5-7-2011

Improvements for Differential Functioning of Items and Tests (DFIT): Investigating the Addition of Reporting an Effect Size Measure and Power

Keith D. Wright
kwright8@att.net

Recommended Citation

Wright, Keith D., "Improvements for Differential Functioning of Items and Tests (DFIT): Investigating the Addition of Reporting an Effect Size Measure and Power" (2011). *Educational Policy Studies Dissertations.* Paper 80.
http://digitalarchive.gsu.edu/eps_diss/80

ACCEPTANCE

This dissertation, IMPROVEMENTS FOR DIFFERENTIAL FUNCTIONING OF ITEMS AND TESTS (DFIT): INVESTIGATING THE ADDITION OF REPORTING AN EFFECT SIZE MEASURE AND POWER, by KEITH DARNELL WRIGHT, was prepared under the direction of the candidate's Dissertation Advisory Committee. It is accepted by the committee members in partial fulfillment of the requirements for the degree Doctor of Philosophy in the College of Education, Georgia State University.

The Dissertation Advisory committee and the student's Department Chair, as representatives of the faculty, certify that this dissertation has met all standards of excellence and scholarship as determined by the faculty. The Dean of the College of Education concurs.

Chris T. Oshima, Ph.D.
Committee Chair

Phillip Gagne', Ph.D.
Committee Member

Raymond Hart, Ph.D.
Committee Member

Philo Hutcheson, Ph.D.
Committee Member

Date

Sheryl A. Gowen, Ph.D.
Chair, Department of Educational Policy Studies

R. W. Kamphaus, Ph.D.
Dean and Distinguished Research Professor
College of Education

AUTHOR'S STATEMENT

Keith Darnell Wright

NOTICE TO BORROWERS

All dissertations deposited in the Georgia State University library must be used in accordance with the stipulations prescribed by the author in the preceding statements. The author of this dissertation is:

Keith Darnell Wright
425 Sugar Gate Court
Lawrenceville, GA 30044

The director of this dissertation is:

Dr. Chris T. Oshima
Department of Educational Policy Studies
College of Education
Georgia State University
Atlanta, GA 30303 - 3083

CURRICULUM VITAE

Keith D. Wright

ADDRESS: 425 Sugar Gate Court
Lawrenceville, Georgia 30044

EDUCATION:

Ph.D.	2011	Georgia State University Educational Policy Studies Concentration: Research, Measurement and Statistics
MBA	1997	Loyola University of Chicago Finance/Marketing
MSCS	1992	Illinois Institute of Technology Computer Science
BSEET	1988	DeVry Institute of Technology Electronics Engineering Technology

PROFESSIONAL EXPERIENCE:

2006 – Present	Associate Dean – College of Engineering & Info. Sciences DeVry University, Atlanta, GA.
2005 – Present	Graduate Teaching Assistant Georgia State University, Atlanta, GA.
2005 – Present	Program Evaluator – Accredited Eng. Tech. Programs ABET, Baltimore, MD.
1999 – 2002	Sales Engineering Manager Cisco Systems, Atlanta, GA.
1997 – 1999	Account Marketing Manager Nortel Networks, Alpharetta, GA.
1992 – 1997	Lead Applications Engineer Tellabs, Lisle, IL.
1988 – 1992	Member of Technical Staff Bell Laboratories, Naperville, IL.

PROFESSIONAL SOCIETIES AND ORGANIZATIONS:

2006 – Present	American Educational Research Association (AERA)
2006 – Present	National Council on Measurement in Education (NCME)

PRESENTATIONS AND PUBLICATIONS:

Oshima, T.C, White, N., & Wright, K. (2010, July). *Differential item functioning among multiple groups using Raju's DFIT*. Paper presented at the 75th Annual Meeting of the Psychometric Society, Athens, GA.

Wright, K. & Oshima, T.C. (2011, April). *Improvements for DFIT: Investigating the addition of reporting an effect size measure, and reporting power.* Paper presented at the 2011 Annual Meeting of the American Educational Research Association, New Orleans, LA.

Oshima, T.C, White, N., & Wright, K. (2010). *Differential item functioning among multiple groups using Raju's DFIT*. Manuscript submitted for publication (copy on file with author).

ABSTRACT

IMPROVEMENTS FOR DIFFERENTIAL FUNCTIONING OF ITEMS AND TESTS (DFIT): INVESTIGATING THE ADDITION OF REPORTING AN EFFECT SIZE MEASURE AND POWER

by

Keith D. Wright

Standardized testing has been part of the American educational system for decades. Controversy from the beginning has plagued standardized testing, is plaguing testing today, and will continue to be controversial. Given the current federal educational policies supporting increased standardized testing, psychometricians, educators and policy makers must seek ways to ensure that tests are not biased towards one group over another.

In measurement theory, if a test item behaves differently for two different groups of examinees, this test item is considered a differential functioning test item (DIF). Differential item functioning, often conceptualized in the context of item response theory (IRT) is a term used to describe test items that may favor one group over another after matched on ability. It is important to determine whether an item is functioning significantly different for one group over another regardless as to why. Hypothesis testing is used to determine statistical significant DIF items; an effect size measure quantifies a statistical significant difference.

This study investigated the addition of reporting an effect size measure for differential item functioning of items and tests' (DFIT) noncompensatory differential item functioning (NCDIF), and reporting empirically observed power. The Mantel-Haenszel (MH) parameter served as the benchmark for developing NCDIF's effect size measure, for reporting moderate and large differential item functioning in test items. In addition, by modifying NCDIF's unique method for determining statistical significance, NCDIF will be the first DIF statistic of test items where in addition to reporting an effect size measure, empirical power can also be reported.

Furthermore, this study added substantially to the body of literature on effect size by also investigating the behavior of two other DIF measures, Simultaneous Item Bias Test (SIBTEST) and area measure. Finally, this study makes a significant contribution to the body of literature by verifying in a large-scale simulation study, the accuracy of software developed by Roussos, Schnipke, and Pashley (1999) to calculate the true MH parameter. The accuracy of this software had not been previously verified.

IMPROVEMENTS FOR DIFFERENTIAL FUNCTIONING OF ITEMS AND TESTS (DFIT): INVESTIGATING THE ADDITION OF REPORTING AN EFFECT SIZE MEASURE AND POWER

by

Keith D. Wright

A Dissertation

Presented in Partial Fulfillment of Requirements of the
Degree of
Doctor of Philosophy
in
Research, Measurement & Statistics
in
the Department of Educational Policy Studies
in
the College of Education
Georgia State University

Atlanta, GA
2011

ACKNOWLEDGEMENTS

I would like to thank first and foremost my Lord and Savior Jesus Christ. My successful completion of this journey would not have been possible without his blessings on a daily basis. It is only because of the mercy and grace of my Savior and my faith, which allowed me to achieve my ultimate educational goal. I realized the importance of education during the summer after my freshmen year of high school and I have never looked back. I give thanks to my mom (Shirley Gunn) and dad (Lorenzo Gunn) who have supported me throughout my educational journey.

I would like to also thank my committee members (Dr. Gagne', Dr. Hart, Dr. Hutcheson, and my Chair Dr. Oshima) for serving on my committee. They provided the unyielding guidance throughout my comprehensive exams, prospectus and dissertation defense. I would like to give a special thanks to Dr. Gagne' and Dr. Oshima, who I have learned a wealth of knowledge related to statistics and measurement theory throughout my time at Georgia State University. I would like to give another special thanks to Dr. Gowen who has given me a graduate teaching/research assistantship position every semester. I will be forever grateful to everyone I mentioned.

It is very important for me to thank my siblings given the unique role each has played in my life at one point or another. I come from a very big family, and needless to say I love each and every one of my siblings dearly. I would like to thank each in order of age from oldest to youngest, Byron, Crystal, Lorenzo (Bernard), Tamiko and my baby brother Gino.

Finally, and most importantly, I would like to thank three other very important individuals in my life. I thank Bernard my brother, my cousin Samad Jaliladdin, and my best friend in the whole world, Sheldon (Corey) Robinson. If it was not for these three individuals encouraging me every day and providing emotional and spiritual support, this dream would not have been realized. These three invidiudals played a pivotal role in my life while with me for many years in Atlanta, GA. It is because of me witnessing how Sheldon dealt with a life threatening diagnosis (brain tumor), his courage and will to live, which gave me the strength to overcome all the adversities I faced during this journey. Thank you very much Sheldon (Corey), you are my inspiration in which I will know that nothing in life is too difficult to overcome if you have faith and determination.

TABLE OF CONTENTS

Page

LIST OF TABLES

Table

Page

LIST OF FIGURES

Figure

Page

ABBREVIATIONS

DFIT	Differential Functioning of Items and Tests
DIF	Differential Item Functioning
ICC	Item Characteristic Curve
IRT	Item Response Theory
1PM	One Parameter Model
2PM	Two Parameter Model
3PM	Three Parameter Model

CHAPTER 1

INTRODUCTION

Standardized testing has been part of the American educational system for decades. Controversy from the beginning has plagued standardized testing, is plaguing testing today, and will plague testing in the future (Gallagher, 2003). In the words of Gallagher, educators today “face a dilemma” (p. 83). The dilemma is associated with the current legislation surrounding increased testing. Given the current federal educational policies supporting increased standardized testing (Hursh, 2008; Millsap & Everson, 1993), psychometricians, educators and policy makers must seek ways to ensure that tests are not biased towards one group over another.

Measurement in Testing

In the field of psychometrics, a test item which separates examinees based on the construct being measured is considered a highly discriminating test item. A test item which discriminates based on the construct being measured and not on personal characteristics (e.g. ethnicity) is desirable. This is considered item impact which is one purpose of testing. The opposite of item impact is item bias, where performance differences are not due to the test item’s construct, but based on group differences (e.g. ethnicity). Many of the standardized tests today are purported to measure a specific ability (Lord, 1980; Kok, 1988; Shealy & Stout, 1993; Ackerman, 1989; Oshima, Raju, & Flowers, 1997; Angoff, 1993). Theoretically as stated by Rudner, Getson, and Knight (1980), “…tests and test items are perfectly unidimensional, that is, an item measures only one ability and all items of a test measure the same ability” (p. 215).

The tenet of unidimensionality is theory-based because in practice, unidimensionality is difficult to attain (Rudner, Getson, & Knight, 1980). For a test measuring vocabulary using sentence completion test questions, this type of test item would require a strong vocabulary and also an understanding of complex sentence structures (Clauser & Mazor, 1998). If the test item purports to measure only vocabulary the primary ability being measured, and sentence structure comprehension is a secondary ability being measured, the test item may favor one group over another. If one group overall has a higher level of sentence structure comprehension, the other group could be at a disadvantage. In measurement theory, this item may be behaving differently for the two groups, hence, a differentially functioning test item (DIF).

Differential item functioning, often conceptualized in the context of item response theory (IRT), is a term used to describe test items that may favor one group over another after matched on ability. A lack of unidimensionality is just one factor that may be causing a test item to exhibit DIF. It is important to determine whether an item is functioning significantly different for one group over another regardless as to why. Hypothesis testing is used to determine statistical significant DIF items (Monahan, McHorney, Stump, & Perkins, 2007).

Statistical Significance versus Practical Significance

When hypothesis testing is conducted and a test item is flagged as significant, this test item is functioning differently for examinees being measured. Typically, when test items are categorized as DIF, test publishers may remove these test items from the test bank.

Constructing standardized tests is an arduous and costly process (Ramsey, 1993). A cost as described by Zieky (1993) is the fact that "…the decisions associated with DIF are likely to be scrutinized in the adversarial arenas of legislation and litigation" (p. 337). Given the laborious nature of test construction and its cost, flagging a test item based only on hypothesis testing is not sufficient evidence to remove the test item. An effect size measure can be used in conjunction with a significant finding, to determine if DIF is large enough to warrant removal of the test item (Cohen, 1988; Kirk, 1996; Hidalgo & Lopez, 2004; Monahan, et al., 2007).

Why use an effect size if an item exhibits statistically significant DIF? DIF statistical techniques require large sample sizes. It is well known, the larger the sample size, the higher the probability of yielding a statistical significant finding. Moreover, an insignificant finding with a small sample may have a meaningful effect size. Statistical significance does not guarantee practical significance; therefore, an effect size helps to quantify an insignificant finding with small samples, and a statistical significant finding with large samples.

The DFIT Framework

Understanding the DIF statistics available and their differences is important for policymakers, practitioners, and researchers. Standardized tests are used to make high-stake decisions and the score an examinee receives can have life changing implications. Research related to DIF can be seen in the literature as early as 1910 (Camilli & Shepard, 1994).

Since 1910 there have been numerous procedures related to the detection of differentially functioning test items (Clauser & Mazor, 1998; Camilli & Shepard, 1994; Shealy & Stout, 1993). But as stated by Clauser and Mazor, "…a relatively small number of these methods have emerged as preferred" (p. 32). Is one DIF method better than another? This is a difficult question to answer given the evolution of the DIF methods.

DIF methods in the beginning were designed to assess dichotomously scored test items. These methods have evolved whereby dichotomous and polytomous test items can be investigated. DIF methods today can also evaluate individual test items as well as the entire test. The methods today can investigate both uniform and non-uniform DIF. Finally, the violation of unidimensionality can be tested, that is testing for multidimensionality. The concept of unidimensionality is related to a test item measuring one ability; the concept of multidimensionality is related to a test item measuring more than one ability. A problem with the many DIF statistics is the specialty nature in which they were initially developed, that is one size does not fit all. The DFIT framework is a new and promising DIF statistic (Raju, 1988; Oshima & Morris, 2008; Osterlind & Everson, 2009).

The DFIT framework can be used for investigating, (a) dichotomous and polytomous test items; (b) individual test items along with the entire test; (c) uniform and non-uniform DIF; and (d) the presence of multidimensionality. Finally, most utilized DIF statistics report an effect size measure (Monahan et al, 2007). The DFIT framework currently does not employ an effect size measure. If DFIT is to continue to gain prominence among practitioners, an effect size measure is highly desirable.

Empirical Observed Power

DFIT's statistical significance test is a highly unique method. The test is called the item parameter replication (IPR) method (Oshima, Raju, & Nanda, 2006). The uniqueness of the IPR method is associated with as stated by Oshima and Morris (2008), "produces an empirical sampling distribution of NCDIF under the null hypothesis that focal and reference groups have identical parameters" (p. 47). If an empirical sampling distribution of NCDIF under the alternative hypothesis is determined, empirical power may be estimated. A DIF technique being able to report a statistical significance or lack of significance finding, with an effect size and power, is a matter of promoting excellent statistical practices (Kirk, 2001). DFIT would be the only DIF technique with this capability.

CHAPTER 2

REVIEW OF THE LITERATURE

DIF methods can be classified into one of two categories, parametric and nonparametric DIF procedures. The parametric category in the literature today is often referred to as item response theory. IRT methods employ explicit measurement models (e.g. 1PL, 2PL, 3PL, etc). Nonparametric procedures do not rely on specific measurement models for assessing DIF. These procedures are referred to in the literature as contingency table approaches or general non-IRT approaches (Camilli & Shepard, 1994). The most utilized nonparametric procedures are (a) Mantel-Haenszel (MH); (b) Standardization; (c) Logistic Regression; and (c) SIBTEST.

Mantel-Haenszel Procedure

In studying the likelihood of getting a disease based on factors that are present or not, the study of matched groups utilizing contingency tables was introduced by Mantel and Haenszel (1959). MH as a practical technique to determine if a test item is functioning different for two groups of examinees was first proposed by Holland (1985). Holland and Thayer (1988) provided the landmark study which explains in great detail the use of MH as a DIF technique.

MH is arguably the most widely used contingency table approach to studying DIF (Clauser & Mazor, 1998). The first step in using the MH approach is to setup a contingency table for each ability group. When analyzing a test item for DIF, it is important to group (i.e. match) examinees based on ability. Typically, total test score is used as the matching criteria to group examinees.

As an example, consider a test item being studied for DIF, whereby 1000 examinees hypothetically answered a test item which was part of a 40 item test. Furthermore, it has been determined to create four ability groups based on total test scores. The first group in this example could be those examinees who had a total test score between 0 – 10 correct, the second group had a total test score between 11 - 20 correct, the third group had a total test score between 21 – 30 correct, and the fourth group had a total test score between 31 – 40 correct. In this example, you would not want to compare those in the first group with any of the other three groups because based on total test score, their ability differs. The importance of matching examinees is a matter of comparing the comparables (Dorans & Holland, 1993). It would not make practical sense to study DIF for examinees with different abilities because this would not be DIF, but impact. As noted by Clauser and Mazor, "…examinees from different groups may in fact differ in ability, in which case differences in performance are to be expected" (p. 31).

The null hypothesis for the MH statistic states that the odds for the focal group answering the test item correctly is the same as the odds for the reference group. Conversely, the alternative hypothesis states that the odds for the focal group answering the test item correctly are not the same as the odds for the reference group. Equations 1 and 2 respectively represent the null and alternative hypotheses for the MH statistic.

$$H_0 : \frac{P_{rj}}{Q_{rj}} = \alpha \frac{P_{fj}}{Q_{fj}} \qquad j = 1, 2, 3, \ldots, k \qquad \alpha = 1 \qquad (1)$$

$$H_1 : \frac{P_{rj}}{Q_{rj}} = \alpha \frac{P_{fj}}{Q_{fj}} \qquad j = 1, 2, 3, \ldots, k \qquad \alpha \neq 1 \qquad (2)$$

In Table 1, A_j represents the total number of reference group examinees in jth group who answered the test item correctly. B_j represents the total number of reference group examinees in jth group who answered the test item incorrectly. N_{rj} represents the total number of reference group examinees for jth group, that is, A_j and B_j summed. Based on these values, Prj can be determined. Prj is the probability of answering the test item correctly, for a reference group examinee in the jth group. P_{rj} can be calculated by dividing A_j by N_{rj}. Q_{rj} is the probability of answering the test item incorrectly. Q_{rj} can be calculated by dividing B_j by N_{rj}. This is the same as 1 minus the probability of answering the test item correctly. C_j, D_j, P_{fj} and Q_{fj} represent focal group values, which are interpreted and calculated as described for the reference group.

The odds for the reference group answering the test item correctly divided by the odds for the focal group answering the test item correctly will be the odds ratio. Alpha (α) in Equation 1 is the odds ratio for the MH statistic, which measures the size of the difference between the reference group odds and the focal group odds. The cross-product of the odds ratio is given in Equation 3. Alpha (α) in Equation 1 and Equation 2 is equal to this cross-product.

$$\text{Odd Ratio Cross Product} = \frac{P_{rj}Q_{fj}}{P_{fj}Q_{rj}} \qquad j = 1, 2, 3, \ldots, k \qquad (3)$$

Table 1 *2x2 Contingency Table - Data for jth Ability Group*

	Test Item Score	1	0	Total
Reference Group		Aj (P_{rj})	Bj (Q_{rj})	N_{rj}
Focal Group		Cj (P_{fj})	Dj (Q_{fj})	N_{fj}
Total		M_{1j}	M_{oj}	T_j

When α is equal to 1, the odds for the focal group answering the test item correctly is the same as the odds for the reference group, hence, the null hypothesis. If the odds for the reference and focal groups are not the same, α ≠ 1. The value (i.e. effect size) of α indicates how much more likely (i.e. multiplicative) the odds for the reference group is for answering the test item correctly over the focal group. The equation in 4 estimates α,

$$\overset{\wedge}{\alpha}_{MH} = \frac{\sum_{j} A_j D_j / T_j}{\sum_{j} B_j C_j / T_j} \qquad (4)$$

The effect size α, is a value with a range from 0 to ∞, where a value of 1 specifies the absence of DIF (Dorans & Holland, 1993). Holland and Thayer (1988) modified $\overset{\wedge}{\alpha}_{MH}$, Equation 5, to make it easier to interpret for those familiar with the Educational Testing Service's (ETS) delta metric for item difficulty. In making an odds ratio (i.e. α) easier to interpret, the odds ratio is converted to log odds. Log odds provide a metric with a range of negative infinity to positive infinity, which is symmetric around zero. Note, when α equals one, indicating the odds for reference and focal are the same, natural log of one is zero, resulting in the Δ_{MH} being zero. When Δ_{MH} is zero, the odds ratio α is one, indicating that the reference and focal groups odds are the same for getting a test item correct. A negative value for Δ_{MH} would indicate a test item favoring the reference group, positive values favoring the focal group (Holland & Thayer, 1988).

$$\Delta_{MH} = -2.35\ln(\overset{\wedge}{\alpha}_{MH}) \qquad (5)$$

The ETS's DIF classification rules based on effect size measured by Δ_{MH}, is categorized as A, B or C.

"A" represents negligible DIF, "B" represents moderate DIF, and "C" represents large DIF (Zwick & Ercikan, 1989; Dorans & Holland, 1993; Hidalgo & Lopez, 2004). Equations 6, 7 and 8 define these classifications based on $\Delta_{MH.}$

$$\text{A (Negligible DIF)} = |\Delta_{MH}| < 1 \tag{6}$$

$$\text{B (Moderate DIF)} = 1 \leq |\Delta_{MH}| < 1.5 \tag{7}$$

$$\text{C (Large DIF)} = |\Delta_{MH}| \geq 1.5 \tag{8}$$

In summary, (a) for Category A, MH Delta not significantly different from 0 (Alpha = .05) or absolute value of MH Delta < 1.0; (b) for Category B, MH Delta not significantly different from 0 and absolute value of MH Delta >= 1.0 or MH Delta significantly different from 0 and absolute value of MH Delta >= 1.0 but < 1.5; (c) for Category C, MH Delta significantly different from 1 and absolute value of MH Delta >= 1.5.

The MH statistic tests the null hypothesis with a chi-square test. . Equation 9 illustrates the formula for testing the null hypothesis, specifically that $\alpha = 1$. All of the variables in Equation 9 are found in Table 1. As with the familiar Pearson's chi-square statistic, the observed and expected cell frequencies are compared for discrepancies. Camilli and Shepard (1994) explain the most important aspect of the MH chi-square test by stating this related to $A_j - E(A_j)$ in Equation 9, "This represents the discrepancy between the observed number of correct responses on the item by the Reference group and the expected number" (p. 120). If the observed correct frequency count for the reference group (i.e. A_j) is higher than the expected count (i.e. $E(A_j)$), the potential for DIF favoring the reference group exists.

