 0.00224 | -0.7372 | -0.1749 | 0.3847 |
| **bdif[62]** | **-0.7437** | **0.3232** | **0.00282** | **-1.378** | **-0.7387** | **-0.1149** |
| bdif[63] | -0.1857 | 0.3103 | 0.00275 | -0.8011 | -0.1858 | 0.4188 |
| bdif[64] | 0.1733 | 0.2883 | 0.002185 | -0.3921 | 0.1719 | 0.7382 |
| bdif[65] | 0.4362 | 0.2808 | 0.002095 | -0.1119 | 0.4352 | 0.9922 |
| bdif[66] | 0.4066 | 0.2854 | 0.002276 | -0.15 | 0.4043 | 0.9719 |
| bdif[67] | -0.0166 | 0.291 | 0.002333 | -0.5871 | -0.0168 | 0.562 |
| bdif[68] | 0.01228 | 0.286 | 0.002209 | -0.5471 | 0.0121 | 0.5779 |
| bdif[69] | -0.1503 | 0.2878 | 0.002073 | -0.7138 | -0.1479 | 0.4113 |
| bdif[70] | 0.1855 | 0.2778 | 0.002192 | -0.3641 | 0.1872 | 0.7297 |

*Note.* Items displaying DIF are in bold font. Node is the item of interest. Mean is the mean of the differences between the difficulties of the gender groups. MC error is an estimate of the Monte Carlo standard error of the mean, $\sigma/N^{1/2}$. SD is the sample standard deviation. 2.50%, median and 97.50% are the quantiles for the node.

| **node** | **mean** | **sd** | **MC error** | **2.50%** | **median** | **97.50%** |
|---|---|---|---|---|---|---|
| bdif[71] | -0.5171 | 0.2869 | 0.00229 | -1.084 | -0.5163 | 0.04141 |
| bdif[72] | -0.0701 | 0.2921 | 0.002232 | -0.6424 | -0.0709 | 0.5016 |
| **bdif[73]** | **-0.5716** | **0.2835** | **0.002202** | **-1.127** | **-0.5732** | **-0.02043** |
| bdif[74] | -0.0675 | 0.35 | 0.003155 | -0.743 | -0.0684 | 0.6163 |
| bdif[75] | 0.08443 | 0.3893 | 0.003574 | -0.6776 | 0.08383 | 0.8539 |
| bdif[76] | 0.1776 | 0.3307 | 0.002732 | -0.4677 | 0.1762 | 0.8303 |
| bdif[77] | -0.1347 | 0.4034 | 0.003611 | -0.9184 | -0.1357 | 0.668 |
| bdif[78] | 0.2972 | 0.4088 | 0.003924 | -0.4924 | 0.292 | 1.116 |
| bdif[79] | -0.4526 | 0.3536 | 0.003061 | -1.142 | -0.4527 | 0.2435 |
| bdif[80] | -0.0898 | 0.2699 | 0.002078 | -0.6258 | -0.0895 | 0.4344 |
| bdif[81] | -0.4401 | 0.2662 | 0.001981 | -0.9585 | -0.4406 | 0.08956 |
| bdif[82] | 0.5213 | 0.3053 | 0.002317 | -0.0695 | 0.52 | 1.125 |
| bdif[83] | -0.0352 | 0.3789 | 0.003298 | -0.7711 | -0.0365 | 0.7135 |
| **bdif[84]** | **0.3759** | **0.1855** | **0.001258** | **0.01505** | **0.3765** | **0.7402** |
| **bdif[85]** | **0.4545** | **0.1832** | **0.001224** | **0.09453** | **0.4559** | **0.8124** |
| bdif[86] | 0.3709 | 0.2532 | 0.001705 | -0.1205 | 0.3692 | 0.8679 |
| bdif[87] | 0.3551 | 0.2536 | 0.001897 | -0.1449 | 0.356 | 0.8528 |
| **bdif[88]** | **0.7473** | **0.2726** | **0.002029** | **0.2082** | **0.7472** | **1.28** |
| bdif[89] | 0.4802 | 0.2637 | 0.001941 | -0.0389 | 0.4811 | 0.9934 |
| bdif[90] | -0.1605 | 0.3502 | 0.002885 | -0.8403 | -0.1611 | 0.5332 |
| bdif[91] | 0.4295 | 0.3752 | 0.00342 | -0.2978 | 0.4248 | 1.169 |
| bdif[92] | -0.6491 | 0.3697 | 0.003301 | -1.381 | -0.6462 | 0.0672 |
| bdif[93] | 0.06515 | 0.3841 | 0.003531 | -0.678 | 0.06422 | 0.829 |
| bdif[94] | -0.2 | 0.3432 | 0.002731 | -0.8635 | -0.1999 | 0.4784 |
| bdif[95] | -0.0236 | 0.3827 | 0.003475 | -0.7693 | -0.0248 | 0.7284 |
| bdif[96] | 0.01678 | 0.2557 | 0.001801 | -0.4852 | 0.01594 | 0.5201 |
| bdif[97] | 0.3596 | 0.2457 | 0.001699 | -0.1219 | 0.3589 | 0.8433 |
| **bdif[98]** | **-0.5432** | **0.2682** | **0.002039** | **-1.071** | **-0.5425** | **-0.01754** |
| bdif[99] | -0.1699 | 0.2724 | 0.001972 | -0.7048 | -0.1689 | 0.3679 |
| bdif[100] | 0.02057 | 0.1309 | 8.38E-04 | -0.2384 | 0.02138 | 0.2782 |
| bdif[101] | -0.3934 | 0.3804 | 0.003472 | -1.131 | -0.3958 | 0.3649 |
| bdif[102] | -0.4232 | 0.2915 | 0.002267 | -0.9963 | -0.4212 | 0.1449 |
| bdif[103] | -0.2368 | 0.3069 | 0.002564 | -0.8303 | -0.2374 | 0.3648 |