Conversely, if the observed correct frequency count for the reference group (i.e. A_j) is less than the expected count (i.e. $E(A_j)$), the potential for DIF favoring the focal group exists.

$$MH\chi 2 = \frac{(\left| \sum_j A_j - \sum_j E(A_j) \right| - \frac{1}{2})^2}{\sum_j \mathrm{var}(Aj)} \quad , \ \mathrm{var}(Aj) = \frac{N_{rj} Nf_j M_{1j} M_{oj}}{T^2{}_j (T_j - 1)} \qquad (9)$$

Standardization Procedure

Dorans and Kulick (1983, 1986) first applied the standardization procedure on the Scholastic Aptitude Test (SAT) to assess DIF on test items. Although in the literature (Camilli & Shepard, 1994; Clauser & Mazor, 1998; Monahan, et al., 2007) the standardization method is described as a procedure used to assess DIF, as stated by Dorans and Holland (1993), "…Mantel-Haenszel was selected as the method for DIF detection and standardization was selected as the method for DIF description" (p. 59). The specific reason for this classification of the two methods was not explicitily clear in Dorans and Holland, but may be attributed to the fact that the standardization procedure lacks a significance test. The standardization procedure as a method used to assess DIF can be found in numerous research studies (Clauser & Mazor, 1998). The popularity of this procedure is more than likely associated with its simplicity in calculating the standardization DIF measure. The major drawback already stated is the lack of a test of significance (Clauser & Mazor, 1998).

Equation 10 specifies the formula for calculating the standardized p-difference (STD P-DIF) DIF measure.

P_{fj}, P_{rj} and j are defined as described in Table 1. K_j and W_j are the only new terms being introduced. The standardization procedure is so named because of the variable W_j (Dorans & Holland, 1993).

$$\text{STD P - DIF} = \frac{\sum_{j=1}^{j} K_j(P_{fj} - \mathrm{P}_{\mathrm{r}\,j})}{\sum_{j=1}^{j} K_j} \qquad , \mathrm{W_j} = \mathrm{K_j}/\Sigma \mathrm{K_j} \tag{10}$$

In calculating the standardized p-difference, the proportion correct on an item for the focal group is subtracted from the proportion correct on the same item for the reference group, for each jth ability group. The standardized p-difference (STD P-DIF) based on the formula in Equation 10 is a value with a range from -1 to +1. If a test item is behaving the same for the focal and reference ability groups, the STD P-DIF measure will be zero indicating no DIF. If the item is favoring the reference group based on the proportions calculated, the difference between (P_{fj} - P_{rj}) will be negative. If a test item is favoring the focal group, the difference between (P_{fj} - P_{rj}) will be positive. This can be seen in Table 2, column 6 for fourth ability groups.

Table 2

Proportions Correct & Frequencies for Reference/Focal Ability Groups

Ability Groups	P_{rj}	N_{rj}	P_{fj}	N_{fj}	$(P_{fj} – P_{rj})$	$W_j = K_j/\Sigma K_j$	$W_j(P_{fj} – P_{rj})$
0 - 10	.6667	3	.5000	4	-.1667	.0656	-.0109
11 - 20	.3684	6	.3539	10	-.0145	.1639	-.0024
21 - 30	.6667	25	.5000	27	-.1667	.4426	-.0738
31 - 40	.5833	18	.7500	20	.1667	.3279	.0547
						STD P-DIF =	-.0324

Standardization as a name describing the standardized p-difference procedure is based on the variable W_j. The standardized p-difference uses a standard weight as defined by W_j in Equation 10. K_j is typically equal to N_{fj} which is the number of examinees at jth ability group for the focal group. W_j is a weighting factor used to discriminate between the calculated differences at each ability group (i.e. $(P_{fj} – P_{rj})$). In Table 2, column 6, the calculated difference between the first and third ability groups is the same, a negative .1667. A greater weight should be given to the difference observed for the third group given the total number of examinees (i.e. 52) in this ability group, as compared to the total number of examinees (i.e. 7) in the first ability group. An average could be used and applied as the weight for each ability group, but this would give equal weight to each difference calculated. Using a weighting factor, Wj at each ability group will result in the greatest weight to differences in P_{fj} and P_{rj} at those ability groups most frequently achieved by the focal group under study (Dorans & Holland, 1993).

This can be seen in Table 2 where the weighting factor for the first ability group is .0656 and for the third ability group is .4426. The third ability group weighting factor is higher given the significant difference in the number of examinees at this ability group, that is, the -.1667 difference is more meaningful as related to this ability group.

In assessing whether or not the difference that exists between P_{fj} and P_{rj} warrant further investigation, an effect size for STD P-DIF is available (Dorans & Holland, 1993). In Table 2, STD P-DIF was calculated as a hypothetical example to demonstrate the simplicity and utility of the standardized p-difference procedure. The calculated value in the example is a value of -.0324. Does this test item based on this value warrant further investigation? Based on Dorans and Kulick's (1986) effect size recommendations the answer is no, differences in proportions between the focal and reference groups for the hypothetical example are negligible. The effect size recommendation is, (a) negligible DIF based on the calculated standardized p-difference having a value between -.05 and +.05; (b) moderate DIF based on the calculated standardized p-difference having a range between -.10 and -.05; and (c) large DIF based on the calculated standardized p-difference having a value beyond -.10 or +.10.

Logistic Regression Procedure

The logistic regression procedure is considered a general non-IRT method for assessing DIF (Camilli & Shepard, 1994). Logistic regression as a DIF detection procedure does not employ specific measurement models like true IRT methods.

Swaminathan and Rogers (1990) introduced logistic regression as a DIF detection procedure, which is arguably comparable if not better than MH in assessing differential item functioning (DIF). A primary advantage of using the logistic regression method is its ability to detect non-uniform DIF (Monahan, et. al., 2007; Clauser & Mazor, 1998; Camilli & Shepard, 1994; Swaminathan & Rogers, 1990). Uniform DIF exists when a test item favors one group over another over the entire ability continuum. Non-uniform DIF exists when a test item favors one group over another for just part of the ability continuum. The group disadvantaged for the first part becomes the group being favored over the second part of the ability continuum. Neither MH nor the standardized method provides the ability to detect non-uniform DIF.

There are two main equations associated with the logistic regression method. Equation 11 represents the first equation, and Equations 12, 13, and 14 represent the second main equation. The differences between Equations 12, 13, and 14 will be discussed when logistic regression hypothesis testing is presented. P_j represents the conditional probability for answering a test item correctly. When P_j differs between the reference group and focal group the test item is exhibiting DIF. The logit(p)' in Equations 12, 13, and 14 is called the logit function for logistic regression. A logit can be transformed into odds by the expression $e^{\text{logit(p)'}}$, with odds the probability can be determined, see Equation 11. When logit(p)' is greater for the reference group, the reference group will have a higher probability of answering a test item correctly, hence, a differential functioning test item.

$$\hat{P}_j(u=1) = \frac{e^{\log it(p)'}}{1+e^{\log it(p)'}} = \frac{\overset{\wedge}{odds}}{1+\overset{\wedge}{odds}}\text{, where j=0 for Ref., j=1 for Foc.} \quad (11)$$

$$\text{Model 1: } \log it(p)' = \beta_0 + \beta_1 X + \beta_2 G + \beta_3 XG \quad (12)$$

$$\text{Model 2: } \log it(p)' = \beta_0 + \beta_1 X + \beta_2 G \quad (13)$$

$$\text{Model 3: } \log it(p)' = \beta_0 + \beta_1 X \quad (14)$$

The logit function in Equations 12, 13 and 14 is a function which specifies the linear combination of the predictor variables, in a logistic regression analysis of DIF. β_0 is the intercept, β_1 is the total test score coefficient, β_2 is the group membership coefficient, β_3 is the interaction coefficient (i.e. a test of non-uniform DIF), X is the observed total score for an examinee, and G represents group membership defined as either reference or focal group. β_2 can also be viewed as the combined log odds ratio as defined by the MH procedure, see Equation 3. If β_2 differs significantly from zero, the odds of getting an item right are not the same for the reference and focal group. Given that X represents total test score for an examinee, it should be no surprise that β_1 is always mostly statistically significant. It should be expected that an examinee with a higher test score have higher odds of getting a test item correct (Camilli & Shepard, 1994). If β_3 is not significant, non-uniform DIF is not present. In summary, when $\beta_3 = 0$, $\beta_2 \neq 0$, and β_2 is significantly different than 0, uniform DIF exist.

Hypothesis testing for logistic regression is conducted in several steps whereby model parsimony is the goal. A model is parsimonious when the least number of coefficients are estimated. Hypothesis testing begins by comparing Model 1 and Model 2 as specified in Equations 12 and 13 respectively. If the term β_3 in Model 1, a test of non-uniform DIF is not significant, Model 2 against Model 3 is then tested.

If the term β_2 in Model 2, a test of uniform DIF is not significant, Model 3 is the final model for specifying the logit (Camilli & Shepard, 1994). MH requires determining and grouping examinees based on ability which is a statistically arbitrary process. Logistic regression does not require groupings by ability. For instance, a test of β_2 uniform DIF is a test of its strength in predicting logit(p)' in and of itself factoring out ability β_1 and non-uniform DIF $\beta_{2.}$ Controlling or factoring (i.e. partial correlation) out other predictors is a tenet of regression analysis.

In the literature related to logistic regression, many different metrics have been reported to assess effect size. These methods do not utilize instinctive metrics that can be derived from logistic regression, more specifically the odds ratio (Monahan, et. al., 2007). The logistic regression odds ratio is defined by Equation 15. It represents the reference to focal group odds of answering a test item correctly, conditioned on ability which is defined by total test score. The expression $\exp\left(\overset{\wedge}{\beta_2}\right)$ represents the multiplicative change in odds for a member of the reference group answering a test item correctly, on average, holding the other predictors in the logit function constant.

$$\overset{\wedge}{\alpha}_{LR} = \exp\left(\overset{\wedge}{\beta_2}\right) \qquad (15)$$

As stated earlier, the null definition of DIF for logistic regression exists when $\beta_2 = 0$, therefore, $\overset{\wedge}{\alpha}_{LR} = 1$.

As with Mantel-Haenszel's $\hat{\alpha}_{MH}$, when $\hat{\alpha}_{LR}$ equals 1, the odds for the reference group answering the test item correctly is the same as the odds for the focal group. As with $\hat{\alpha}_{MH}$, $\hat{\alpha}_{LR}$ is not symmetric around 0. Using Holland and Thayer's (1988) conversion formula, $\hat{\Delta}_{LR}$ can be defined similar to $\hat{\Delta}_{MH}$, see Equation 16.

$$\hat{\Delta}_{LR} = -2.35\ln(\hat{\alpha}_{LR}) \qquad (16)$$

The ETS's DIF classification rules based on effect size (Zwick & Ercikan, 1989; Hidalgo & Lopez, 2004) now measured by $\hat{\Delta}_{LR}$ can be summarized similarly to MH, (a) for Category A, $\hat{\Delta}_{LR}$ is not significantly different from 0 or $\hat{\Delta}_{LR}$ absolute value is less than 1; (b) for Category B, $\hat{\Delta}_{LR}$ is significantly different from 0, $\hat{\Delta}_{LR}$ absolute value is at a minimum 1, and $\hat{\Delta}_{LR}$ absolute value is less than 1.5; (c) for Category C, $\hat{\Delta}_{LR}$ is significantly different from 0, $\hat{\Delta}_{LR}$ absolute value is at a minimum 1.5. Note $\hat{\Delta}_{LR}$ absolute value is at a minimum 1.5 when $\beta_2 = .4255$, that is $\hat{\alpha}_{LR} = e^{.4255} = 1.53$.

SIBTEST Procedure

The DIF statistics presented in this dissertation, MH, Standardization, and Logistic Regression, were each developed with the premise of determining and measuring DIF, but each fail to address the underlying causes of DIF.

Shealy and Stout (1993) introduced Simultaneous Item Bias Test (SIBTEST) as a procedure to measure DIF, but also as a method to determine a possible underlying cause of DIF, specifically multidimensionality. Multidimensionality in the literature is identified as one factor contributing to test items functioning differently between groups (Oshima & Miller, 1992; Shealy & Stout, 1993; Roussos & Stout, 1996a). SIBTEST closely resembles Dorans and Kulick's (1983, 1986) standardization DIF procedure, but with many important improvements (Clauser & Mazor, 1998). SIBTEST provides a mechanism for not only detecting single item DIF, but multiple item DIF, known in the literature as differential test functioning (DTF). Dorans and Kulick's standardization DIF procedure lacks a test of significance which is another improvement provided with SIBTEST. Finally, unlike MH and standardization where observed scores are used to match examinees, SIBTEST provides a regression correction procedure to mitigate the limitation of using observed scores which contain measurement error (Gierl, Gotzmann, & Boughton, 2004).

SIBTEST null and alternative statistical hypotheses are represented in Equation 17. The parameter β_{UNI} specifies the presence or absence of DIF. As can be seen in Equation 17, DIF is innocuous when $\beta_{UNI} = 0$.

$$H_0 : \beta_{UNI} = 0 \text{ vs. } H_1 : \beta_{UNI} \neq 0 \tag{17}$$

The specifics of β_{UNI} are defined by Equation 18.

$$\beta_{UNI} = \int B(\theta) fF(\theta) d(\theta) \tag{18}$$

β_{UNI} is defined by three parts, $B(\theta)$, $fF(\theta)$ and $d(\theta)$.

As with the standardization procedure, DIF is measured by the difference in probability of answering a test item correctly between the reference and focal groups conditional on ability, that is, $B(\theta)$ equals the difference between $P(\theta,\mathrm{R})$ - $P(\theta,\mathrm{F})$. The variable $fF(\theta)$ is a probability density function for the focal group's theta θ and $d(\theta)$ is the differential of theta (Gierl, Gotzmann, & Boughton, 2004). Theta is considered a continuous random variable which can assume an unbounded range of values. Therefore, having defined a probability density function of theta along with the differential of theta, the difference in probability of a correct answer on a test item, between the reference and focal group can be calculated for any focal group examinee's ability level between negative infinity and positive infinity. More eloquently stated by Gierl, Gotzmann, and Boughton, "$B(\theta)$ is integrated over θ to produce $\beta_{UNI,}$ a weighted expected mean difference in probability of a correct response on an item between reference and focal group examinees who have the same ability" (p. 244).

An estimate of β_{UNI} is provided by $\overset{\wedge}{\beta}_{UNI}$ defined in Equation 19.

$$\overset{\wedge}{\beta}_{UNI} = \sum_{k=0}^{K} p_k d_k \qquad (19)$$

Examinees are divided into subgroups conditional on ability. The total number of subgroups is defined by K, and a specific ability subgroup is defined by k as illustrated in Equation 19. As with the standardization procedure a weighting factor is specified by P_k which is the proportion of focal group members in subgroup k. The variable d_k equals $\overset{*}{P}_{R_K}$ - $\overset{*}{P}_{F_K}$, which specifies the difference in adjusted means on the test item under study for the reference group and focal groups based on each subgroup k (Gierl, Gotzmann, &

Boughton, 2004). The means are adjusted using a regression correction procedure as outlined in Gierl, Gotzmann, and Boughton. An overall statistical test for β_{UNI} is defined by Equation 20. The statistic SIB has a normal distribution where the mean is 0 and a standard deviation is 1 when the null hypothesis is true (Gierl, Gotzmann, & Boughton, 2004). The standard error of β_{UNI} is represented in Equation 20 by $\sigma^{\wedge}\left(\overset{\wedge}{\beta}_{UNI}\right)$.

$$SIB = \frac{\overset{\wedge}{\beta}_{UNI}}{\sigma^{\wedge}\left(\overset{\wedge}{\beta}_{UNI}\right)} \tag{20}$$

SIBTEST's effect size guidelines were initially defined by Nandakumar (1993). These guidelines are not comparable to the ETS's classification of negligible, moderate and large DIF. Given the extensive research, popularity, and familiarity of the ETS's classifications of DIF, Roussos and Stout (1996b) devised a method by which values of $\overset{\wedge}{\beta}_{UNI}$ could be interpreted using the ETS's classifications. $\overset{\wedge}{\Delta}_{\mathrm{MH}}$ and $\overset{\wedge}{\beta}_{UNI}$ are different metrics not on the same scale, therefore, as stated by Roussos and Stout (1996b), "no strict mathematical relationship exists between the two estimators that allows $\overset{\wedge}{\Delta}$ cutoff values to be converted to equivalent $\overset{\wedge}{\beta}$ values" (p. 219). Research has shown that these two estimators are highly correlated (Dorans & Holland, 1993). Given that the absence of DIF for both metrics, their values equaling zero, Roussos and Stout (1996b) defined an approximate linear relationship as $\overset{\wedge}{\beta}_{UNI} = \mathrm{K}^{*}\overset{\wedge}{\Delta}$.

The constant K is defined as a constant with an approximate value of -17 for 3PL data based on research by Roussos and Stout (1996b). K is defined as a constant with an approximate value of -15 for 1PL and 2PL data based on research by Shealy and Stout (1993).

The ETS's DIF classification rules based on effect size can now be measured by $\hat{\beta}_{UNI}$, summarized similarly to MH for 1PL/2L models, (a) for Category A, $\hat{\beta}_{UNI}$ is not significantly different from 0 (Alpha = .05) or absolute value of $\hat{\beta}_{UNI}$ < .067; (b) for Category B, $\hat{\beta}_{UNI}$ not significantly different from 0 and absolute value of $\hat{\beta}_{UNI}$ >= .067 or $\hat{\beta}_{UNI}$ significantly different from 0 and absolute value of $\hat{\beta}_{UNI}$ >= .067 but < .10; (c) for Category C, $\hat{\beta}_{UNI}$ significantly different from 0 and $\hat{\beta}_{UNI}$ >= .10.

The ETS's DIF classification rules based on effect size can now be measured by $\hat{\beta}_{UNI}$, summarized similarly to MH for the 3PL model, (a) for Category A, $\hat{\beta}_{UNI}$ is not significantly different from 0 (Alpha = .05) or absolute value of $\hat{\beta}_{UNI}$ < .059; (b) for Category B, $\hat{\beta}_{UNI}$ not significantly different from 0 and absolute value of $\hat{\beta}_{UNI}$ >= .059 or $\hat{\beta}_{UNI}$ significantly different from 0 and absolute value of $\hat{\beta}_{UNI}$ >= .059 but < .088; (c) for Category C, $\hat{\beta}_{UNI}$ significantly different from 0 and $\hat{\beta}_{UNI}$ >= .088. In concluding the discussion on SIBTEST, it is important to note that this statistic also lacks a non-uniform test of DIF.

Nonparametric procedures were explained in great detail as statistical methods for assessing whether a test item behaves differently for different groups of examinees. Table 3 provides a summary of the most utilized non-parametric DIF statistics today with its effect size.

Parametric DIF Procedures

Parametric procedures' foundation is based on estimating ability and test item parameters for reference group and focal group examinees. Depending on the model selected to fit the data, the number of parameters being estimated can vary. In discussing different IRT models, Oshima and Morris (2008) state, "A variety of IRT models have been developed to address different types of item response formats" (p. 44). For instance, the 1PL model (Rasch, 1960) defines one parameter, the 2PL model (Choppin, 1983) defines two parameters, and the 3PL (Birnbaum, 1968) model defines three parameters. There are numerous IRT Models typically categorized as dichotomous or polytomous. Dichotomous IRT models handle test response data in the format of a correct response (i.e., 1) or an incorrect response (i.e., 0). Polytomous IRT models can estimate probabilities beyond just either correct or incorrect answers. Polytomous models can estimate probabilities based on an examinee choosing a specific answer. In other words, what is the probability of an examinee selecting a specific answer out of five choices? IRT models the functional relationship between item responses from a test and an examinee's position on the underlying latent ability purported to be measured by the test (Oshima & Morris, 2008).

Table 3

Summary of DIF Procedures and Effect Sizes

DIF Procedure	Effect Size	Range	No DIF	Negligible DIF (A)	Moderate DIF (B)	Large DIF (C)
(1985) Mantel-Haenszel	Δ_{MH}	-∞ to +∞ Midpoint 0	0	< 1	≥ 1 < 1.5	≥ 1.5
(1986) Standardization	STDP-DIF	-1 to +1 Midpoint 0	0	-.05 to +.05	-.10 to -.05	< -.10 or > +.10
(1990) Logistic Regression	$\hat{\Delta}_{LR}$	-∞ to +∞ Midpoint 0	0	< 1	≥ 1 < 1.5	≥ 1.5
(1993) SIBTEST (1PL/2PL)	$\hat{\beta}_{UNI}$	-∞ to +∞ Midpoint 0	0	< .067	≥ .067 < .10	≥ .10
(1996) SIBTEST (3PL)	$\hat{\beta}_{UNI}$	-∞ to +∞ Midpoint 0	0	< .059	≥ .059 < .088	≥ .088

The probability of an examinee in a specific ability group answering a question correctly is still calculated similarly to the contingency table procedures. For each ability level measured defined by theta θ, the proportion of examinees getting the answer correct is used to determine the initial probability for that ability level. Although this method is similar to the contingency table procedures, important differences exist. The true ability of an examinee from a conceptual perspective is measured on a continuous scale (see Figure 1), as opposed to a discrete scale. The parametric item characteristic curve (ICC) in Figure 1 is interpreted as the probability correct for a randomly identified examinee in the population, not the probability correct based on proportions as defined with contingency table approaches. Once parameters are estimated using likelihood statistics, the probability determined is referred to as the likelihood of a randomly selected examinee in the population of ability θ (Camilli & Shepard, 1994). This interpretation is made possible because the proportions are used a priori.

IRT methods define DIF as a significant difference between ICCs, see Figure 1. In the case of dichotomous models, there are two ICCs, one for the reference group and one for the focal group. If DIF is not present the ICCs will overlap, therefore, the example in Figure 1 is a case where DIF exists. Throughout the ability continuum, a member of the reference group in comparison to a member of the focal group at the same ability level, the reference group member has a higher probability of answering this test item correctly.

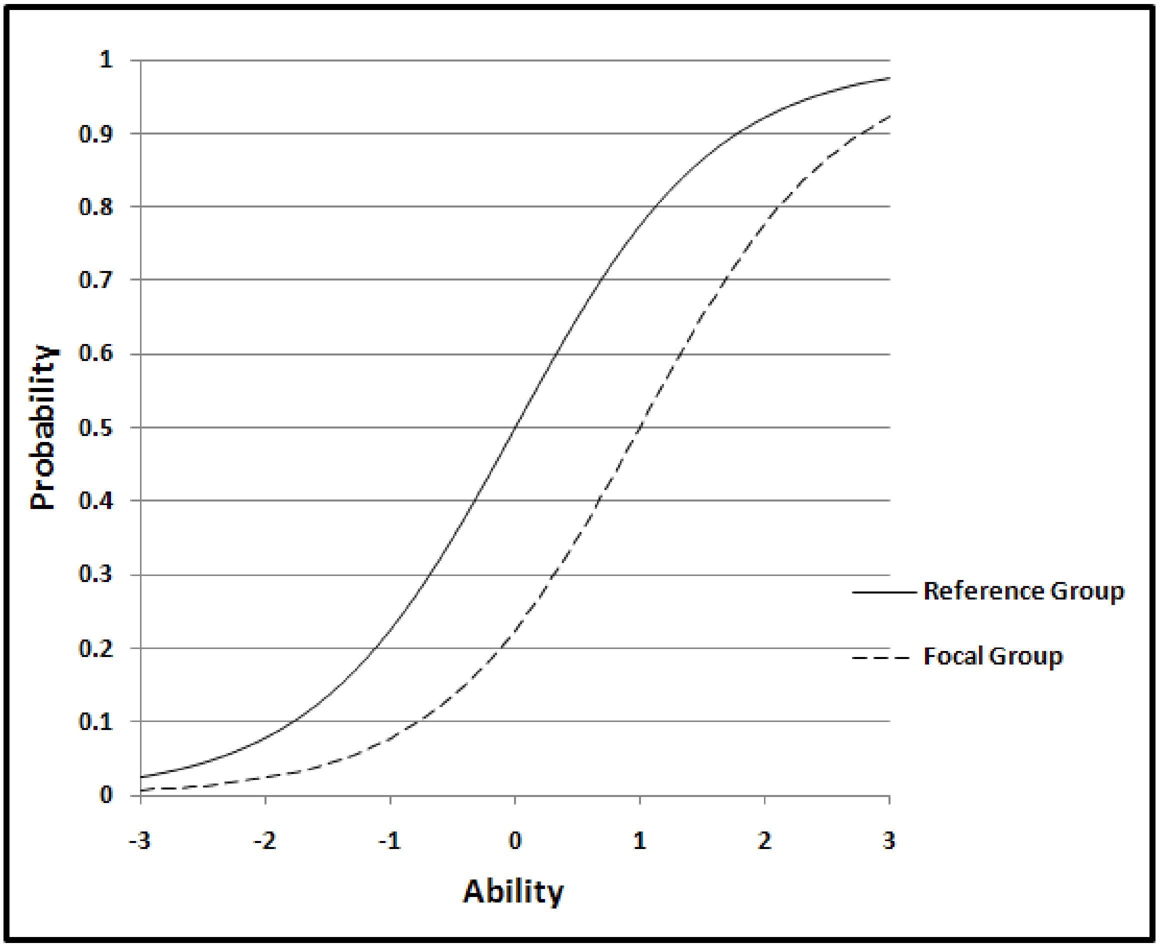


Figure 1. Illustration of a Test Item ICC for Reference and Focal Groups Displaying DIF.

Before determining if DIF exists using the IRT approach, as noted by Oshima and Morris (2008), "One has to, of course, allow for sampling error. However, the gap can be larger than what would be expected due to sampling" (p. 46). Several statistical techniques were developed to determine if the difference between the two ICCs is statistically significant.

Lord (1980) chi-square method compares the item parameters between the two groups, Raju (1988) area measure estimates the area between the two ICCs, Thissen, Steinberg, and Wainer (1988) likelihood ratio test compares the fit of the model with and without separate group parameter estimates, and differential functioning of items and tests (DFIT) framework methods (Flowers, Oshima, & Raju, 1999; Oshima, Raju, & Flowers, 1997; Raju, van der Linden & Fleer's, 1995) uses a cutoff score for each test item to flag DIF. The cutoff score is determined by producing a 95 or 99 percentile rank score from a frequency distribution under the DIF = 0 (null hypothesis) condition.

Parametric versus Nonparametric Procedures

There have been many studies investigating the strengths and weaknesses between parametric versus nonparametric DIF procedures. All DIF methods regardless of the classification yields aberrant results when assumptions associated with the DIF procedure are violated (Camilli & Shepard, 1994; Clauser & Mazor, 1998; Hambleton, Swaminathan, & Rogers, 1991; Millsap & Everson, 1993; Osterlind & Everson, 2009; Shepard, Camilli, & Averill, 1981; Teresi & Fleishman, 2007; Wiberg, 2007). In reviewing the literature related to the advantages associated with parametric procedures, the focus will be on those advantages deemed as most important related to the efficacy of reporting DIF or no DIF. The property of invariance, matching variable, and the importance of item parameters will be discussed.

Property of Invariance

The tenet of invariance is central to parametric procedures (Hambleton et al., 1991). Simply stated, if item parameters and ability estimates are determined for a random sample of examinees in a population, these estimated item parameters and ability estimates will not change for a different random sample of examinees from a different population. In many of the nonparametric procedures discussed, this is not possible because the proportions used to determine whether differences in probabilities exist are related to the group of examinees. When the groups of examinees change, the proportions change. The property of invariance is one of the main distinctions between parametric and nonparametric DIF procedures. Based on this review of literature, Lord and Novick (1968) were the first to highlight the property of invariance related to educational testing. In discussing Lord and Novick's assertion related to the property of invariance, Bejar (1980) provides this description:

> A test is population invariant if the characteristic curve (i.e., the regression of probability of success on achievement) of every item in the test within one population is a linear transformation of the characteristic curve for that item in the other population. (p. 514)

Lord (1980) argues that an ICC can also be considered a regression function, whereby the probability of success on a test item can be regressed on the latent construct being measured. If this is the case, as noted by Lord, "…regression functions remain unchanged when the frequency distribution of the predictor variable is changed" (p. 34). The probability of an examinee answering a test item correctly based on the 2PL model is given by Equation 21.

$$P(\theta) = \frac{e^{a(\theta-b)}}{1+e^{a(\theta-b)}} \tag{21}$$

The regression function where the probability of success on a test item can be regressed on the latent construct being measured, is equal to $a(\theta\text{-}b)$. It is clear to see and should be expected that when ability defined by theta is equal to the item difficulty, an examinee has a .5 probability of chance in getting the test item correct. If an examinee's ability exceeds the item difficulty parameter *b*, the chance of getting the item correct increases. Conversely, if the item difficulty parameter *b* exceeds the examinee's ability, the chance of getting the item correct decreases.

The chances described above for an examinee in one population should not differ for an examinee in another population based on a linear transformation (Bejar, 1980; Lord, 1980; Shepard et al., 1981). As an example, consider the item parameters *a* and *b* to be defined for examinees in population 1: Item parameters $a*$ and $b*$ for examinees in population 2 based on a linear transformation, is defined by Equations 22 and 23 (Bejar, 1980). In these two equations, α is the slope of the linear conversion, and β is the intercept of the linear conversion.

$$a^{*} = \left(\frac{1}{\alpha}\right)a \tag{22}$$

$$b^{*} = \alpha b + \beta \tag{23}$$

In discussing this linear relationship in great details, Lord (1980) uses the notion that the regression function where the probability of success on a test item can be regressed on the latent construct being measured, is equal to $a(\theta\text{-}b)$.

If this is the case, adding a constant to theta, and adding the same constant to the item difficulty parameter *b*, the regression function remains the same, hence, the probability of success is unchanged (Lord, 1980). As stated by Lord related to this case, “This means that the choice of origin for the ability scale is purely arbitrary; we can choose any origin we please for measuring ability as long as we use the same origin for measuring item difficulty…” (p. 36). This is why examinees from two different populations where the ability distributions differ as related to the means and variances will still have the same probability of success on a test item at any given ability level. This is not to say that the item parameter estimates from two different populations will be the same; they will be different, but as stated by Lord (1980), “The invariance of item parameters…clearly holds only as long as the origin and unit of the ability scale is fixed” (p. 36). The invariance of these different parameters is made possible by their linear relationship. Several studies have been conducted related to the property of invariance.

The property of invariance hypothesis is supported by several empirical studies (Rudner & Covey, 1978; Ironson & Subkoviak, 1979; Rudner, Getson, & Knight, 1980; Lord, 1980; Hambleton et al., 1991). Rudner and Covey in evaluating different DIF procedures, demonstrated the property of invariance by considering two different populations; one population consisted of 2637 hearing impaired students and 1607 normal students. Ironson and Subkoviak in comparing several methods to assess item bias demonstrated the property of invariance by utilizing two different populations; one population consisted of 1691 12th grade black students and 1794 12th grade white students. In conducting a Monte Carlo study comparing seven DIF techniques, Rudner et al. validated the property of invariance by using two different simulated populations.

The simulated populations' ability distributions differed by one standard deviation. As noted by Rudner et al., the one standard deviation was appropriate based on what is, "frequently encountered in actual data" (p. 5). Finally, in researching the property of invariance, Lord (1980) compared item parameter estimates from 2250 white students with item parameter estimates from 2250 black students for an 85 verbal item SAT test.

Matching Variable

In the literature related to matching variable, observed score versus latent variable has also been used to distinguish the differences between nonparametric and parametric DIF procedures (Potenza & Dorans, 1995). There has been extensive research related to the matching variable required for DIF analyses (Bolt, 2002; Clauser & Mazor, 1998; Donoghue, Holland, & Thayer, 1993; Potenza & Dorans, 1995; Mazor, Kanjee, & Clauser, 1995; Penfield & Lam, 2000; Penny & Johnson, 1999; Wiberg, 2007; Zwick, 1990). The matching variable constitutes what is required to accurately identify the presence or absence of DIF. It should be expected that if comparing groups with different abilities, a difference would exist in their performance on a test item.

In the context of observed score, total test score is often used as the matching variable. An examinee is grouped with other examinees based on the examinee's ability. Ability in this context is determined based on performance on the test related to the items being studied for DIF. Given this definition of matching variable, examinees with similar total test scores would be grouped together; hence, ability groups are determined based on total test score. Determining an ability group based on total test score is as stated earlier a statistically arbitrary process.

Furthermore, it is not uncommon for the two groups being compared in a DIF analysis to have unequal mean and variances related to ability (Penny & Johnson, 1999). Related to flagging DIF, if this is the case as stated by Penfield and Lam (2000), "…the Type I error rates increases, and this increase becomes more extreme as the discrimination of the item increases and as the reliability of matching variable decreases" (p. 10). There have been many recommendations proposed in the literature related to increasing the reliability of the matching variable when total test score is used. Holland and Thayer (1988) recommended including the studied test item in the total test score regardless if it is identified as a DIF item. Mazor, Kanjee, and Clauser (1995) proposed using an external measure in conjunction with the internal measure (i.e. total test score) when assessing ability. Clauser and Mazor (1998) discussed the idea associated with thick versus thin matching, essentially this is using wider score categories when determining ability.

All of the recommended solutions potentially can increase the reliability of the matching variable when total test score is used. It is the opinion of many that parametric IRT DIF methods based on the latent measure of ability approach, provides a more statistically eloquent solution when the data fits the IRT model being used. Potenza and Dorans (1995) in discussing the latent measure approach state, "A fundamental difference between the latent variable approaches and the observed score approaches is the use of estimates, derived from observed data, of the latent trait or true score instead of observed score as either an implicit or explicit matching variable" (p. 28). Unlike the observed score approaches, the latent variable approaches utilize the joint estimation of item and ability parameters when ability and item parameters are unknown which is commonly the case, see Equation 24.

$$L(u_1, u_2, \ldots..u_N \mid \theta, a, b, c) = \prod_{i=1}^{N} \prod_{j=1}^{n} P_{ij}^{u_{ij}} Q_{ij}^{1-u_{ij}} \tag{24}$$

The ICC is a result of Equation 24, hence, the importance of the model chosen to fit the data. The notation *L* (i.e. likelihood) would be replaced with *P* for probability in Equation 24, if the calculation was based on a randomly selected examinee responding to a set of test items. Equation 24 is known as the likelihood as oppose to the probability given that u_1, u_2, u_3,…u_N is the actual response pattern observed from an examinee (Hambleton et al., 1991). Hambleton et al. provide a detailed discussion related to ability and item parameter estimation using parametric statistics.

Importance of Item Parameters

Accurately modeling the test data prior to assessing whether or not DIF exists is of utmost importance in any DIF analysis. If the data is not modeled accurately to reflect the responses to the test items, inaccurate conclusions may be purported. Many simulation studies have been conducted with the purpose of determining the importance of all three test item parameters (Reckase, 1978; Penny & Johnson, 1999). The three test item parameters often considered most important related to providing a sufficient modeling of the test response data are, (a) item difficulty parameter; (b) item discrimination parameter; and (c) pseudo-guessing parameter. For details related to these parameters, see Hambleton et al. (1991).

Parametric DIF procedures basic foundation hinges on the use of measurement models which can incorporate all three test item parameters if necessary.

This is important because Reckase (1978) in a comparison of using a one-parameter model versus a three-parameter model, concluded that using more than one-parameter provided a better fit to the test data. This conclusion was based on the comparison of sixteen different datasets, both real and simulated test data. In all comparisons studied, the three-parameter model was superior to the one-parameter model in fitting the data. In another study by Penny and Johnson (1999), it was determined that when between group differences exist in ability which is often the case with test data, not considering the discrimination and pseudo-guessing parameters could lead to an inflated Type I error rate when using the Mantel-Haenszel DIF statistic. Having the ability to model the test response data by incorporating all three test item characteristic parameters if necessary, is important to ensure accurate identification of DIF items.

The DFIT Framework

The history of developments related to the DFIT framework is shown in Figure 2. DFIT as a statistical method primarily was developed to overcome limitations associated with Raju's (1988) DIF area measure technique.

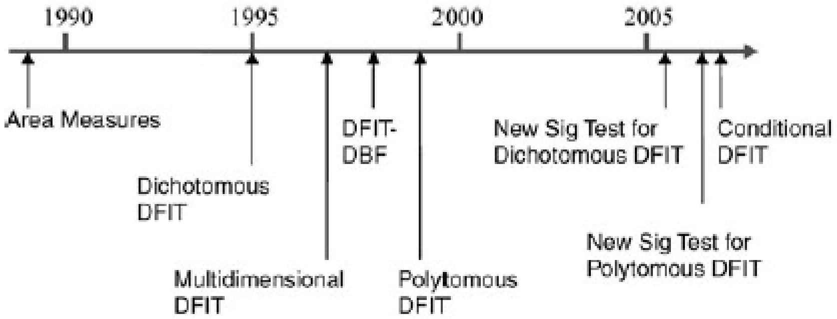


Figure 2. Historical Overview of the DFIT Framework (Oshima & Morris, 2008).

The DFIT framework as of today consists of a comprehensive set of methods for assessing differential item functioning. Dichotomous and polytomous test items can be investigated. Unidimensional and multidimensional models can be the bases for investigating differential item functioning. Individual test items as well as the entire test can be analyzed for differential item/test functioning. Uniform and non-uniform DIF can be detected equally effectively. Additional capabilities are also possible as stated by Oshima and Morris, "…it has been extended to a variety of applications such as differential bundle functioning (DBF) and conditional DIF" (p. 44). Table 4 provides a summary of the most utilized DIF procedures based on six different capabilities. Of the most utilized DIF statistics listed, DFIT is the only parametric technique capable of handling multidimensional models. As argued already, there are many advantages to utilizing DIF methods based on parametric principles. Furthermore, related to the capabilities listed in Table 4, DFIT only lacks an effect size measure.

Table 4
Summary of most utilized DIF procedures based on six different capabilities. 1. (P)arametric or (N)on-parametric IRT. 2. (L)atent or (O)bserved matching variable. 3. (D)ichotomous or (P)olytomous test items. 4. (S)ignificant test, (E)ffect size measure. 5. (U)niform, (N)onuniform DIF. 6. (Uni)dimensional models, (Mu)ltidimensional models.

Method	**(1) P/N**	**(2) L/O**	**(3) D/P**	**(4) S/E**	**(5) U/N**	**(6) Uni/Mu**
Lord's Chi-Square	P	L	D	S	U/N	Uni
Mantel-Haenszel	N	O	D/P	S/E	U	Uni
Area Measure	P	L	D	S	U/N	Uni
Logistic Regression	-	O	D/P	S/E	U/N	Uni/Mu
SIBTEST	N	L	D/P	S/E	U/N	Uni/Mu
DFIT	P	L	D/P	S	U/N	Uni/Mu

Note: Logistic Regression is considered a general non-IRT method.

It was stated earlier, IRT methods define DIF as a significant difference between ICCs. In its simplest form, DIF can be regarded as differences observed between the item parameters between the two groups of interest. A no DIF condition (null hypothesis) would result in Equation 25 for a 3PL model (Hambleton et al., 1991).

$$H_0 : b_r = b_f; a_r = a_f; c_r = c_f; \quad \text{r = ref. group, f = foc. group} \qquad (25)$$

A direct comparison of item parameters is intuitive, but the simplistic nature of this method is not without limitations (Lord, 1980; Rudner et al., 1980; Linn, Levine, Hastings, & Wardrop, 1981). Linn, Levine, Hastings, and Wardrop demonstrated a false negative DIF analysis within the ability range of (-3, 3), when true item parameters differences existed. The area measure of determining DIF goes a step beyond the direct comparison of item parameters (Rudner et al., 1980; Raju, 1988). The area measure involves calculating the exact area between two ICCs. Raju developed precise formulas for calculating the area between two item characteristic curves, taking into account the entire ability continuum. Raju’s (1988) area measure works well for the 1PL, 2PL and 3PL model when the c-parameter is equal. If the c-parameter is not equal, there are also limitations with Raju’s area measure method, hence, one of Raju’s motivations to develop the DFIT framework.

Dichotomous DFIT

Dichotomous DFIT was the first significant development within the DFIT framework (Raju et al., 1995). The development consisted of noncompensatory DIF (NCDIF), compensatory DIF (CDIF) and differential test functioning (DTF).

Oshima and Morris (2008) provide this specific definition of NCDIF in stating, "…is defined as the average squared distance between the ICFs for the focal and reference groups" (p. 46). NCDIF measures the difference in probability of selecting a correct response to a test item, between examinees from two different groups of interest (e.g. members from different ethnicity groups). In other words, is there a difference in probability for members of different groups endorsing a test item, while having the same latent ability? The difference in probability is taken over the entire latent ability continuum, denoted by E_F in Equation 27. NCDIF functions similarly to other item-level DIF statistics, in that all items are assumed to be DIF free with the exception of the item being investigated. In calculating NCDIF, squaring the difference between the item characteristic functions allows for both uniform and nonuniform DIF to be detected, see Equations 26 and 27.

$$d_i(\theta_s) = P_{iF}(\theta_s) - P_{iR}(\theta_s) \tag{26}$$

$$NCDIF_i = E_F[d_i(\theta_s)^2] \tag{27}$$

The DFIT framework offers the advantage for researchers and practitioners not only the ability to assess item-level DIF, but DIF can also be investigated at the test-level. CDIF and DTF are the two DFIT statistics developed for this purpose. CDIF is an important new novel development in DIF research. Osterlind and Everson (2009) discuss this importance in stating:

> The idea of compensatory DIF, as represented by the CDIF index, has the advantage of allowing researchers to study the overall effect of removing particular test items on the estimation of DTF, the differential functioning of the test as a whole. Thus, within this framework, test developers and psychometric specialists may be able to develop tests with the least amount of differential impact at the test score level. (p. 73)

Unlike item-level DIF where the difference is based on the item characteristic curves, test-level DIF is the difference between the two groups' test characteristic curves (TCC). A test characteristic curve is computed by summing the item response functions for each group in the DIF analysis. DTF and CDIF are related by Equations 28, 29 and 30.

$$d_{is}(\theta_s) = P_{iF}(\theta_s) - P_{iR}(\theta_s) \tag{28}$$

$$DTF = E[(\sum_{i=1}^{n} d_{is})^2] \tag{29}$$

$$DTF = \sum_{i=1}^{n}\left[Cov(d_i, D) + \mu_{di}\mu_D\right], \qquad DTF = \sum_{i=1}^{n} CDIF_i \tag{30}$$

Equations 26 and 28 are similarly defined as measuring the difference in probability of selecting a correct response to a test item, between examinees from two different groups of interest. Equations 27 and 29 are similarly defined in that the difference in probabilities is taken over the entire latent ability continuum, but for each test item as related to DTF. CDIF differs from NCDIF in that removing significant CDIF items results in direct changes in DTF. Oshima et al. (1997) explain CDIF in this way as related to Equation 30, "…is additive in the sense that differential functioning at the test level is simply the sum of compensatory differential functioning at the test level" (p. 255). Once again, NCDIF differs from CDIF given the fact that with NCDIF all items are considered to be DIF free. This is not the case with CDIF, items related to CDIF takes into consideration the correlation between DIF items (Raju et al., 1995; Oshima et al., 1997; Oshima & Morris, 2008). This is represented in Equation 30, where item covariances are taken into account when calculating CDIF, hence, DTF.

Multidimensional DFIT

The DFIT framework is based on parametric procedures, which utilizes IRT models to investigate the relationship between test item responses conditioned on the ability of an examinee. Given this, extending dichotomous DIF to multidimensional DIF is a matter of employing a multidimensional model for the DIF analysis. The 1PL, 2PL and 3PL models assume that the construct being measured is unidimensional, so only one latent trait is required. There are situation in which a test item must measure more than one latent trait, an example would be mathematical word problems. There are many psychological and educational tests which measure by design more than one latent trait (Oshima et al., 1997; Snow & Oshima, 2009). Conducting the DIF analysis with unidimensional models when multidimensionality is intended, would potentially produce false positives for those multidimensional test items. Reckase (1985) specified a 2PL multidimensional model (M2PL) to use when test items are known to be multidimensional. DFIT has been shown to work reasonably well within the framework of the M2PL model (Oshima et al., 1997).

DFIT-DBF

Identifying DIF items is important to ensure tests are fair, but just as important is to understand why items are identified as DIF. Explaining the sources of DIF will aid test developers in creating tests that are not bias (Douglas, Roussos, & Stout, 1996; Oshima, Raju, Flowers, & Slinde, 1998; Gierl, Bisanz, Bisanz, & Boughton, 2001). Differential bundle functioning (DBF) parallels the tenet of multidimensionality. DIF is assumed to occur if a test item is multidimensional.

A multidimensional test item typically consists of a primary latent construct and a secondary latent construct. If the secondary construct is intentional, it is considered auxiliary, conversely if the secondary construct is unintentional; it is a nuisance dimension reflecting item bias (Gierl, Bisanz, Bisanz, & Boughton, 2001).

Item-level DIF analysis as stated a few times already, operates under the premise that all other test items are DIF free. When item-level DIF analyses are conducted under this premise, small differences across many items may appear benign. In fact, when these small differences are considered together in the case of CDIF, significant DTF may be observed, hence the monumental importance of these two DFIT measures. CDIF measures the relationship between test items, on the other hand, DBF bundles items with the assumption that the items are related. Based on this test bundle, groups can be compared related to their performance on the test bundles. Evaluating test item bundles using DFIT is a natural extension; for the specific details see Oshima, Raju, Flowers, and Slinde (1998).