*Note.* Items displaying DIF are in bold font. Node is the item of interest. Mean is the mean of the differences between the difficulties of the gender groups. MC error is an estimate of the Monte Carlo standard error of the mean, $\sigma/N^{1/2}$. SD is the sample standard deviation. 2.50%, median and 97.50% are the quantiles for the node.

| **node** | **mean** | **Sd** | **MC error** | **2.50%** | **median** | **97.50%** |
|---|---|---|---|---|---|---|
| bdif[104] | -0.0356 | 0.2984 | 0.00261 | -0.6184 | -0.0366 | 0.5496 |
| bdif[105] | -0.146 | 0.3518 | 0.003124 | -0.835 | -0.1475 | 0.5517 |
| bdif[106] | 0.6864 | 0.3284 | 0.002729 | 0.04291 | 0.6864 | 1.337 |
| bdif[107] | 0.3096 | 0.3001 | 0.002494 | -0.2761 | 0.3088 | 0.905 |
| bdif[108] | -0.2312 | 0.2941 | 0.002362 | -0.8142 | -0.2315 | 0.3473 |
| bdif[109] | 0.1631 | 0.3118 | 0.00249 | -0.4484 | 0.1629 | 0.7778 |
| bdif[110] | -0.0419 | 0.325 | 7.97E-04 | -0.6799 | -0.0414 | 0.5989 |
| muf | 0.4971 | 0.05074 | 3.77E-04 | 0.3984 | 0.497 | 0.5963 |
| mur | 0.373 | 0.04188 | 2.87E-04 | 0.2904 | 0.3734 | 0.454 |

*Note.* Items displaying DIF are in bold font. Node is the item of interest. Mean is the mean of the differences between the difficulties of the gender groups. MC error is an estimate of the Monte Carlo standard error of the mean, $\sigma/N^{1/2}$. SD is the sample standard deviation. 2.50%, median and 97.50% are the quantiles for the node.