Polytomous DFIT

Educational reform efforts during the 1980s led to an increased focus on evaluating students using alternative assessment methods (e.g. portfolios, etc). These alternative methods are not scored from a 1-0 binary perspective. Osterlind and Everson (2009) provides a useful example for understanding the difference between binary versus polytomous items by stating, "…suppose an item is graded on a four-point continuum, leaving three score levels" (p. 66). In this example, DIF can be anywhere within the score levels.

With score levels, you would expect for groups with the same ability to have the same probability of choosing a specific answer, when this does not occur, understanding why is a matter of a DIF analysis.

Extending DFIT to polytomously scored test items is also seamless. There are many polytomous models available for researchers. Some of the more common polytomous models are, (a) Samejima's (1969) graded response model (GRM); (b) Bock's (1972) nominal response model; (c) Andrich's (1978) rating scale model; and (d) Muraki's (1992) generalized partial credit model. Extending DFIT to investigate polytomously scored test items requires employing a polytomous IRT model for the DIF analysis. NCDIF within the DFIT framework was shown to work reasonably well within the framework of the graded response model (Flowers et al., 1999).

Effect Size - DFIT

DFIT as a DIF technique is a promising new statistic in the area of DIF analysis (Osterlind & Everson, 2009). The statistic as discussed provides breadth and depth in many important areas lacking with other DIF statistics, see Table 4. DFIT provides a significance test of DIF, but lacks a very important measure, an effect size. A significance test answers only one important research question. In discussing significance testing, Hays (1981) states, "virtually any study can be made to show statistically significant results if one uses enough subjects" (p. 293). There are two other important questions that must be answered beyond significance testing. If the observance is real, than how large is it? Next, is the size large enough to be useful (Kirk, 2001)?

DIF analyses require a large sample size, typically greater than 200. Given that the recommended sample size for many statistics utilizing the normal probability distribution is 30, a sample size of 200 is large. Large sample sizes are known to cause Type I errors (i.e. false positives) when in fact a test item is unbiased (Cohen, 1990, 1994; Thompson, 1999, 2002; Finch, Cumming, & Thomason, 2001).

A large sample size is just one factor that may contribute to unreliable DIF findings. Other factors to consider are the types of ability distributions and the distribution of the population. An assumption of the DIF methodology, hence the statistics measuring DIF, is that the ability distributions of the reference group and focal groups are the same. Three studies demonstrated that when incongruence exists between the reference and focal groups' ability distributions, detecting DIF may not be reliable (Pommerich, Spray, & Parshall, 1994; Sweeney, 1996; Penny & Johnson, 1999). Another assumption held by many prominent researchers is the tenet of normality in the population. In investigating the departure from normality, Micceri (1989) found that normal distributions were rare related to achievement and psychometric measures. Of the 440 large-samples investigated, only 3.2% at a 99% confidence were normal. Based on these arguments presented, it is obvious why an effect size measure used in conjunction with a statistical significance test is a vital requirement.

Additional Improvement – Power

In reviewing the literature related to power being reported in DIF analyses, power is similarly defined as the statistically accepted statement of not committing a Type II error.

If a test item indeed exhibits DIF, and the DIF technique does not flag it as a DIF item, this is considered a false negative or statistically speaking, committing a Type II error. The studies reviewed during this literature review calculated power based on the proportion of correct rejections, when the null hypothesis of DIF is false (Ross, 2007; Awuor, 2008; Guler & Penfield, 2009). In assessing power related to the SIBTEST DIF statistic, Awuor (2008) stated, "The average of the percent of the proportions of flagging of the DIF items were calculated to represent statistical power of the SIBTEST procedure…"(p. 41). In comparing the efficacy between several DIF techniques, Guler and Penfield (2009) similarly defined power as, "…these rejection rates serve as an approximation of power…" (p. 324). DFIT's uniqueness related to its statistical method (IPR), will allow power to be calculated beyond a simple statement related to proportions. Empirically observed power may be determined. Again, being able to report power with a significance test of DIF and an effect size is a powerful statement related to the reliability and validity of any DIF analysis.

CHAPTER 3

METHOD

A Monte Carlo simulation study served as the overall framework for determining an effect size measure for DFIT's NCDIF, in simulating a 61-item test where item number 61 represented the DIF item. The MH statistic and parameter served as the basis by which an effect size measure was developed for DFIT's NCDIF. The MH DIF statistic is arguably the most widely used measure for DIF. Furthermore, researchers and practitioners are very familiar with the MH DIF effect size guidelines for measuring the size of DIF. The Mantel-Haenszel statistic has been shown to be stable in measuring the size of DIF for certain conditions (Hidalgo & Lopez, 2004). If the magnitude of DIF increases, one would expect for the effect size measure to also increase.

Similar to Donoghue, Holland, and Thayer (1993), DIF was embedded in item 61 by manipulating the b-parameter, and all other items were free of DIF. This approach allowed DIF to be measured by the difference in b-parameters for the focal and reference groups (i.e., $b_f - b_r$). The amount of DIF in item 61 (see Appendix A) the studied item, varied depending on the condition. The amount of DIF varied in increments of .025, .05, .10 or .20; see Appendix B. The a-parameter and c-parameters related to item 61 were the same for both the focal and reference groups. The a-parameter was modeled with 8 different values, the b-parameter was modeled with 11 different values and the c-parameter was either 0 for the 1PL/2PL models or .20 for the 3PL model; see Appendix B. The choice of .20 for the pseudo-guessing parameter is associated with typical multiple choice exams having five choices.

This resulted in the 1PL model having 11 different difficulty levels being studied. Related to the 2PL and 3PL models, the b-parameters were fully crossed with the 8 different a-parameters.

The combination of parameters resulted in 11 conditions for the 1PL model, 88 conditions for the 2PL model, and 88 conditions for the 3PL model. The number of conditions investigated in this Monte Carlo simulation totaled 187; see Appendix B. Given that each condition was manipulated by embedding DIF in increments of .025, .05, .10, or .20 each condition could have 10, 20, 40 or 60 items being studied. These increments hereafter will be referenced to as "within conditions." This resulted in 5750 DIF items being studied; see Appendix B. Unequal and equal ability distributions were also investigated which resulted in an additional 5750 DIF items being estimated for MH and SIBTEST. Additional calculations were not required for NCDIF and area measure.

Study Design

Effect Size – DFIT(NCDIF)

Item Parameters. Ducan's (2006) estimated item parameters from a 60-item American College Testing (ACT) administration were used for this study. The 1-0 item responses for the test are from a simple random sample of 40,000 examinees. The examinees took an equivalent form of the ACT math subtest on the same national test date, presumably with the same testing conditions. Per Ducan (2006), the 1-0 data were imported into BILOG-MG 3 (Scientific Software International [SSI], 2003) software which produced the estimated item parameters.

Appendix A contains the estimated parameters of the 60 items used in this study. Table 5 provides the summary statistics for the item parameters.

Table 5
Means and standard deviations for item parameters used in the study

$\bar{a}$	σ_a	$\bar{b}$	σ_b	$\bar{c}$	σ_c	N
1.8	.54	.152	.91	.20	0	60

Sample Size. Fixed sample size pairs of (1000, 1000) for the reference and focal groups are used. In this study, the impact of sample size was not a factor being considered; therefore, the sample size was fixed throughout the study. The choice of using a sample size of 1000 is based on sample sizes in actual testing scenarios ranging from 250 to 3000 (Shealy & Stout, 1993).

Monte Carlo Simulation Study (Estimating MH and SIBTEST). The 1-0 data were generated for the 60-item test based on a sample size of 1000. An additional test item was used whereby DIF was embedded into the test item for the focal group utilizing the b-parameter. The a-parameter for this test item took on eight different values to simulate a comprehensive range of discrimination levels. The b-parameter for this test item took on eleven different values in simulating a comprehensive range of difficulty levels. Furthermore, each of the difficulty levels was varied for the focal group in increments of .025, .05, .10 or .20 in effect producing several studied test items.

The different a-parameters were crossed with the different b-parameters producing 5750 studied test items, see Appendix B. For the 1PL case, this produced 350 different data points. For the 2PL case this produced 2760 different data points. For the 3PL case this produced 2640 data points. Total score categorized into a certain number of categories served as the matching criteria for calculating the MH statistic.

True parameters were calculated for Raju's (1988) area measure, and Raju, van der Linden and Fleer's (1995) NCDIF and Holland and Thayer's (1988) MH. Statistics were also estimated for Holland and Thayer's (1988) MH and Shealy and Stout's (1993) SIBTEST. An approximate linear relationship was determined by plotting the two parameters (i.e. NCDIF and MH) to determine the formula NCDIF = K*MH, where K was defined as a constant. The correlation index for NCDIF and MH was also determined based on the conditions for this study.

Calculating DIF based on Area Measure. Raju's (1988) DIF measure based on the area formulas are used as an additional DIF measure in this study for comparison purposes. The item parameters in Appendix B, with the Equations 31 through Equation 33 (Hambleton et al., 1991), were used to calculate the area between the two ICCs, where D = 1.7.

$$\text{3PL: } Area = (1-c)\left|\left[2(a_2 - a_1)/Da_1a_2\right]\ln\left[1 + e^{Da_1a_2(b_2-b_1)/(a_2-a_1)}\right] - (b_2 - b_1)\right| \quad (31)$$

$$\text{2PL: } Area = \left|\left[2(a_2 - a_1)/Da_1a_2\right]\ln\left[1 + e^{Da_1a_2(b_2-b_1)/(a_2-a_1)}\right] - (b_2 - b_1)\right| \quad (32)$$

$$\text{1PL: } Area = \left|(b_2 - b_1)\right| \quad (33)$$

Swaminathan and Rogers (1990) and Hambleton, Swaminathan, and Rogers (1991) in studies assert moderate DIF (Category B) if the area measure is .6 or more.

Finally, the subscript one and two in the formulas represent the reference group's and focal group's a and b-parameters, respectively.

Calculating DIF based on NCDIF. Raju et al. (1995) noncompensatory DIF (NCDIF) was calculated using Equations 26 and 27.

Calculating the MH Parameter. Roussos, Schnipke, and Pashley (1999) developed a generalized formula which calculates the true MH parameter, see Equations 34 and 35. Equation 34 is a derivation of Equation 4 when many assumptions are considered. The specific details can be found in Roussos, Schnipke, and Pashley.

$$\alpha = \frac{\int_{-\infty}^{\infty} P_F(\theta) Q_R(\theta) \frac{\mathcal{F}_R(\theta)\mathcal{F}_F(\theta)}{\gamma_F \mathcal{F}_F(\theta) + \gamma_R \mathcal{F}_R(\theta)} \alpha(\theta) d\theta}{\int_{-\infty}^{\infty} P_F(\theta) Q_R(\theta) \frac{\mathcal{F}_R(\theta)\mathcal{F}_F(\theta)}{\gamma_F \mathcal{F}_F(\theta) + \gamma_R \mathcal{F}_R(\theta)} d\theta} \qquad (34)$$

where

$$\alpha(\theta) = \frac{P_R(\theta) Q_F(\theta)}{P_F(\theta) Q_R(\theta)} \qquad (35)$$

Equations 3 and 35 are equivalent when the assumption is made that matching examinees on observed proportion-right score is equal to matching examinees on θ. Software was developed by Roussos, Schnipke, and Pashley incorporating this formula which this study utilized. In a review of literature, this software has not been validated in a large-scale simulation study. A purpose of estimating the MH parameter served to validate the accuracy of the software which purports to calculate the MH parameter.

Estimating SIBTEST and MH . Shealy and Stout's (1993) SIBTEST was estimated with the simulation study in conjunction with a statistical software package. More specifically, DIFPACK$^{©}$ (Assessment Systems Corporation) software which implements the algorithm for calculating SIBTEST was integrated into the simulation study, see Appendix J. The MH statistic was calculated by developing a SAS routine to calculate the chi-square statistic (see Appendix I). Again, Roussos and Stout (1996b) defined an approximate linear relationship for SIBTEST related to MH Delta based on an IRT 3PL model as $\overset{\wedge}{\beta}_{UNI}$ = K* $\overset{\wedge}{\Delta}$. K for the 3PL model is defined as a constant with an approximate value of -17 based on research by Roussos and Stout (1996b). K is defined as a constant with an approximate value of -15 for 1PL and 2PL data based on research by Shealy and Stout (1993).

In evaluating the effectiveness of SIBTEST, Shealy and Stout (1993) determined K in $\overset{\wedge}{\beta}_{UNI}$ = K* $\overset{\wedge}{\Delta}$ based on a priori measure of potential bias. Based on the predetermined amount of bias, the parameter values for SIBTEST (see Equation 18) was calculated. Shealy and Stout defined unidirectional test bias as $B(\theta) = T_{SR}(\theta) - T_{SF}(\theta)$, where $B(\theta)$ represents the difference in the studied subtest response function between the reference and focal groups. MH parameter value was calculated based on Shealy and Stout's assertion that MH Delta based on a predetermined amount of bias is, "proportional to the horizontal distance between $T_{SR}(\theta)$ and $T_{SF}(\theta)$ …" (p. 182). Based on these priori calculations and research showing a high correlation between the two statistics, K was defined.

Shealy and Stout’s study showed a high correlation between the true parameters and the estimated SIBTEST and MH statistics. This study took a similar approach with the exception that SIBTEST was only estimated.

Study Specification and Factors. Table 6 lists the specifications and factors used in this study.

Table 6

Specifications and Factors of the Study

A. Number of Replications for estimating MH and SIBTEST : 100

B. Ability Distribution

No Impact Case

Mean value for ref. group and focal group theta respectively, $\mu_R = 0$, $\mu_F = 0$

Standard deviation for ref. group and focal group theta respectively $\sigma_R=\sigma_F= 1$

Impact Case

Mean value for ref. group and focal group theta respectively, $\mu_R = 0$, $\mu_F = -1$.

Standard deviation for ref. group and focal group theta respectively $\sigma_R=\sigma_F= 1$

C. Generating Model: 1PL, 2PL, 3PL

D. Discrimination Levels: .3, .5, .75, .95, 1.25, 1.50, 1.75, 2.0

E. Difficulty Levels: -3, -2, -1.5, -1, -.5, 0, .5, 1, 1.5, 2, 3

F. Number of Items

60 NO DIF ITEMS, 1 DIF ITEM (See Appendix A and B)

G. Item Score Type: Dichotomous

H. Sample Size: 1000

I. Magnitude of DIF

Increments of .025, .05, .10 or .20 (See Appendix B)

Data generation. The IRTGEN software algorithm (Whittaker, Fitzpatrick, Williams, & Dodd, 2003) which incorporates Monte Carlo simulation techniques was used to generate the item responses. IRTGEN generates item responses and known trait scores for the 1PL, 2PL and 3PL models which were necessary for this study.

Power - DFIT

Utilizing the item parameter replication (IPR) method (Oshima et al., 2006) an empirical sampling distribution of NCDIF under the alternative hypothesis was determined. The IPR algorithm already produces an empirical sampling distribution of NCDIF under the null hypothesis. The area beyond the null critical value, under the alternative distribution may be viewed as empirical power. The IPR method currently replicates item parameters for the focal group to build the null distribution for determining the .001, .01, .05 and .10 NCDIF critical values. The IPR method was modified to replicate item parameters for both the focal and reference group to build the alternative distribution.

CHAPTER 4

RESULTS

There was a voluminous amount of data associated with this study; Appendix C summarizes the key data points related to this study. In Appendix C several results are reported for each of the 187 conditions. Appendix C contains only the within condition which corresponded to moderate DIF (Category B) related to the specific condition. Table 7 illustrates an example. Condition 1 is associated with the 1PL model where the b-parameter for the reference group is equal to -3 (see Appendix B). Condition 1 consisted of 40 within conditions by embedding DIF in increments of .10. In Table 7, only 20 of the 40 within conditions are illustrated in an effort to conserve space. See Table 7 where the within condition corresponds to moderate DIF (Category B) and condition 1 in Appendix C.

The definition of moderate DIF as defined by the MH parameter is 1. Large DIF is defined as 1.5. The closest MH parameter value equal to 1 but not equal to or greater than 1.5 was used. All other conditions should be interpreted in a similar manner. There were 46 conditions in which moderate DIF (Category B) could not be accurately estimated. These conditions are easily identified in Appendix C where “Indeterminate” is labeled in the “Congruent” column. In addition, associated with the 46 conditions, 22 of these were 3PL conditions and the MH parameter never reached moderate DIF (Category B).

Table 7

How to Interpret Appendix C – Condition 1

Reference b-parameter (-3)

Focal b-param.	AREA	b-diff	Estimated MH	Estimated NO DIF	Adjusted Estimated MH	True MH	MH (DIF) Category
-2.9	0.10	0.10	-0.949	-0.723	-0.226	-.399	A
-2.8	0.20	0.20	-1.348	-0.624	-0.724	-.799	A
-2.7	**0.30**	**0.30**	**-1.676**	**-0.641**	**-1.035**	**-1.198**	**B**
-2.6	0.40	0.40	-2.084	-0.532	-1.552	-1.598	C
-2.5	0.50	0.50	-2.445	-0.576	-1.869	-1.997	C
-2.4	0.60	0.60	-2.763	-0.455	-2.308	-2.397	C
-2.3	0.70	0.70	-3.121	-0.573	-2.548	-2.797	C
-2.2	0.80	0.80	-3.511	-0.483	-3.028	-3.196	C
-2.1	0.90	0.90	-3.888	-0.364	-3.524	-3.596	C
-2.0	1.00	1.00	-4.222	-0.378	-3.844	-3.995	C
-1.9	1.10	1.10	-4.603	-0.402	-4.201	-4.394	C
-1.8	1.20	1.20	-4.960	-0.234	-4.726	-4.794	C
-1.7	1.30	1.30	-5.379	-0.320	-5.059	-5.193	C
-1.6	1.40	1.40	-5.727	-0.314	-5.413	-5.992	C
-1.5	1.50	1.50	-6.104	-0.278	-5.826	-6.392	C
-1.4	1.60	1.60	-6.525	-0.263	-6.262	-6.792	C
-1.3	1.70	1.70	-6.884	-0.247	-6.637	-7.191	C
-1.2	1.80	1.80	-7.327	-0.199	-7.128	-7.591	C
-1.1	1.90	1.90	-7.705	-0.171	-7.534	-7.990	C
-1.0	2.00	2.00	-8.063	-0.204	-7.859	-8.389	C

Furthermore, the results in Appendix C correspond to the unequal ability distribution investigation. Corresponding to the identification of moderate DIF in Appendix C, the corresponding (a) area measure is calculated; (b) difference in difficulty level is reported ($b_f - b_r$); (c) estimated MH statistic which is based on the average of 100 replicates; (d) estimated MH statistic for the "No DIF" condition, which is also based on the average of 100 replicates; (e) adjusted estimated MH statistic which is the difference between the estimated MH statistic and "No DIF" condition; (f) true parameter for the within condition; and (g) congruency indicator.

For each of the 5750 DIF items, the true MH parameter was also determined. Congruency met is defined for this study as, when the adjusted estimated MH statistic agrees with the true parameter related to the size of DIF for a given condition (i.e. Negligible, Moderate or Large). As an example, the MH estimate for condition 1 is -1.035 and the corresponding MH true parameter is -1.198. Related to the size of DIF both the adjusted estimated statistic and true parameter are considered moderate DIF (Category B), see Appendix C.

As did Allen and Donoghue (1996) in their study, the MH statistic estimate for this study was determined by also simulating for each within condition the "No DIF" scenario, hereafter referred to as the null condition. By subtracting the null condition from the MH estimate, an adjusted MH estimate is reported. Roussos, Schnipke, and Pashley (1999) referred to this null condition as a rough estimate of the bias associated with estimating the true MH parameter Δ.

Effect Size Recommendation for NCDIF

The effect size recommendation is based on the fact that a clear relationship exists between the MH parameter and the NCDIF parameter. The Monte Carlo simulation study and MH parameter software produced 10, 20, 40, or 60 data points for the MH statistic and parameter for each of the 5750 DIF items. Equations 26 and 27 were used to calculate true NCDIF for these same items.

Scatterplots showed that the relationship between the two measures was curvilinear in nature, see Figure 3. Only condition 4 is illustrated, but all of the conditions investigated revealed through scatter plots a curvilinear relationship between MH and NCDIF.

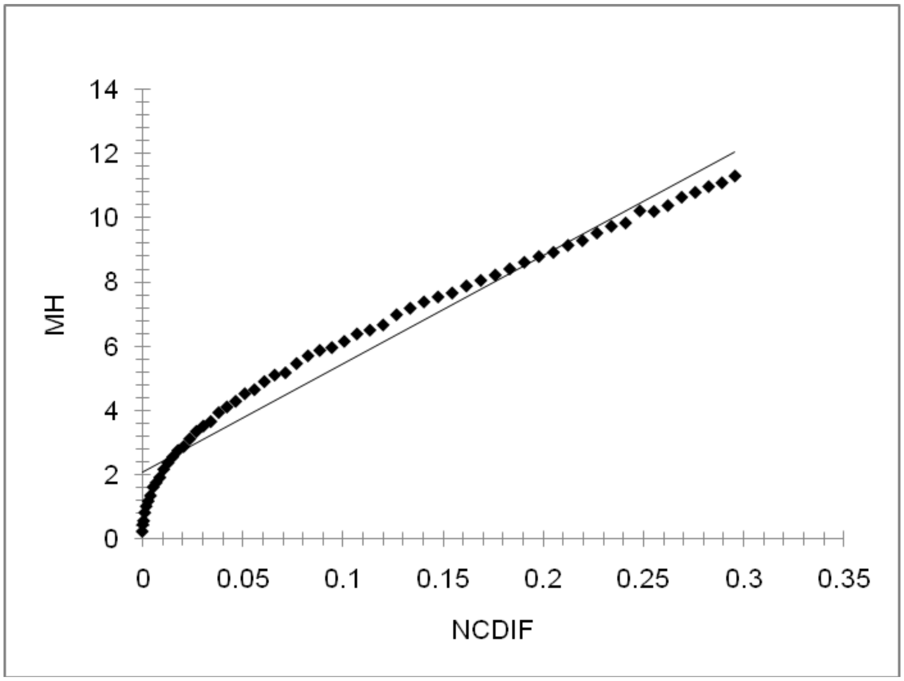


Figure 3. Scatter Plot (NCDIF without transformation) – Condition 4 (See Appendix B)

Polynomial block regression analysis was applied to each condition. Related to condition 4, the linear component accounted for 96% of the variance $F(1, 58) = 1302$; $R^2 = .96$, $p < .01$. The quadratic component was entered in the second step; it accounted for an additional 3 percent of the variance, R^2 change = .03, $F(1, 57) = 1075$, $p < .01$. The cubic component was entered in the third step which accounted for a very small percentage of the variance, but significant, $F(1, 56) = 282$, $p < .01$. Each of the three beta coefficients were significant, p < .01. The quadratic component was statistically significant for all of the conditions. The cubic component was statistically significant for approximately 70% of the conditions, but in all cases explained a very small percentage of the variance between the two measures.

The linear and quadratic components explained almost 100% of the variance between the two statistics revealed through the polynomial block regression analyses. It was then determined that a simpler approach could be used to correct the curvilinear relationship. NCDIF by definition is the average squared distance between the focal and reference group's ICCs. Applying a nonlinear transformation to NCDIF by taking the square root of each data point produced an acceptable linear relationship, see Figure 4.

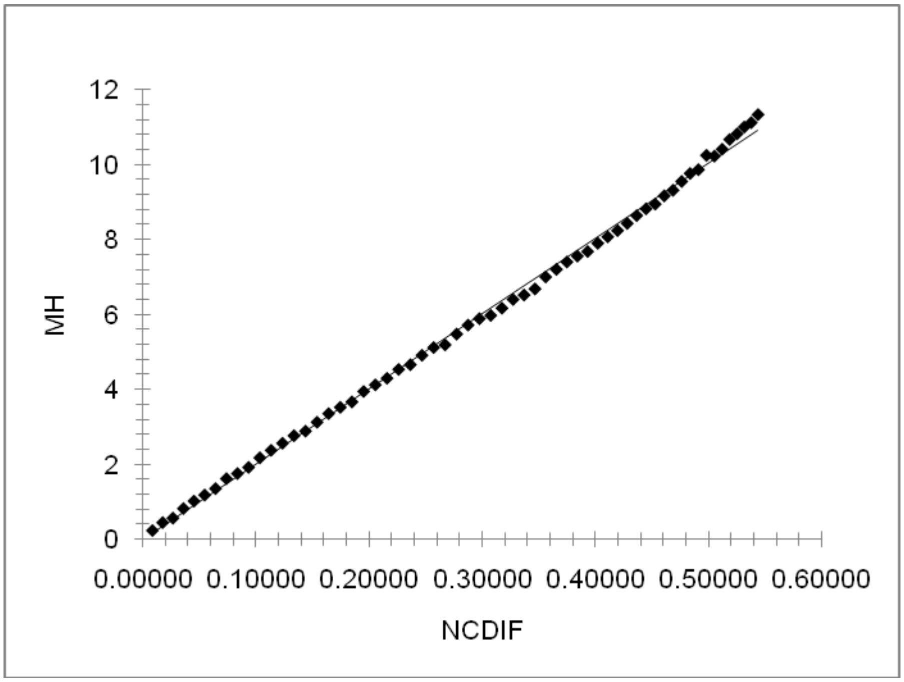


Figure 4. Scatter Plot (NCDIF with transformation) – Condition 4 (See Appendix B)

Each of the other conditions had similar results after applying the transformation. Correlation matrices are provided in Appendix D for several of the conditions. The conditions are identified by the condition number. For each condition, the correlation between the MH parameter and NCDIF was .87 or higher, with the majority being .99 after the transformation. In general, the 3PL conditions had the lower correlation indexes. In this study the a-parameter was held constant between the focal and reference groups, hence, essentially modeling a special case of the 1PL model. Past research has showed the MH statistic to be reliable for 1PL and 2PL data.

The recommended effect sizes for DFIT's NCDIF are presented in Tables 8, 9, 10, 11, 12 and 13. The constants in Tables 12 and 13 indicate a one-size-fits-all approach is not advisable. The effect size of NCDIF is influenced by the model, the discrimination parameter and the difficulty parameter.

Table 8

NCDIF Recommended Effect Size – Moderate DIF: Category B (1PL/2PL)

Model	*1PL*	*2PL*							
Discrimination	N/A	0.30	0.50	0.75	0.95	1.25	1.50	1.75	2.00
Difficulty									
-3	<.001	0.005	0.001	0.001	<.001	<.001	<.001	<.001	<.001
-2	0.001	0.006	0.003	0.002	0.001	0.001	0.001	<.001	<.001
-1.5	0.001	0.007	0.005	0.003	0.002	0.001	0.001	0.001	0.001
-1	0.002	0.009	0.007	0.005	0.004	0.002	0.002	0.001	0.001
-0.5	0.003	0.009	0.008	0.006	0.006	0.003	0.003	0.002	0.002
0	0.003	0.009	0.008	0.007	0.006	0.003	0.004	0.004	0.002
0.5	0.002	0.008	0.007	0.006	0.006	0.003	0.003	0.002	0.002
1	0.002	0.007	0.006	0.004	0.003	0.003	0.002	0.001	0.001
1.5	0.001	0.007	0.004	0.003	0.002	0.001	0.001	<.001	<.001
2	0.001	0.006	0.003	0.001	0.001	<.001	<.001	<.001	<.001
3	<.001	0.003	0.001	<.001	<.001	<.001	<.001	<.001	<.001

For values < .001, see Appendix F, Detailed Values for NCDIF, seven decimal places listed.

Table 9

NCDIF Recommended Effect Size – Moderate DIF: Category B (3PL)

Model	*3PL*							
Discrimination	0.30	0.50	0.75	0.95	1.25	1.50	1.75	2.00
Difficulty								
-3	0.003	0.001	<.001	<.001	<.001	<.001	<.001	<.001
-2	0.007	0.003	0.001	0.001	0.001	0.001	<.001	<.001
-1.5	0.008	0.006	0.002	0.002	0.001	0.001	0.001	0.001
-1	0.009	0.006	0.004	0.003	0.003	0.002	0.001	0.001
-0.5	0.011	0.010	0.008	0.007	0.006	0.005	0.004	0.004
0	0.010	0.010	0.009	0.011	0.009	0.008	0.009	0.009
0.5	0.010**	0.014	0.014	0.013	0.016	0.017	0.018	0.019
1	0.010**	0.014**	0.014**	0.013**	0.016**	0.017**	***	***
1.5	0.010**	0.014**	0.014**	***	***	***	***	***
2	0.010**	***	***	***	***	***	***	***
3	***	***	***	***	***	***	***	***

For values < .001, see Appendix F, Detailed Values for NCDIF, seven decimal places listed
**MH underestimation of DIF. Used the preceding NCDIF value where area ≤ .80
***MH never reached moderate DIF (Category B) - Indeterminate

Table 10

NCDIF Recommended Effect Size – Large DIF: Category C (1PL/2PL)

Model	1PL	2PL							
Discrimination	N/A	0.30	0.50	0.75	0.95	1.25	1.50	1.75	2.00
Difficulty									
-3	0.001	0.011	0.002	0.002	<.001	<.001	<.001	<.001	<.001
-2	0.002	0.014	0.007	0.005	0.002	0.002	0.002	0.001	0.001
-1.5	0.002	0.016	0.011	0.007	0.005	0.002	0.002	0.002	0.002
-1	0.005	0.020	0.016	0.011	0.009	0.005	0.005	0.002	0.002
-0.5	0.007	0.020	0.018	0.014	0.014	0.007	0.007	0.005	0.005
0	0.007	0.020	0.018	0.016	0.014	0.007	0.009	0.009	0.005
0.5	0.005	0.018	0.016	0.014	0.014	0.007	0.007	0.005	0.005
1	0.005	0.016	0.014	0.009	0.007	0.007	0.005	0.002	0.002
1.5	0.002	0.016	0.009	0.007	0.005	0.002	0.002	0.001	0.001
2	0.002	0.014	0.007	0.002	0.002	0.001	<.001	<.001	<.001
3	<.001	0.007	0.002	0.001	<.001	<.001	<.001	<.001	<.001

For values < .001, see Appendix F, Detailed Values for NCDIF, seven decimal places listed

Table 11

NCDIF Recommended Effect Size – Large DIF: Category C (3PL)

Model	3PL							
Discrimination	0.30	0.50	0.75	0.95	1.25	1.50	1.75	2.00
Difficulty								
-3	0.007	0.002	0.001	<.001	<.001	<.001	<.001	<.001
-2	0.016	0.007	0.002	0.002	0.002	0.002	<.001	<.001
-1.5	0.018	0.014	0.005	0.005	0.002	0.002	0.002	0.002
-1	0.020	0.014	0.009	0.007	0.007	0.005	0.002	0.002
-0.5	0.025	0.023	0.018	0.016	0.014	0.011	0.009	0.009
0	0.023	0.023	0.020	0.025	0.020	0.018	0.020	0.020
0.5	0.023**	0.032	0.032	0.029	0.036	0.038	0.041	0.043
1	0.023**	0.032**	0.032**	0.029**	0.036**	0.038**	***	***
1.5	0.023**	0.032**	0.032**	***	***	***	***	***
2	0.023**	***	***	***	***	***	***	***
3	***	***	***	***	***	***	***	***

For values < .001, see Appendix F, Detailed Values for NCDIF, seven decimal places listed.
**MH underestimation of DIF. Used the preceding NCDIF value where area ≤ .80
***MH never reached moderate DIF (Category B) - Indeterminate

Table 12

1PL/2PL Linear Constant Relating - NCDIF = (MH / K)^2

Model	1PL	2PL							
Discrimination		0.30	0.50	0.75	0.95	1.25	1.50	1.75	2.00
Difficulty									
-3	52	14	32	32	93	142	190	235	528
-2	32	13	18	22	32	32	32	61	65
-1.5	32	12	14	18	22	32	32	32	32
-1	22	11	12	14	16	22	22	32	32
-0.5	18	11	11	13	13	18	18	22	22
0	18	11	11	12	13	18	16	16	22
0.5	22	11	12	13	13	18	18	22	22
1	22	12	13	16	18	18	22	32	32
1.5	32	12	16	18	22	32	32	46	55
2	32	13	18	32	32	55	68	85	103
3	70	18	32	64	103	213	309	439	587

Table 13

3PL Linear Constant Relating - NCDIF = (MH / K)^2

Model	3PL							
Discrimination	0.30	0.50	0.75	0.95	1.25	1.50	1.75	2.00
Difficulty								
-3	18	32	62	116	178	238	294	660
-2	12	18	32	32	32	32	77	82
-1.5	11	13	22	22	32	32	32	32
-1	11	13	16	18	18	22	32	32
-0.5	10	10	11	12	13	14	16	16
0	10	10	11	10	11	11	11	11
0.5	10	8	8	9	8	8	7	7
1	10	8	8	9	8	8	***	***
1.5	10	8	8	***	***	***	***	***
2	10	***	***	***	***	***	***	***
3	***	***	***	***	***	***	***	***

***MH never reached moderate DIF (Category B) - Indeterminate

Appendix E illustrates the relationship between the model, discrimination parameter and the difficulty parameter for the 187 conditions. Several relationships are apparent: (a) at difficulty level of b = 0, the NCDIF value at this point is either equal to or higher than at any other difficulty level for the 1PL and 2PL conditions. Given that in this study the mean ability distributions were *N*(0, 1) and *N*(-1, 1), the majority of the examinees would be in this region. Therefore, at the extreme ends of the difficulty levels, the NCDIF value will be lower; (b) related to the discrimination parameter, NCDIF value is highest at the lowest discrimination value, and decreasing as the discrimination parameter increases; and (c) related to model, the 2PL/3PL NCDIF values are equal or differ by no more than .001 until the difficulty level is approximately b=0. At this point, as the difficulty level increases, the 3PL NCDIF values are significantly higher; a possible explanation is the psudeo-gusessing parameter.

The noise associated with random guessing may be contributing to the difficulty in measuring DIF between the focal and reference groups (Donoghue, Holland, & Thayer, 1993; Lord, 1980). Zwick, Thayer, and Wingersky (1994) provide this as a possible explanation, “the more difficult the item, the closer the probability of correct response is to guessing value, and the more difficult the groups are to differentiate” (p. 135). Roussos et al. (1999) debunk this hypothesis because the same phenomenon is not happening with easy 3PL items. Roussos et al. study demonstrated that the very parameter being estimated is shrinking with increased difficulty, where sparseness of examinees is not an issue. This study corroborates Roussos et al.’s findings. NCDIF is based on where MH is reporting moderate DIF (Category B), and MH may not be reliable for specific conditions. Figure 5 illustrates these observed relationships for one condition where a = .95. The 2PL case is represented by the solid line; conversely the 3PL case is represented by the dash line.

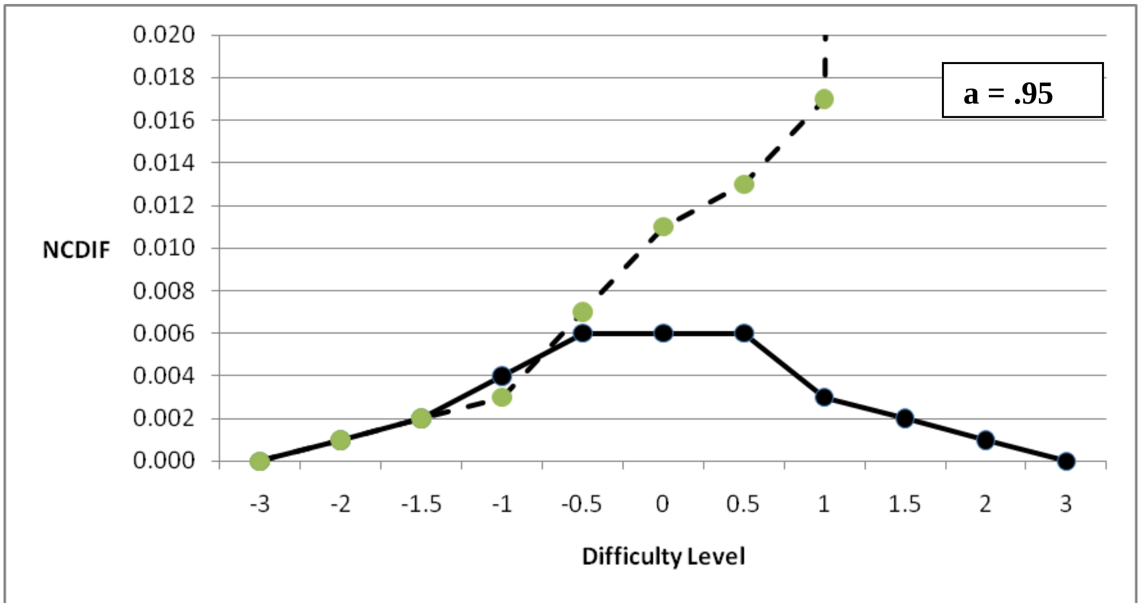


Figure 5.
Relationship between NCDIF (Moderate DIF) and Difficulty Level by Model

In Equation 36, MH is equal to 1 for moderate DIF (Category B), MH is equal to 1.5 for large DIF (Category C), and K is a constant, see Tables 12 and 13.

$$NCDIF = (MH / K)^2 \quad (36)$$

There were 29 conditions where the MH estimate corresponded to moderate DIF (Category B) size, where the corresponding NCDIF value was less than .001; see Tables 8, 9, and also Appendix F for these conditions and more specific NCDIF values. The null condition for NCDIF is 0, and the MH estimate is reporting for these cases moderate DIF. Recall, for the reference and focal groups the ability (*θ)* distributions for this study were randomly drawn as *N*(0, 1) and *N*(-1, 1) respectively. In applying Lord's (1980) formula, *a(θ –b)* to each of these conditions, the corresponding z-scores will be on the extreme ends of the distributions. The number of examinees in the extreme regions are limited, hence, the very small NCDIF values.

The corresponding difference in difficulty level ($b_f – b_r$) for the two groups for these conditions is between .16 and .30, which may be an indication of differential item functioning between the two groups. Given the lack of agreement between NCDIF and MH in interpreting the effect size for these conditions, the following guidelines are recommended for moderate DIF (Category B): (a) significance is reported for these conditions; and (b) empirically observed power is .80.

Equal Ability Distributions. In concluding the effect size recommendation for NCDIF, it is important to note that as part of this study, equal ability distributions were also investigated. In investigating equal ability distributions, the reference and focal groups' ability (*θ)* distributions were randomly drawn as *N*(0, 1) and *N*(0, 1) respectively. The same Monte Carlo procedures were applied. Figure 6 illustrates that the results in Appendix C would be identical for the equal ability distribution case. In plotting the relationship between the pairs of MH estimates 5040 for the equal ability conditions and 5040 for the unequal conditions, the Pearson r coefficient indicates an almost perfect relationship. This was further corroborated by the fact that for each of the 116 out of 187 estimated, the MH estimate for the equal and unequal conditions converged at the same location for reporting moderate (Category B) and large (Category C) DIF. Prior research (Spray & Miller, 1992; Donoghue, Holland, & Thayer, 1993; Demars, 2009) supports these findings.

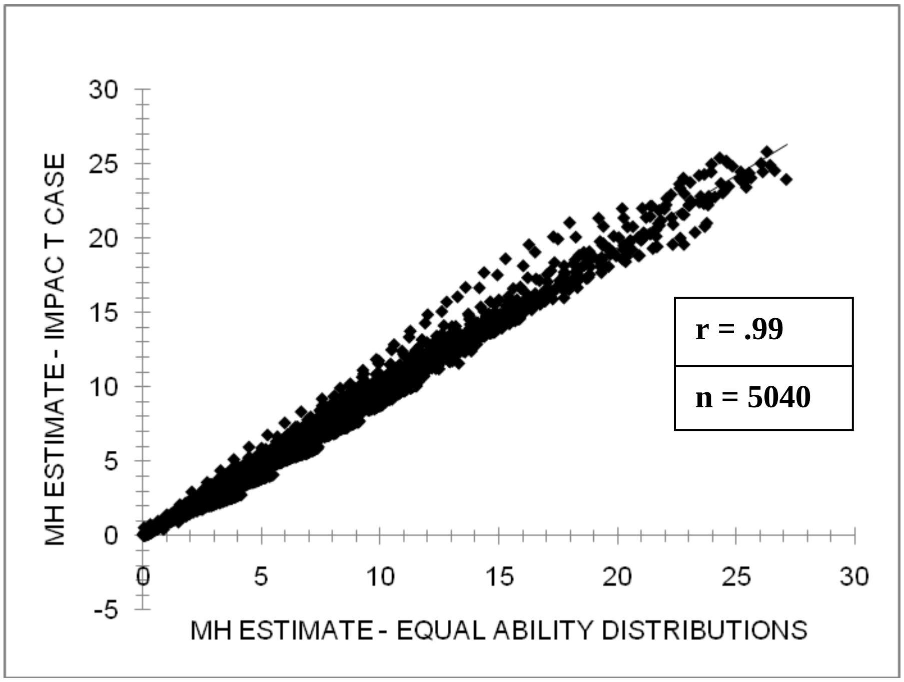


Figure 6.
Scatter Plot- MH Estimates Impact versus MH Estimates No Impact.

SIBTEST

An effect size recommendation is not being made based on the results of this study for SIBTEST. The purpose of including SIBTEST in the investigation was for comparison purposes only and an evaluation of previously established guidelines based on the MH statistic. The effect sizes based on this study for SIBTEST are presented in Table 14 and Appendix G. Equal and unequal ability distributions were also investigated for SIBTEST. In Equation 37, MH is equal to 1 for moderate DIF (Category B), MH is equal to 1.5 for large DIF (Category C), and K is a constant, see Table 14.

$$\text{SIBTEST} = (\text{MH} / \text{K}) \qquad (37)$$

Table 14

Impact Condition: Linear Constant Relating - SIBTEST = (MH / K)

Model	1PL	2PL							
Discrimination	N/A	0.30	0.50	0.75	0.95	1.25	1.50	1.75	2.00
Difficulty									
-3	17	10	11	14	16	20	22	25	26
-2	14	10	11	13	14	16	18	20	22
-1.5	13	10	10	11	13	14	16	17	19
-1	12	10	11	13	14	16	17	19	20
-0.5	16	10	11	13	15	18	22	24	27
0	27	11	13	19	24	37	48	77	94
0.5	33	9	13	20	30	54	86	N/A	N/A
1	72	10	19	32	58	106	193	N/A	N/A

Model	3PL							
Discrimination	0.30	0.50	0.75	0.95	1.25	1.50	1.75	2.00
Difficulty								
-3	11	12	12	15	13	18	18	10
-2	10	11	12	12	11	11	13	13
-1.5	10	11	11	12	11	12	12	12
-1	10	11	12	13	13	14	14	15
-0.5	10	11	11	15	13	19	20	21
0	11	13	16	20	27	33	41	20

A correlation matrix is provided in Appendix D relating the two statistics MH and SIBTEST. The correlation matrix is only for the impact conditions. The correlations for the 116 conditions estimated range from a low of .81 to a high of 1.0. There were 92% of the conditions which had a correlation index of .9 or higher. These results support a previous finding (Shealy & Stout, 1993). The conditions where the correlations were lower than .9 were 48, 54, 56, 60, 61, 62, 66, 67, and 68. For these conditions the scatter plots revealed a curvilinear relationship, typically in the middle of the data points or at the tail end of the data points, see Figure 7. This observation had not previously been noted based on a limited review of the literature.

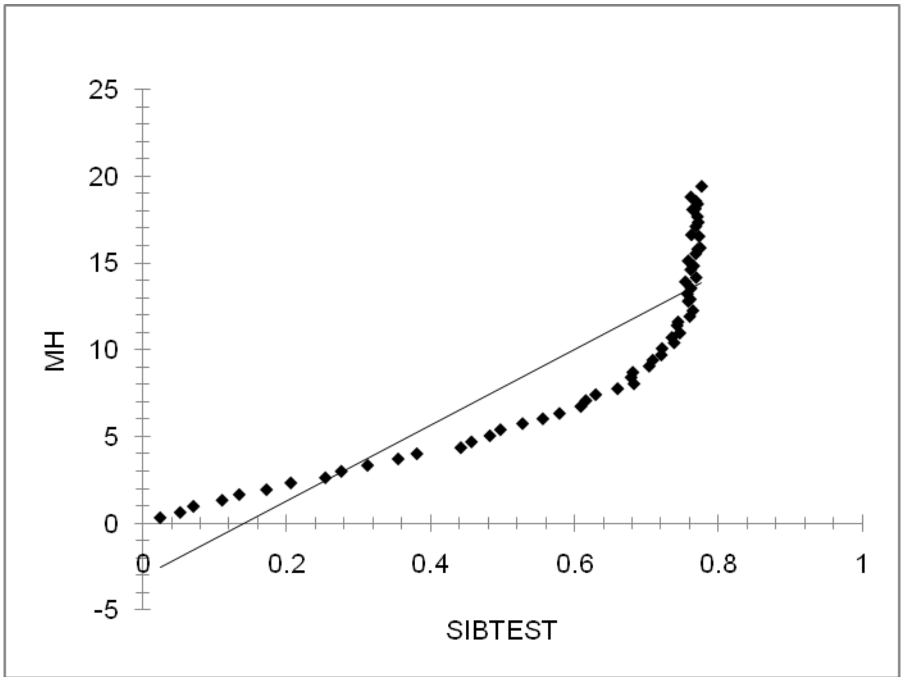


Figure 7. Scatter Plot - SIBTEST – Condition 82 (See Appendix B)

Each of these conditions, hence, test items are considered hard or either highly discriminating. In a simulation study investigating Type I error performance associated with MH and SIBTEST, Roussos and Stout (1996b) reported inflated Type I error rates for MH. In their study, Type I error was reported as .26 for condition 135; see Appendix

B. In this study, Type I error was also calculated for each condition, and for condition 135 which is identical to Roussos and Stout's condition, the Type I error rate was .23. This is an important observation related to how MH estimates DIF items considered extremely easy, hard or highly discriminating, when the two groups being studied ability distributions are incongruent.

Shealy and Stout (1993) used a constant of -15 for the 1PL/2PL models in relating an effect size for SIBTEST based on the MH parameter. Roussos and Stout (1996b) used a constant of -17 for the 3PL model. Based on this study a one-size-fit-all approach may not be advisable, see Table 14. As stated when discussing an effect size recommendation for NCDIF, the size of DIF is influenced by the model, the discrimination parameter and the difficulty parameter.

Area Measure

Area measure also served for comparison and observational purposes. The results for the area measure calculations for the 116 estimated conditions are presented in Figure 8, Figure 9, and Appendix C. These area measure calculations correspond to where the adjusted MH estimates corresponded to the moderate DIF location (Category B). The histograms in Figures 8 and 9 provide frequencies for the 2PL and 3PL conditions related to moderate DIF based on the MH estimate. There were 8 1PL conditions, the area measures were approximately .30 for all 8 conditions, see Appendix C. If using area measure to interpret the size of DIF, Swaminathan and Rogers (1990) used the guideline for medium DIF as .6. Using this point of view, it was expected for the histograms to peak around .6. Given the conditions in this study, area measure related to MH's definition of moderate DIF does not peak at .6, but appears to be a function of the model, difficulty parameter and discrimination parameter (see also Appendix C column labeled "Area Measure").

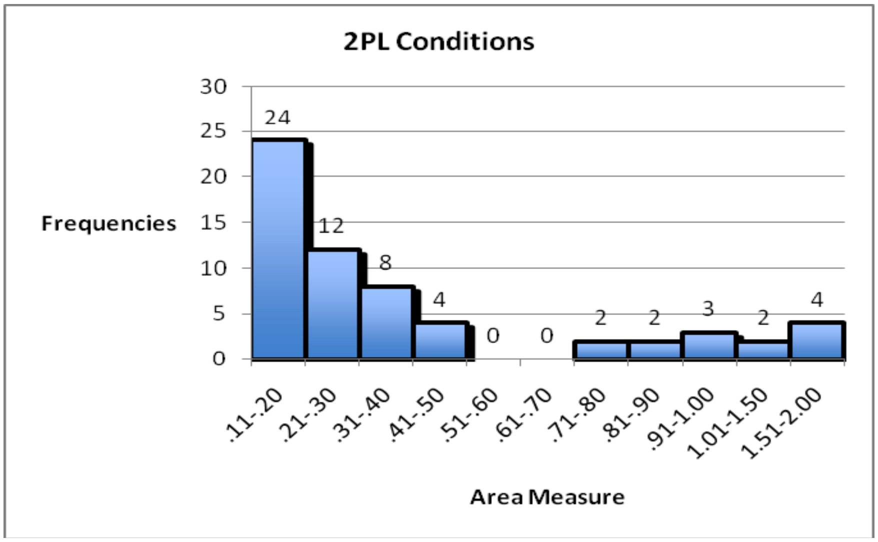


Figure 8.
Area Measure frequencies of 2PL Conditions.

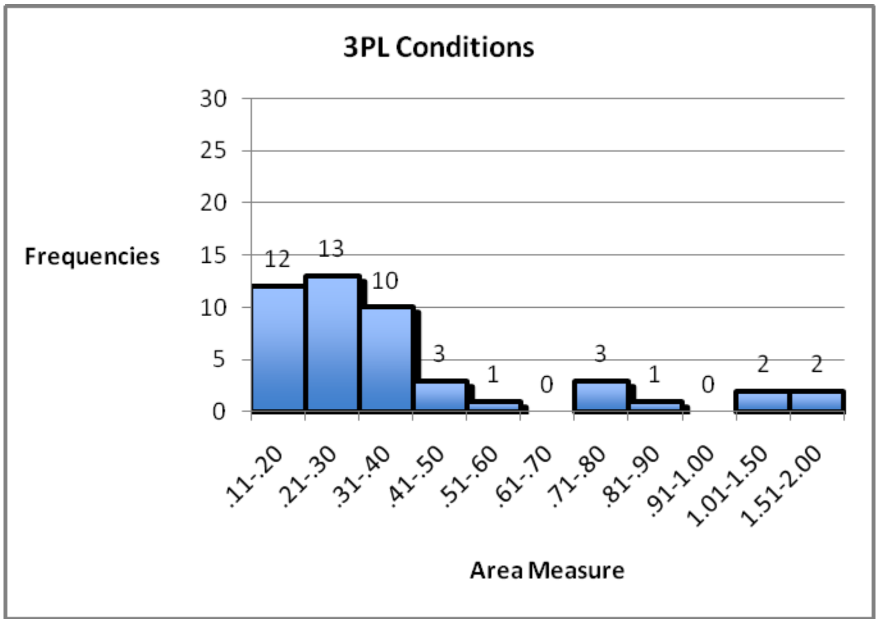


Figure 9.
Area Measure frequencies of 3PL Conditions.

Empirical Observed Power

The NCDIF statistical test is based on the item parameter replication algorithm (IPR). Essentially, using the focal group's item parameters for a test item, 1000 pairs of these parameters are reproduced. NCDIF for each of these pairs is calculated. These replicated pairs represent the "No DIF" condition, and hence, any extreme differences observed would be considered beyond chance. The 1000 pairs form the null distribution, and cutoffs are determined at the 90%, 95%, 99% and 99.9% percentile rank scores. The NCDIF values at any of these levels will be used to determine statistical significance at .10, .05, .01, and .001, respectively. For a detailed description of the IPR procedure, see Oshima, Raju, and Nanda (2006). In modifying Oshima et al. (2006) item parameter replication algorithm (IPR), an empirical sampling distribution of NCDIF under the alternative hypothesis was determined.

In determining the alternative distribution, the IPR algorithm was modified to reproduce 1000 pairs of the focal group and reference groups' item parameters. These pairs of parameters represent the DIF case, and the NCDIF value determined using these pairs represent a distribution under the alternative hypothesis. The power of a statistical test in this study is defined by the probability of correctly rejecting a false null condition when NCDIF is not 0. The probability of correctly rejecting a false null condition is determined by calculating the area to the right of the null distribution, related to the alternative distribution for the specified alpha level for the statistical test. Cohen (1988) provided power tables for other statistical test (e.g. Student's t-Ratio). Also, there are many applets available for calculating power. The uniqueness of the IPR method made calculating empirical observed power simple.

The area to the right of the NCDIF alternative distribution was calculated by determining the number of NCDIF values under the alternative distribution which is greater than the NCDIF value (on the Null Distribution) at .05 divided by 1000, see Figure 10. For simulated example 2 (see Table 15), the NCDIF value under the null condition at α = .05 was .001.

Table 15
Results – Empirical Observed Power (α = .05)

	Ref. b-param.	Foc. b-param.	b-diff.	Est. NCDIF	True NCDIF	Null Distribution NCDIF Value α = .05	Power
#1	-3	-2.7	.3	.0002	.0003	.00065	19%
#2	-3	-2.4	.6	.003	.002	.001	90%
#3	-3	-2.2	.8	.004	.003	.004	98%
#4	0	.3	.3	.003	.003	.001	96%

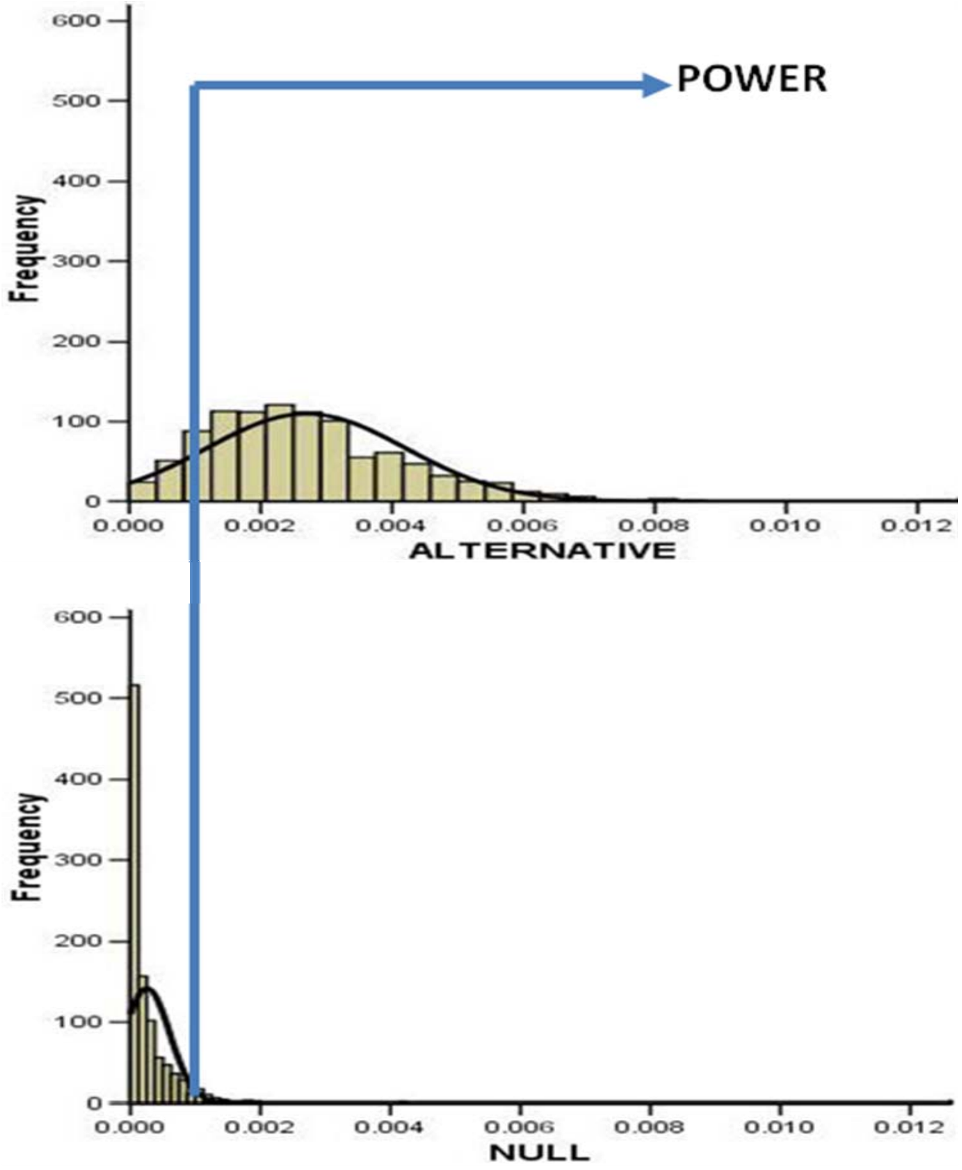


Figure 10.
Empirical Null and Alternative Distribution (Table 15- #2)

In having the null distribution and the alternative distribution, empirical observed power was estimated for two of the 187 conditions. Three of the within conditions for condition 1 and 1 of the within conditions for condition 6 are presented in Table 15.

Condition 1 was selected based on what has already been discussed related to easy test items. Condition 6 was selected based on a difficulty level of 0 representing the ability level of the majority of the 1000 examinees. Recall, the NCDIF value for condition 1 was less than .001 where the b-difference between the focal group and reference group was .30; MH corresponds to a b-difference of .30 to be moderate DIF. As an example of how empirical observed power was determined, for number 2 in Table 15, NCDIF value at α = .05 under the null distribution was .001. There were 895 NCDIF values equal to or greater than .001 under the alternative distribution (see Figure 10), therefore, power would equal 895/1000 or 90%. As would be expected as the b-difference in difficulty level increases between the two groups, hence, essentially an effect size increase, power increase gradually. Examples 1, 2 and 3 illustrate the increase in power. As the effect size increases (i.e. b difference) the statistical test would have more power in accurately identifying a departure from the null hypothesis. Finally, increasing the sample size would also increase power. Related to example 1 and example 4, both are related to a b-difference of .30, but starkly different power. Example 4 is related to condition 6 where the b-parameter equals 0. Discussed earlier, given the mean ability distributions chosen for this study, there would be more examinees in this region, hence, power increases as the sample size increases.

Summary

The primary goal of this study was to determine an effect size for NCDIF, whereby the MH parameter served as the benchmark. The effect size for NCDIF is based on several factors investigated in this study (see Table 6).

The MH measure has sporadic behavior for easy and hard test items, also for low and highly discriminating test items. This behavior should not be surprising that the MH measure does not work well as a function of discrimination, given that it was designed for 1PL data. This sporadic behavior was considered in recommending an effect size for NCDIF. In the cases where the MH measure underestimated the size of DIF, the effect size for NCDIF is based on the preceding NCDIF effect size recommendation, where the area measure was calculated to be less than or equal to .80. Furthermore, in the cases where the MH measure never reached moderate DIF (Category B), the effect size guidelines are based on statistical signicance and empirically observed power.

CHAPTER 5

DISCUSSION

This study investigated the addition of reporting an effect size measure for DFIT's NCDIF and reporting empirically observed power. The MH parameter served as the benchmark for developing NCDIF's effect size measure, for reporting moderate and large differential item functioning in test items. In addition, by modifying NCDIF's unique method for determining statistical significance, NCDIF will be the first DIF statistic of test items where in addition to reporting an effect size measure, empirical power can also be reported (see Appendix H). This study added substantially to the body of literature on effect size by also investigating the behavior of two other DIF measures, SIBTEST and area measure. Finally, this study makes a significant contribution to the body of literature by verifying in a large-scale simulation study the accuracy of software developed by Roussos, Schnipke, and Pashley (1999) to calculate the true MH parameter; see Equation 34. The accuracy of this software had not been previously verified in a large-scale simulation study.

Behavior of MH Measure

In determining a comparable effect size for DFIT's NCDIF, the MH statistic which is widely used today served as the benchmark for this study. It is important to understand the results already presented related to the behavior of the MH parameter.

There is a plethora of empirical research on the MH statistic in observing its behavior (Holland & Thayer, 1988; Donoghue, Holland & Thayer, 1993; Clauser, Mazor & Hambleton, 1994; Allen & Donoghue, 1996; Roussos & Stout, 1996b; Roussos, Schnipke, & Pashley, 1999).

Donoghue, Holland and Thayer (1993) determined that the MH statistic $\overset{\wedge}{\Delta}$ which estimates the underlying parameter Δ can be explained by Equation 38 for 1PL and 2PL models. Equation 38 does not hold true for 3PL data which has also been verified by Roussos, Schnipke, and Pashley (1999). In Equation 38, "a" is common for all test items, and "b" is defined by $b_f - b_r$ (i.e. the difference in difficulty) for the studied test item between the focal and reference group examinees. In Equation 38, "b" is the difference in difficulty between the focal and reference groups' b-parameter.

$$\Delta = -4ab \qquad (38)$$

As noted by Donoghue, Holland and Thayer, several conditions must be satisfied: (a) the a-parameter is common for both groups; (b) the studied item is included when matching the focal and reference group examinees on ability; and (c) none of the other test items used to match examinees are contaminated with DIF. These three conditions were satisfied for this study. The relationship expressed in (38) was observed for many of the 1PL and 2PL conditions considering estimation error. The exceptions were conditions 16, 17, 18, 19, 27, 28, 29, 30, 89, 90. These conditions are either low or highly discriminating test items. Furthermore, the b-parameter for conditions 16 - 19 and 27 - 30 range from -.5 to 1. Conditions 89 and 90 have b-parameters of -3 and -2 respectively. In Allen and Donoghue (1996), it was purported that a b-parameter of 0, 1, or 2 corresponds respectively with a z-score of .875, 2.125 and 3.375.

Given these z-scores, the area to the right, hence, the number of examinees in this region, is limited. Allen and Donoghue further assert that given difficult test items, “it is not surprising that MH has little power to detect DIF” (p. 248). Although not stated by Allen and Donoghue, this should also apply to easy test items. The conditions noted as exceptions to Equation (38) which relates to the 1PL and 2PL models would all have z-scores approximately at or above ±.875, hence a possible explanation to the underestimation of the true parameter.

In concluding the discussion on the behavior of MH, the MH measure of DIF overestimates the amount of DIF for easy and hard test items related to the 1PL and 2PL models. MH overestimates the amount of DIF for easy test items related to the 3PL model. Once the b-parameter difficulty level increases for the 3PL model, MH underestimates the amount of DIF for hard test items. This behavior was identified in another study by Donoghue, Holland and Thayer (1993), in which the behavior is contributed to the fact of using a fixed c-parameter for the reference and focal groups. This study utilized a fixed c-parameter.

Why Use MH for Determining NCDIF’s Effect Size

DIF studies are conducted by large-scale testing organizations such as ETS the makers of many high-stakes exams. These exams are used for entry into institutions of higher education, K-12 statewide assessments, etc. The MH statistic has been used for over a half century as a tool for assessing DIF. Practitioners in K-12 education are very familiar with its use and interpretation of measuring the size DIF for test items.

SIBTEST and Logistic Regression are other statistical techniques for assessing DIF of test items, and have also based its results on the MH guidelines of, (a) negligible DIF (Category A); (b) moderate DIF (Category B); and (c) large DIF (Category C). DIF studies are critically important related to standardize testing. In an effort to ensure fairness related to standardize testing, more than one method should be employed. If NCDIF is going to become a statistical tool of choice for measuring DIF, being able to interpret the size of DIF using already familiar guidelines is important.

General Discussion on Effect Size

Today, an effect size measure is of critical importance. In the 6th edition of the APA Publication Manual, reporting an effect size measure is recommended (APA, 2009). Differential item functioning of test item studies requires large sample sizes, hence, a potential propensity to report significance for practically insignificant results. Most importantly, large-scale testing companies typically only discard test items which display moderate to large DIF. An effect size measure in conjunction with a significant finding today is necessary, especially in DIF studies.

This study revealed that many factors influence the size of DIF, and one size does not fit all. Furthermore, the agreement of the size of DIF is complicated given that each of the measures investigated in this study measures DIF using a different scale as discussed in Chapter 2. Based on the MH guidelines for judging moderate to large DIF, these same test items would be considered negligible DIF (Category A) when using area measure guidelines of .6 and .8 respectively. Previous research provided guidelines for paralleling SIBTEST measure of DIF with MH.

This study revealed that those previously established guidelines may be too general. It was corroborated in this study where MH behavior becomes unstable for specific test items. A goal of this study was also to parallel NCDIF measure of DIF with MH.

Effect Size Recommendation

The effect size recommendations for NCDIF are based on many factors considered in this investigation. There were 11 different difficulty levels investigated, 8 different discrimination levels, and 3 ICC models (1PL, 2PL and 3PL). The effect size recommendations will allow researchers and practitioners the ability to provide an integrity check if using MH and NCDIF to evaluate differential item functioning in test items. Given the importance of balancing test fairness and the cost of constructing test items, it is highly recommended to use more than one measure to evaluate DIF. This is being done today at ETS by using the STD-P difference in conjunction with MH (Sinharay & Dorans, 2010) given the unstable behavior of MH with certain types of test items. This study will now allow NCDIF to be used in conjunction with MH in evaluating DIF. Finally, in addition to reporting statistical significance and the effect size of DIF, researchers and practitioners will now be able to judge how much power the statistical test had in assessing DIF.

Limitations and Directions for Future Research

In this study, unequal sample sizes between the focal and reference groups were not considered.

The discrimination parameter in this study was fixed for the focal and reference groups, hence, non-uniform DIF was not investigated. The pseudo-guessing parameter was also fixed for the focal and reference groups. DIF was embedded in only one test item, and all other test items were free of DIF which does not consider contamination of a test. Prior to calculating NCDIF, both the focal and reference groups' ability estimates must be put on the same scale. The true NCDIF parameter was calculated in this study which does not factor in linking error when placing the ability estimates on the same scale. Future studies can investigate the impact of these factors on the recommendations developed for this study.

Conclusion (Yesterday, Today, and Tomorrow)

In the 1970s the U.S. government saw standardized testing as a means to ensure its scientific competitiveness in the world during the accountability era (Pulliam & Van Patten, 1999). The 1970s also saw increased attention to standardized testing by the state governments. State governments were also funding public schools, therefore, similar to Title I from a federal perspective, states also were holding schools accountable for receiving state funds. Colleges were still utilizing standardized scores for evaluating applicants, but reliance solely on them had not yet developed. The 1980s ushered in two significant events impacting standardized testing. The first was a report, *A Nation at Risk*, which criticized public schools in the United States for failing to adequately prepare the country's future scientists and leaders (Pulliam & Van Patten, 1999).

Standardized tests were encouraged as a tool to measure educational progress.

Then the 1980s witnessed the formation of the first organized movement lambasting standardized testing. The mission of the Center for Fair and Open Testing has been and still is today to ensure tests are fair and valid (Curano, n.d.). Today, the Center for Fair and Open Testing remains the leading organization for making the public aware of any misuses or abuses of using testing scores for high-stakes decisions (Chandler, 1999). For example, FairTest criticizes any college which relies solely on SAT scores for admission decisions. During the 1990s, nothing really significant happened either positive or negative to shift ETS's momentum related to more and more testing.

The 21st century witnessed the birth of one, if not the most significant law related to education in the United States. The No Child Left Behind (NCLB) Act mandates that all schools show adequate yearly progress (AYP) toward a goal of 100 percent academic proficiency by 2014. All of this translates into more standardized testing. Many who oppose more and more testing in education would call this era the "teaching to the test era." Despite many objections and cautions related to the use of standardized testing throughout its history, beginning with those opposed to the eugenicists' movement early in the 20th century, the use of standardized tests for college admissions increased. ETS came to be the dominant force in the United States of America's educational system. Standardized testing became controversial with the eugenicist movement, and standardized testing will continue to be controversial if more is not done to educate all students equally.

The SAT is just one of the many standardized test given in the United States. Elite institutions of higher learning place a high emphasis on high SAT scores.

Students without cultural capital will be at a disadvantage as argued by Sacks (2007). What is cultural capital? Cultural capital is any additional resource afforded to those with higher education and money. As discussed by Sacks (drawing on Bourdieu), cultural capital is subtle. In discussing the subtle nature of cultural capital Sacks states, "Cultural capital is of no intrinsic value. Its utility comes in using, manipulating, and investing it for socially valued and difficult-to-secure purposes and resources" (p. 15). Affluent parents use their cultural capital to ensure that their children are well prepared to apply to the elite colleges such as UC Berkely, Stanford and Harvard. How is this cultural capital manifested into advantages for those with it? Taking advanced placement classes in high school and SAT test preparation are just two tools used by those with cultural capital to gain advantages. Given the competitive nature of attracting the best and brightest high school seniors, high SAT scores are considered a "jewel crown." Elite colleges are in competition for the illustrious rankings as published by U.S. News (Sacks, 2007). The single most important factor in getting a high SAT score probably would be associated with learning about the SAT, and how to take the SAT. Students, who come from families with cultural capital, in this case cultural capital as the specific knowledge about standardized tests, learn early on about the importance of getting a high SAT score. Furthermore, these culturally advantage students learn how to take the SAT (Sacks, 2007).

Today, standardized testing is a high-stakes measure with serious implications. The score a student receives determines which student advances to the next level in grade school; which student moves on to high school and which high school; and which student moves on to college and which college.

Students who come from a less advantaged economic background will lack the cultural capital defined by Sacks (2007). Many would argue that these students are already at a disadvantage related to taking standardize tests. A DIF analysis is just one tool that can be used to attempt to equal the playing field between economically advantaged and disadvantaged students. If fairness is one of the goals of standardized testing, then investigating and improving various statistical measures to assess DIF in test items should be highly encouraged.

.

References

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

Allen, N., & Donoghue, J. (1996). Applying the Mantel-Haenszel procedure to complex samples of items. *Journal of Educational Measurement, 33*(2), 231-251.

American Psychological Association. (2009). Publication manual of the American Psychological Association (6th ed.). Washington, DC.

Andrich, D. (1978). A rating formulation for ordered response categories. *Psychometrika, 43*, 561-573.

Angoff, W. (1993). Perspectives on differential item functioning methodology. In P. Holland & H. Wainer (Eds.), *Differential item functioning,* (pp. 3-24). Hillsdale, NJ: Erlbaum.

Awuor, R. (2008). *Effect of unequal sample sizes on the power of DIF detection: An IRT-Based Monte Carlo study with SIBTEST and Mantel-Haenszel procedures*. Doctoral dissertation, Virginia Polytechnic Institute and State University, 2008.

Bejar, I. (1980). Biased assessment of program impact due to psychometric artifacts. *Psychological Bulletin, 87*(3), 513-524.

Birnbaum, A. (1968). Some latent trait models and their uses in inferring an examinee's ability. In F. M. Lord & M. R. Novick, *Statistical theories of mental rest scores,* (Part 5, pp. 397-479). Reading, MA: Addison-Wesley.

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

Bolt, D. M. (2002). A Monte Carlo comparison of parametric and nonparametric polytomous DIF detection methods. *Applied Measurement in Education, 15*(2), 113-141.

Camilli, G., & Shepard, L. (1994). *Methods for identifying biased test items*. New Park, CA: Sage.

Chandler, M. (1999). *Frontline: Secrets of the SAT.* [Television broadcast]. Boston, MA: Public Broadcasting Service.

Choppin, B. (1983). *A two-parameter latent trait model* (CSE Report No. 197). Los Angeles, CA: University of California, Center for the Study of Evaluation, Graduate School of Eduation.

Clauser, B., Mazor, K., & Hambleton, R. (1994). The effects of score group width on the Mantel-Haenszel procedure. *Journal of Educational Measurement, 31*(1), 67-78.

Clauser, B., & Mazor, K. (1998). Using statistical procedures to identify differentially functioning test items. *Educational Measurement: Issues and Practice, 17*(1), 31-44.

Cohen, J. (1988). *Statistical power analysis for the behavioral sciences*. Hillsdale, NJ: Lawrence Erlbaum.

Cohen, J. (1990). Things I have learned (so far). *American Psychologist, 45*, 1304-1312.

Cohen, J. (1994). The earth is round (p<.05). *American Psychologist, 49*, 997-1003.

Caruano, R. M. (n.d.). An historical overview of standardized educational testing in the United States. Retrieved February 22, 2009, from http://www.gwu.edu/~gjackson/caruano.PDF

DeMars, C. (2009). Modification of the Mantel-Haenszel and Logistic Regression DIF procedures to incorporate the SIBTEST regression correction. *Journal of Educational and Behavioral Statistics, 34*(2), 149-170.

Donoghue, J. R., Holland, P. W., & Thayer, D. T. (1993). A Monte Carlo study of factors that affect the Mantel-Haenszel and Standardization measures of differential item functioning. In P. Holland & H. Wainer (Eds.), *Differential item functioning,* (pp. 137-166). Hillsdale, NJ: Erlbaum.

Dorans, N., & Holland, P. (1993). DIF Detection and description: Mantel_Haenszel and Standardization. In P. Holland & H. Wainer (Eds.), *Differential item functioning,* (pp. 35-66). Hillsdale, NJ: Erlbaum.

Dorans, N., & Kulick, E. (1983). *Assessing unexpected differential item performance of female candidates on SAT and TSWE forms administered in December 1977: An application of the standardization approach* (ETS Research Report RR-83-9). Princeton, NJ: Educational Testing Service.

Dorans, N., & Kulick, E. (1986). Demonstrating the utility of the standardization approach to assessing unexpected differential item performance on the Scholastic Aptitude Test. *Journal of Educational Measurement, 23*, 355-368.

Douglas, J., Roussos, L., & Stout, W. (1996). Item bundle DIF hypothesis testing: Identifying suspect bundles and assessing their differential item functioning. *Journal of Educational Measurement, 33*, 465-484.

Ducan, S. (2006). Improving the prediction of differential item functioning: A comparison of the use of an effect size for logistic regression DIF and Mantel-Haenszel DIF methods (Doctoral dissertation, Texas A&M University).

Finch, S., Cumming, G., & Thomason, N. (2001). Reporting statistical inference in the Journal of Applied Psychology: Little evidence of reform. *Educational and Psychological Measurement, 61*, 181-210.

Flowers, C. P., Oshima, T. C., & Raju, N. S. (1999). A description and demonstration of the polytomous-DFIT framework. *Applied Psychological Measurement, 23*, 309-326.

Gallagher, C. (2003). Reconciling a tradition of testing with a new learning paradigm. *Educational Psychology Review, 15*(1), 83-99.

Gierl, M., Gotzmann, A., & Boughton, K. (2004). Performance of SIBTEST when the percentage of DIF items is large. *Applied Measurement in Education, 17*(3), 241-264.

Gierl, M., Bisanz, J., Bisanz, G., & Boughton, K. (2001). Illustrating the utility of differential bundle functioning analyses to identify and interpret group differences on achievement tests. *Educational Measurement: Issues and Practice, 20*(2), 26-36.

Guler, N., & Penfield, R. (2009). A comparison of the logistic regression and contingency table methods for simultaneous detection of uniform and nonuniform DIF. *Journal of Educational Measurement, 46*(3), 314-329.

Hambleton, R. K., Swaminathan, H., & Rogers, H. J. (1991). *MMSS fundamentals of Item Response Theory*. Newbury Park, CA: Sage Publications.

Hays, W. L. (1981). *Statistics* (3rd ed.). New York, NY: Rinehart and Winston.

Hidalgo, M., & Lopez, A. (2004). Differential item functioning detection and effect size: A comparison between logistic regression and Mantel-Haenszel procedures. *Educational and Psychological Measurement, 64*(6), 903-915.

Holland, P. (1985). *On the study of differential item performance without IRT*. Proceedings of the Military Testing Association.

Holland, P.W., & Thayer, D. T. (1988). Differential item performance and the Mantel-Haenszel procedure. In H. Wainer & H. I. Braun (Eds.), *Test validity,* (pp. 129-145). Hillsdale, NJ: Erlbaum.

Hursh, D. (2008). *High-Stakes testing and the decline of teaching and learning*. Lanham, MD: Rowman & Littlefield Publishers.

Ironson, G. & Subkoviak, M. (1979). A comparison of several methods of assessing item bias. *Journal of Educational Measurement, 16*(4), 209-225.

Kirk, R. E. (1996). Practical significance: A concept whose time has come. *Educational and Psychological Measurement, 56*, 746-759.

Kirk, R. E. (2001). Promoting good statistical practices: Some suggestions. *Educational and Psychological Measurement, 61*(2), 213-218.

Kok, F. (1988). Item bias and test multidimensionality. In R. Langeheine, & J. Rost (Eds.), *Latent trait and latent class models,* (pp. 263-274). New York: Plenum Press.

Linn, R. L., Levine, M. V., Hastings, C. N., & Wardrop, J. L. (1981). Item bias in a test of reading comprehension. *Applied Psychological Measurement, 5*, 159-173.

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

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

Mantel, N., & Haenszel, W. (1959). Statistical aspects of the analysis of data from retrospective studies of disease. *Journal of the National Cancer Institute, 22*, 719-748.

Mazor, K., Kanjee, A., & Clauser, B. (1995). Using logistic regression and the Mantel-Haenszel with multiple ability estimates to detect differential item functioning. *Journal of Educational Measurement, 32*(2), 131-144.

Millsap, R., & Everson, H. (1993). Methodology review: Statistical approaches for assessing measurement bias. *Applied Psychological Measurement, 17*(4), 297-334.

Micceri, T. (1989). The unicorn, the normal curve, and other improbable creatures. *Psychological Bulletin, 105*(1), 156-166.

Monahan, P., McHorney, C., Stump, T., & Perkins, A. (2007). Odds ratio, delta, ETS classification, and standardization measures of DIF magnitude for binary logistic regression. *Journal of Educational and Behavioral Statistics, 32*(1), 92-109.

Muraki, E. (1972). A generalized partial credit model: Application of an EM algorithm. *Applied Psychological Measurement, 16*, 159-176.

Nandakumar, R. (1993). Simultaneous DIF amplification and cancellation: Shealy-Stout's test for DIF. *Journal of Educational Measurement, 30*, 293-311.

Oshima, T. C., & Miller, M. D. (1992). Multidimensionality and item bias in item response theory. *Applied Psychological Measurement, 16*, 237-248.

Oshima, T. C., & Morris, S. (2008). Raju's differential functioning of items and tests (DFIT). *Educational Measurement: Issues and Practice, 27*(3), 43-50.

Oshima, T. C., Raju, N. S., & Flowers, C. P. (1997). Development and demonstration of multidimensional IRT-Based internal measures of differential functioning of items and tests. *Journal of Educational Measurement, 34*(3), 253-272.

Oshima, T. C., Raju, N. S., Flowers, C. P., & Slinde, J. A. (1998). Differential bundle functioning using the DFIT framework: Procedures for identifying possible sources of differential functioning. *Applied Measurement in Education, 11*(4), 353-369.

Oshima, T. C., Raju, N. S., & Nanda, A. O. (2006). A new method for assessing the statistical significance in the differential functioning of items and tests (DFIT) framework. *Journal of Educational Measurement, 43*(1), 1-17.

Osterlind, S., & Everson, H. (2009). *Differential Item Functioning*. New Park, CA: Sage.

Penny, J., & Johnson, R. (1999). How group differences in matching criterion distribution and IRT item difficulty can influence the magnitude of the Mantel-Haenszel Chi-Square DIF index. *Journal of Experimental Education, 67*(4), 343-367.

Penfield, R., & Lam, T. (2000). Assessing differential item functioning in performance assessment: Review and recommendations. *Educational Measurement: Issues & Practices, 19*(3), 5-15.

Pommerich, M., Spray, J. A., & Parshal, C. G. (1994). *The performance of the Mantel-Haenszel DIF statistic when comparison group distributions are incongruent*. Paper presented at the Annual Meeting of the National Council on Measurement in Education, New Orleans. (ERIC Document Reproduction Service No. ED37097)

Potenza, M., & Dorans, N. (1995). DIF assessment for polytomously scored items: A framework for classification and evaluation. *Applied Psychological Measurement, 19*(1), 23-37.

Pulliam, J., & Van Patten, J. (1999). *History of education in America*. Upper Saddle River, NJ: Prentice Hall, Inc.

Raju, N. S. (1988). The area between two item characteristic curves. *Psychometrika, 54*, 495-502.

Raju, N. S., van Der Linden, W. J., & Fleer, P. F. (1995). An IRT-based internal measure of test bias with applications for differential item functioning. *Applied Psychological Measurement, 19*, 353-368.

Ramsey, P. (1993). Sensitivity review: The ETS experience as a case study. In P. Holland & H. Wainer (Eds.), *Differential item functioning,* (pp. 367-388). Hillsdale, NJ: Erlbaum.

Rasch, G. (1960). *Probabilistic models for some intelligence and attainment tests*. Copenhagen: Denmarks Paedagogiske Institute.

Reckase, M. D. (1978). *A comparison of the one- and three-parameter logistic models for item calibration.* Paper presented at the annual meeting of the American Educational Research Association, Toronto, Canada. (ERIC Document Reproduction Service No. ED 155203)

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

Ross, T. (2007). *The impact of multidimensionality on the detection of differential bundle functioning using SIBTEST*. Doctoral dissertation, Georgia State University, 2007.

Roussos, L. A., Schnipke, D., & Pashley, P. (1999). A generalized formula for the Mantel-Haenszel differential item functioning parameter. *Journal of Educational and Behavioral Statistics, 24*(3), 293-322.

Roussos, L. A., & Stout, W. F. (1996a). A multidimensionality-based DIF analysis paradigm. *Applied Psychological Measurement, 20*, 355-371.

Roussos, L. A., & Stout, W. F. (1996b). Simulation studies of the effect of small sample size and studied item parameters on SIBTEST and Mantel-Haenszel Type I error performance. *Journal of Educational Measurement, 33*(2), 215-230.

Rudner, L., & Convey, J. (1978). *An evaluation of select approaches for biased item identification.* Paper presented at the annual meeting of the American Educational Research Association, Toronto, Canada. (ERIC Document Reproduction Service No. ED 157942)

Rudner, L., Getson, P., & Knight, D. (1980). Biased item detection techniques. *Journal of Educational Statistics, 5*(2), 213-233.

Sacks, P. (2007). *Tearing down the gates: Confronting the class divide in American education*. Los Angeles, CA: University of California Press.

Samejima, F. (1969). Estimation of latent ability using a response pattern of graded scores. *Psychometrika Monograph, No 17.*

Shealy, R., & Stout, W. F. (1993). A model-based standardization approach that separates true bias/DIF from group differences and detects test bias/DIF as well as item bias/DIF. *Psychometrika, 58*, 159-194.

Shepard, L., Camilli, G., & Averill, M. (1981). Comparison of procedures for detecting test-item bias with both internal and external ability criteria. *Journal of Educational Statistics, 6*(4), 317-375.

Sinharay, S., & Dorans, N. (2010). Two simple approaches to overcome a problem with the Mantel-Haenszel statistic. *Journal of Educational and Behavioral Statistics, 35*(4), 474-488.

Snow, T., & Oshima, T. C. (2009). A comparison of unidimensional and three-dimensional differential item functioning analysis using two-dimensional data. *Educational and Psychological Measurement, 69*, 732-747.

Spray, J., Miller, T. (1992). *Performance of the Mantel-Haenszel statistic and the Standardized Difference in proportions correct when population ability distributions are incongruent* (ACT Research Report Series No. 92-1). ACT.

Swaminathan, H., & Rogers, J. (1990). Detecting differential item functioning using logistic regression procedures. *Journal of Educational Measurement, 27*(4), 361-370.

Sweeney, K. P. (1996). *A Monte Carlo investigation of the likelihood-ratio procedure in the detection of differential item functioning*. Unpublished doctoral dissertation, Fordham University, New York, NY.

Teresi, J., & Fleishman, J. (2007). Differential item functioning and health assessment. *Quality of Life Research, 16*, 33-42.

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. I. Braun (Eds.), *Test validity,* (pp. 147-169). Hillsdale, NJ: Erlbaum.

Thompson, B. (1999). Statistical significance tests, effect size reporting, and the vain pursuit of pseudo-objectivity. *Theory & Psychology, 9*(2), 191-196.

Thompson, B. (2002). "Statistical," "practical," and "clinical": How many kinds of significance do counselors need to consider? *Journal of Counseling and Development, 80*, 64-71.

Whittaker, T., Fitzpatrick, S. J., Williams, N. J., & Dodd, B. G. (2003). IRTGEN: A SAS macro program to generate known trait scores and item responses for commonly used Item Response Theory models. *Applied Psychological Measurement, 27*(4), 299-300.

Wiberg, M. (2007). *Measuring and detecting differential item functioning in criterion-referenced licensing test: A theoretic comparison of methods* (Department of Educational Measurement Research Report EM No. 60). Sweden: UMEA University.

Zieky, M. (1993). Practical questions in the use of DIF statistics in test development. In P. Holland & H. Wainer (Eds.), *Differential item functioning,* (pp. 337-347). Hillsdale, NJ: Erlbaum.

Zwick, R. (1990). When do item response function and Mantel-Haenszel definitions of differential item functioning coincide? *Journal of Educational Statistics, 15*(3), 185-197.

Zwick, R., & Ercikan, K. (1989). Analysis of differential item functioning in the NAEP history assessment. *Journal of Educational Measurement, 26*, 55-66.

Zwick, R., Thayer, D. T., & Wingersky, M. (1994). A simulation study of methods for assessing differential item functioning in computerized adaptive tests. *Applied Psychological Measurement, 18*, 121-140.

APPENDIXES

APPENDIX A

Item Parameters - Reference and Focal Groups

Item	*a*	*b*	*c*
1	0.94	-1.76	0.20
2	1.99	-0.21	0.20
3	1.24	-1.21	0.20
4	1.47	-1.40	0.20
5	2.22	-0.78	0.20
6	1.21	-1.56	0.20
7	1.14	-1.10	0.20
8	1.51	-0.92	0.20
9	1.56	-1.14	0.20
10	2.28	-0.23	0.20
11	2.16	-0.91	0.20
12	1.60	-0.52	0.20
13	1.89	0.26	0.20
14	2.09	0.03	0.20
15	2.26	0.04	0.20
16	1.40	-0.25	0.20
17	2.50	-0.21	0.20
18	1.76	-0.26	0.20
19	1.78	-0.54	0.20
20	2.42	-0.15	0.20
21	1.12	-1.08	0.20

22	0.60	0.84	0.20
23	2.17	-0.44	0.20
24	1.55	0.30	0.20
25	1.32	-0.63	0.20
26	2.32	0.31	0.20
27	2.11	-0.18	0.20
28	1.28	-0.02	0.20
29	2.04	0.14	0.20
30	2.92	0.08	0.20
31	1.76	0.47	0.20
32	1.86	0.30	0.20
33	1.20	0.37	0.20
34	1.76	-0.11	0.20
35	2.09	0.34	0.20
36	1.41	-0.04	0.20
37	1.71	0.11	0.20
38	1.50	0.70	0.20
39	1.49	-0.18	0.20
40	1.76	-1.01	0.20
41	1.13	2.24	0.20
42	2.59	0.30	0.20
43	1.70	0.87	0.20
44	2.67	0.26	0.20
45	0.61	0.36	0.20
46	1.29	0.07	0.20

47	2.03	0.88	0.20
48	2.50	0.82	0.20
49	2.02	0.80	0.20
50	2.04	0.48	0.20
51	1.91	1.57	0.20
52	1.80	1.39	0.20
53	2.03	1.03	0.20
54	2.44	1.42	0.20
55	1.16	1.58	0.20
56	3.07	1.43	0.20
57	1.80	1.33	0.20
58	2.25	1.05	0.20
59	2.71	1.53	0.20
60	2.47	2.26	0.20
61	Variable	Variable	0 or .20

APPENDIX B

Conditions – Test Item 61

Condition	*a*	*b*	*c*	Amount of DIF	# of Increments
					Within Conditions
1	N/A	-3	N/A	0.1	40
2	N/A	-2	N/A	0.1	40
3	N/A	-1.5	N/A	0.05	60
4	N/A	-1	N/A	0.05	60
5	N/A	-0.5	N/A	0.05	40
6	N/A	0	N/A	0.05	40
7	N/A	0.5	N/A	0.05	20
8	N/A	1	N/A	0.05	20
9	N/A	1.5	N/A	0.025	10
10	N/A	2	N/A	0.025	10
11	N/A	3	N/A	0.025	10
12	0.3	-3	N/A	0.1	40
13	0.3	-2	N/A	0.1	40
14	0.3	-1.5	N/A	0.05	60
15	0.3	-1	N/A	0.05	60
16	0.3	-0.5	N/A	0.05	40
17	0.3	0	N/A	0.05	40
18	0.3	0.5	N/A	0.05	20
19	0.3	1	N/A	0.05	20
20	0.3	1.5	N/A	0.1	10
21	0.3	2	N/A	0.1	10
22	0.3	3	N/A	0.1	10
23	0.5	-3	N/A	0.1	40
24	0.5	-2	N/A	0.1	40
25	0.5	-1.5	N/A	0.05	60
26	0.5	-1	N/A	0.05	60
27	0.5	-0.5	N/A	0.05	40
28	0.5	0	N/A	0.05	40
29	0.5	0.5	N/A	0.05	20
30	0.5	1	N/A	0.05	20

31	0.5	1.5	N/A	0.1	10
32	0.5	2	N/A	0.1	10
33	0.5	3	N/A	0.1	10
34	0.75	-3	N/A	0.1	40
35	0.75	-2	N/A	0.1	40
36	0.75	-1.5	N/A	0.05	60
37	0.75	-1	N/A	0.05	60
38	0.75	-0.5	N/A	0.05	40
39	0.75	0	N/A	0.05	40
40	0.75	-0.5	N/A	0.05	20
41	0.75	1	N/A	0.05	20
42	0.75	1.5	N/A	0.1	10
43	0.75	2	N/A	0.1	10
44	0.75	3	N/A	0.1	10
45	0.95	-3	N/A	0.1	40
46	0.95	-2	N/A	0.1	40
47	0.95	-1.5	N/A	0.05	60
48	0.95	-1	N/A	0.05	60
49	0.95	-0.5	N/A	0.05	40
50	0.95	0	N/A	0.05	40
51	0.95	-0.5	N/A	0.05	20
52	0.95	1	N/A	0.05	20
53	0.95	1.5	N/A	0.1	10
54	0.95	2	N/A	0.1	10
55	0.95	3	N/A	0.1	10
56	1.25	-3	N/A	0.1	40
57	1.25	-2	N/A	0.1	40
58	1.25	-1.5	N/A	0.05	60
59	1.25	-1	N/A	0.05	60
60	1.25	-0.5	N/A	0.05	40
61	1.25	0	N/A	0.05	40
62	1.25	0.5	N/A	0.05	20
63	1.25	1	N/A	0.05	20
64	1.25	1.5	N/A	0.025	10
65	1.25	2	N/A	0.025	10
66	1.25	3	N/A	0.025	10

67	1.5	-3	N/A	0.1	40
68	1.5	-2	N/A	0.1	40
69	1.5	-1.5	N/A	0.05	60
70	1.5	-1	N/A	0.05	60
71	1.5	-0.5	N/A	0.05	40
72	1.5	0	N/A	0.05	40
73	1.5	0.5	N/A	0.05	20
74	1.5	1	N/A	0.05	20
75	1.5	1.5	N/A	0.025	10
76	1.5	2	N/A	0.025	10
77	1.5	3	N/A	0.025	10
78	1.75	-3	N/A	0.1	40
79	1.75	-2	N/A	0.1	40
80	1.75	-1.5	N/A	0.05	60
81	1.75	-1	N/A	0.05	60
82	1.75	-0.5	N/A	0.05	40
83	1.75	0	N/A	0.05	40
84	1.75	0.5	N/A	0.025	10
85	1.75	1	N/A	0.025	10
86	1.75	1.5	N/A	0.025	10
87	1.75	2	N/A	0.025	10
88	1.75	3	N/A	0.025	10
89	2	-3	N/A	0.1	40
90	2	-2	N/A	0.1	40
91	2	-1.5	N/A	0.05	60
92	2	-1	N/A	0.05	60
93	2	-0.5	N/A	0.05	40
94	2	0	N/A	0.05	40
95	2	0.5	N/A	0.025	10
96	2	1	N/A	0.025	10
97	2	1.5	N/A	0.025	10
98	2	2	N/A	0.025	10
99	2	3	N/A	0.025	10
100	0.3	-3	0.2	0.1	40
101	0.3	-2	0.2	0.1	40
102	0.3	-1.5	0.2	0.05	60

103	0.3	-1	0.2	0.05	60
104	0.3	-0.5	0.2	0.05	40
105	0.3	0	0.2	0.05	40
106	0.3	0.5	0.2	0.2	10
107	0.3	1	0.2	0.2	10
108	0.3	1.5	0.2	0.2	10
109	0.3	2	0.2	0.2	10
110	0.3	3	0.2	0.2	10
111	0.5	-3	0.2	0.1	40
112	0.5	-2	0.2	0.1	40
113	0.5	-1.5	0.2	0.05	60
114	0.5	-1	0.2	0.05	60
115	0.5	-0.5	0.2	0.05	40
116	0.5	0	0.2	0.05	40
117	0.5	0.5	0.2	0.2	10
118	0.5	1	0.2	0.2	10
119	0.5	1.5	0.2	0.2	10
120	0.5	2	0.2	0.2	10
121	0.5	3	0.2	0.2	10
122	0.75	-3	0.2	0.1	40
123	0.75	-2	0.2	0.1	40
124	0.75	-1.5	0.2	0.05	60
125	0.75	-1	0.2	0.05	60
126	0.75	-0.5	0.2	0.05	40
127	0.75	0	0.2	0.05	40
128	0.75	0.5	0.2	0.2	10
129	0.75	1	0.2	0.2	10
130	0.75	1.5	0.2	0.2	10
131	0.75	2	0.2	0.2	10
132	0.75	3	0.2	0.2	10
133	0.95	-3	0.2	0.1	40
134	0.95	-2	0.2	0.1	40
135	0.95	-1.5	0.2	0.05	60
136	0.95	-1	0.2	0.05	60
137	0.95	-0.5	0.2	0.05	40
138	0.95	0	0.2	0.05	40

139	0.95	0.5	0.2	0.2	10
140	0.95	1	0.2	0.2	10
141	0.95	1.5	0.2	0.2	10
142	0.95	2	0.2	0.2	10
143	0.95	3	0.2	0.2	10
144	1.25	-3	0.2	0.1	40
145	1.25	-2	0.2	0.1	40
146	1.25	-1.5	0.2	0.05	60
147	1.25	-1	0.2	0.05	60
148	1.25	-0.5	0.2	0.05	40
149	1.25	0	0.2	0.05	40
150	1.25	0.5	0.2	0.2	10
151	1.25	1	0.2	0.2	10
152	1.25	1.5	0.2	0.2	10
153	1.25	2	0.2	0.2	10
154	1.25	3	0.2	0.2	10
155	1.5	-3	0.2	0.1	40
156	1.5	-2	0.2	0.1	40
157	1.5	-1.5	0.2	0.05	60
158	1.5	-1	0.2	0.05	60
159	1.5	-0.5	0.2	0.05	40
160	1.5	0	0.2	0.05	40
161	1.5	0.5	0.2	0.2	10
162	1.5	1	0.2	0.2	10
163	1.5	1.5	0.2	0.2	10
164	1.5	2	0.2	0.2	10
165	1.5	3	0.2	0.2	10
166	1.75	-3	0.2	0.1	40
167	1.75	-2	0.2	0.1	40
168	1.75	-1.5	0.2	0.05	60
169	1.75	-1	0.2	0.05	60
170	1.75	-0.5	0.2	0.05	40
171	1.75	0	0.2	0.05	40
172	1.75	0.5	0.2	0.2	10
173	1.75	1	0.2	0.2	10
174	1.75	1.5	0.2	0.2	10

175	1.75	2	0.2	0.2	10
176	1.75	3	0.2	0.2	10
177	2	-3	0.2	0.1	40
178	2	-2	0.2	0.1	40
179	2	-1.5	0.2	0.05	60
180	2	-1	0.2	0.05	60
181	2	-0.5	0.2	0.05	40
182	2	0	0.2	0.05	40
183	2	0.5	0.2	0.2	10
184	2	1	0.2	0.2	10
185	2	1.5	0.2	0.2	10
186	2	2	0.2	0.2	10
187	2	3	0.2	0.2	10

TOTAL 5750

APPENDIX C

Comprehensive Results – Impact Case

Notes:

(1) Only 116 of the 187 conditions were also estimated for MH. If N/A is in the "CONGRUENT" column, these conditions were not selected to be estimated. N/A does not indicate estimation issues with these conditions.

(2) There were 46 of the 187 conditions that could not be accurately estimated for MH, these conditions are noted by the "Indeterminate" label in the "CONGRUENT" column.

(3) There were 22 of the 46 conditions were the MH parameter for moderate DIF (Category B) could not be determined. "Indeterminate" is indicated in the "MH" column.

COND.	AREA	b-diff	EST	EST	ADJ		CONGRUENT
			MH	NO DIF	EST MH	MH	
1	0.3	0.3	-1.676	-0.641	-1.035	-1.198	√
2	0.3	0.3	-1.505	-0.281	-1.224	-1.198	√
3	0.25	0.25	-1.171	-0.194	-0.977	-0.999	√
4	0.25	0.25	-1.172	-0.157	-1.015	-0.999	√
5	0.3	0.3	-1.346	-0.177	-1.169	-1.198	√
6	0.3	0.3	-1.41	-0.202	-1.208	-1.198	√
7	0.25	0.25	-1.245	-0.17	-1.075	-0.999	√
8	0.3	0.3	-1.519	-0.349	-1.17	-1.199	√
9	0.25	0.25	N/A	N/A	N/A	-0.999	N/A
10	0.25	0.25	N/A	N/A	N/A	-0.999	N/A
11	0.25	0.25	N/A	N/A	N/A	-0.999	Indeterminate
12	0.9	0.9	-1.097	0.032	-1.129	-1.073	√
13	0.8	0.8	-0.977	0.034	-1.011	-0.956	√
14	0.8	0.8	-0.93	0.068	-0.998	-0.958	√
15	0.85	0.85	-0.989	0.063	-1.052	-1.02	√
16	1.8	1.8	-1.036	0.01	-1.046	-2.164	x
17	1.7	1.7	-1.024	0.004	-1.028	-2.046	x
18	1.7	1.7	-0.978	0.013	-0.991	-2.048	x
19	1.7	1.7	-1.048	0.038	-1.086	-2.049	x
20	0.9	0.9	N/A	N/A	N/A	-1.079	N/A
21	0.9	0.9	N/A	N/A	N/A	-1.079	N/A

22	0.9	0.9	N/A	N/A	N/A	-1.079	N/A
23	0.5	0.5	-1.052	-0.081	-0.971	-0.992	√
24	0.5	0.5	-1.066	-0.004	-1.062	-0.996	√
25	0.5	0.5	-1.035	-0.012	-1.023	-0.997	√
26	0.5	0.5	-1.024	-0.025	-0.999	-0.999	√
27	1	1	-1.055	0.016	-1.071	-2.001	x
28	1.1	1.1	-1.161	-0.069	-1.092	-2.203	x
29	1.1	1.1	-1.157	-0.095	-1.062	-2.204	x
30	1	1	-1.123	-0.082	-1.041	-2.006	x
31	0.9	0.9	N/A	N/A	N/A	-1.079	N/A
32	0.9	0.9	N/A	N/A	N/A	-1.079	N/A
33	0.9	0.9	N/A	N/A	N/A	-1.079	N/A
34	0.4	0.4	-1.43	-0.288	-1.142	-1.144	√
35	0.4	0.4	-1.342	-0.064	-1.278	-1.172	√
36	0.35	0.35	-1.1	-0.091	-1.009	-1.034	√
37	0.35	0.35	-1.087	-0.094	-0.993	-1.047	√
38	0.35	0.35	-1.16	-0.101	-1.059	-1.06	√
39	0.35	0.35	-1.191	-0.11	-1.081	-1.073	√
40	0.35	0.35	-1.194	-0.172	-1.022	-1.086	√
41	0.35	0.35	-1.272	-0.188	-1.084	-1.1	√
42	0.35	0.35	N/A	N/A	N/A	-1.049	N/A
43	0.35	0.35	N/A	N/A	N/A	-1.049	N/A
44	0.35	0.35	N/A	N/A	N/A	-1.049	Indeterminate
45	0.3	0.3	-1.566	0.538	-1.028	-1.027	√
46	0.3	0.3	-1.349	-0.277	-1.072	-1.114	√
47	0.3	0.3	-1.243	-0.163	-1.08	-1.126	√
48	0.3	0.3	-1.279	-0.142	-1.137	-1.137	√
49	0.3	0.3	-1.268	-0.132	-1.136	-1.148	√
50	0.3	0.3	-1.342	-0.114	-1.228	-1.159	√
51	0.3	0.3	-1.334	-0.301	-1.033	-1.17	√
52	0.3	0.3	-1.494	-0.403	-1.091	-1.181	√
53	0.3	0.3	N/A	N/A	N/A	-1.139	N/A
54	0.3	0.3	N/A	N/A	N/A	-1.139	N/A
55	0.3	0.3	N/A	N/A	N/A	-1.139	Indeterminate

56	0.3	0.3	-2.344	-0.944	-1.4	-1.421	√
57	0.3	0.3	-1.777	-0.336	-1.441	-1.463	√
58	0.2	0.2	-1.263	-0.239	-1.024	-0.976	√
59	0.2	0.2	-1.16	-0.208	-0.952	-0.993	√
60	0.2	0.2	-1.196	-0.192	-1.004	-1.01	√
61	0.2	0.2	-1.178	-0.219	-0.959	-1.027	√
62	0.2	0.2	-1.368	-0.319	-1.049	-1.044	√
63	0.25	0.25	-1.754	-0.36	-1.394	-1.316	√
64	0.2	0.2	N/A	N/A	N/A	-0.999	N/A
65	0.2	0.2	N/A	N/A	N/A	-0.999	Indeterminate
66	0.2	0.2	N/A	N/A	N/A	-0.999	Indeterminate
67	0.3	0.3	-2.756	-1.296	-1.46	-1.406	√
68	0.2	0.2	-1.708	-0.611	-1.097	-1.151	√
69	0.2	0.2	-1.373	-0.269	-1.104	-1.173	√
70	0.2	0.2	-1.366	-0.226	-1.14	-1.194	√
71	0.2	0.2	-1.343	-0.228	-1.115	-1.214	√
72	0.2	0.2	-1.427	-0.26	-1.167	-1.235	√
73	0.2	0.2	-1.637	-0.409	-1.228	-1.256	√
74	0.2	0.2	-1.531	-0.33	-1.201	-1.28	√
75	0.17	0.17	N/A	N/A	N/A	-1.049	N/A
76	0.17	0.17	N/A	N/A	N/A	-1.049	Indeterminate
77	0.17	0.17	N/A	N/A	N/A	-1.049	Indeterminate
78	0.2	0.2	-2.736	-1.574	-1.162	-1.317	√
79	0.2	0.2	-1.893	-0.707	-1.186	-1.361	√
80	0.15	0.15	-1.399	-0.409	-0.99	-1.026	√
81	0.15	0.15	-1.278	-0.286	-0.992	-1.043	√
82	0.15	0.15	-1.212	-0.197	-1.015	-1.06	√
83	0.15	0.15	-1.692	-0.33	-1.362	-1.431	√
84	0.15	0.15	N/A	N/A	N/A	-1.049	N/A
85	0.15	0.15	N/A	N/A	N/A	-1.049	N/A
86	0.15	0.15	N/A	N/A	N/A	-1.049	Indeterminate
87	0.15	0.15	N/A	N/A	N/A	-1.049	Indeterminate
88	0.15	0.15	N/A	N/A	N/A	-1.049	Indeterminate
89	0.2	0.2	-3.085	-2.025	-1.06	-1.49	x

90	0.2	0.2	-2.203	-0.864	-1.339	-1.549	x
91	0.15	0.15	-1.57	-0.498	-1.072	-1.169	√
92	0.15	0.15	-1.43	-0.342	-1.088	-1.193	√
93	0.15	0.15	-1.397	-0.293	-1.104	-1.217	√
94	0.15	0.15	-1.525	-0.341	-1.184	-1.24	√
95	0.13	0.13	N/A	N/A	N/A	-0.999	N/A
96	0.13	0.13	N/A	N/A	N/A	-0.999	N/A
97	0.13	0.13	N/A	N/A	N/A	-0.999	Indeterminate
98	0.13	0.13	N/A	N/A	N/A	-0.999	Indeterminate
99	0.13	0.13	N/A	N/A	N/A	-0.999	Indeterminate
100	0.72	0.9	-1.018	-0.015	-1.003	-0.988	√
101	0.8	1	-1.045	0.011	-1.056	-1.05	√
102	0.8	1	-1.016	0.015	-1.031	-1.019	√
103	0.84	1.05	-1.015	-0.015	-1	-1.029	√
104	1.68	2.1	-0.997	0.023	-1.02	-1.845	x
105	1.92	2.4	-1.015	0.044	-1.059	-1.925	x
106	1.12	1.4	N/A	N/A	N/A	-1.126	Indeterminate
107	1.12	1.4	N/A	N/A	N/A	-1.036	Indeterminate
108	1.28	1.6	N/A	N/A	N/A	-1.052	Indeterminate
109	1.6	1.8	N/A	N/A	N/A	-1.03	Indeterminate
110	*	*	N/A	N/A	N/A	Indeterminate	Indeterminate
111	0.48	0.6	-1.302	-0.238	-1.064	-1.128	√
112	0.48	0.6	-1.212	-0.118	-1.094	-1.074	√
113	0.52	0.65	-1.162	-0.103	-1.059	-1.116	√
114	0.48	0.6	-1.052	-0.077	-0.975	-0.981	√
115	1.12	1.4	-1.09	-0.033	-1.057	-1.972	x
116	1.12	1.4	-0.962	0.004	-0.966	-1.762	x
117	0.32	0.9	N/A	N/A	N/A	-1.063	N/A
118	0.96	1.2	N/A	N/A	N/A	-1.141	Indeterminate
119	1.12	1.4	N/A	N/A	N/A	-1.034	Indeterminate
120	*	*	N/A	N/A	N/A	Indeterminate	Indeterminate
121	*	*	N/A	N/A	N/A	Indeterminate	Indeterminate
122	0.32	0.4	-1.6	-0.523	-1.077	-1.095	√
123	0.32	0.4	-1.425	-0.294	-1.131	-1.06	√

124	0.32	0.4	-1.241	-0.18	-1.061	-1.021	√
125	0.32	0.4	-1.115	-0.139	-0.976	-0.964	√
126	0.4	0.5	-1.183	-0.084	-1.099	-1.093	√
127	0.4	0.5	-0.998	-0.01	-0.988	-0.96	√
128	0.56	0.7	N/A	N/A	N/A	-1.043	N/A
129	0.8	1	N/A	N/A	N/A	-1.038	Indeterminate
130	0.8	1	N/A	N/A	N/A	-1.038	Indeterminate
131	*	*	N/A	N/A	N/A	Indeterminate	Indeterminate
132	*	*	N/A	N/A	N/A	Indeterminate	Indeterminate
133	0.24	0.3	-1.926	-0.851	-1.075	-1.044	√
134	0.32	0.4	-1.672	-0.303	-1.369	-1.342	√
135	0.28	0.35	-1.407	-0.37	-1.037	-1.2	√
136	0.28	0.35	-1.291	-0.178	-1.113	-1.044	√
137	0.32	0.4	-1.21	-0.136	-1.074	-1.064	√
138	0.4	0.5	-1.154	-0.104	-1.05	-1.106	√
139	0.48	0.6	N/A	N/A	N/A	-0.995	N/A
140	0.8	1	N/A	N/A	N/A	-0.996	Indeterminate
141	*	*	N/A	N/A	N/A	Indeterminate	Indeterminate
142	*	*	N/A	N/A	N/A	Indeterminate	Indeterminate
143	*	*	N/A	N/A	N/A	Indeterminate	Indeterminate
144	0.24	0.3	-2.253	-0.855	-1.398	-1.352	√
145	0.24	0.3	-1.555	-0.503	-1.052	-1.292	√
146	0.24	0.3	-1.623	-0.246	-1.377	-1.226	√
147	0.24	0.3	-1.396	-0.284	-1.112	-1.129	√
148	0.28	0.35	-1.29	-0.124	-1.166	-1.148	√
149	0.32	0.4	-1.099	-0.06	-1.039	-1.057	√
150	0.48	0.6	N/A	N/A	N/A	-1.056	N/A
151	1.12	1.2	N/A	N/A	N/A	-1.024	Indeterminate
152	*	*	N/A	N/A	N/A	Indeterminate	Indeterminate
153	*	*	N/A	N/A	N/A	Indeterminate	Indeterminate
154	*	*	N/A	N/A	N/A	Indeterminate	Indeterminate
155	0.16	0.2	-2.536	-1.522	-1.014	-1.036	√
156	0.24	0.3	-2.167	-0.634	-1.533	-1.519	√
157	0.2	0.25	-1.461	-0.33	-1.131	-1.193	√

158	0.2	0.25	-1.409	-0.363	-1.046	-1.09	√
159	0.24	0.3	-1.304	-0.127	-1.177	-1.123	√
160	0.28	0.35	-1.091	-0.071	-1.02	-1.027	√
161	0.48	0.6	N/A	N/A	N/A	-1.077	N/A
162	1.6	2	N/A	N/A	N/A	-0.971	Indeterminate
163	*	*	N/A	N/A	N/A	Indeterminate	Indeterminate
164	*	*	N/A	N/A	N/A	Indeterminate	Indeterminate
165	*	*	N/A	N/A	N/A	Indeterminate	Indeterminate
166	0.24	0.3	-3.722	-2.181	-1.541	-1.878	√
167	0.16	0.2	-2.161	-1.13	-1.031	-1.157	√
168	0.2	0.25	-1.932	-0.646	-1.286	-1.356	√
169	0.16	0.2	-1.36	-0.363	-0.997	-0.983	√
170	0.2	0.25	-1.198	-0.2	-0.998	-1.039	√
171	0.28	0.35	-1.165	-0.099	-1.066	-1.09	√
172	0.48	0.6	N/A	N/A	N/A	-1.086	N/A
173	*	*	N/A	N/A	N/A	Indeterminate	Indeterminate
174	*	*	N/A	N/A	N/A	Indeterminate	Indeterminate
175	*	*	N/A	N/A	N/A	Indeterminate	Indeterminate
176	*	*	N/A	N/A	N/A	Indeterminate	Indeterminate
177	0.16	0.2	-3.66	-2.345	-1.315	-1.365	√
178	0.16	0.2	-2.396	-1.288	-1.108	-1.284	√
179	0.16	0.2	-1.838	-0.791	-1.047	-1.2	√
180	0.16	0.2	-1.465	-0.425	-1.04	-1.078	√
181	0.2	0.25	-1.314	-0.176	-1.138	-1.124	√
182	0.28	0.35	-1.196	-0.023	-1.173	-1.148	√
183	0.48	0.6	N/A	N/A	N/A	-1.088	N/A
184	*	*	N/A	N/A	N/A	Indeterminate	Indeterminate
185	*	*	N/A	N/A	N/A	Indeterminate	Indeterminate
186	*	*	N/A	N/A	N/A	Indeterminate	Indeterminate
187	*	*	N/A	N/A	N/A	Indeterminate	Indeterminate

APPENDIX D

Correlation Matrix – Impact Case (The number represents a specific condition). For those conditions in which MH was not estimated, correlations are not provided.

	1			2			3			4		
MH	1			1			1			1		
NCDIF	**0.99**	1		**0.99**	1		**1**	1		**0.99**	1	
SIBTEST	0.98	0.99	1	0.97	0.99	1	0.98	0.98	1	0.95	0.96	1

	5			6			7			8		
MH	1			1			1			1		
NCDIF	**0.99**	1		**0.99**	1		**0.99**	1		**0.99**	1	
SIBTEST	0.97	0.97	1	0.94	0.96	1	0.97	0.97	1	0.94	0.95	1

	12			13			14			15		
MH	1			1			1			1		
NCDIF	**0.99**	1		**1**	1		**1**	1		**1**	1	
SIBTEST	1	0.99	1	1	0.99	1	0.99	0.99	1	0.99	1	1

	16			17			18			19		
MH	1			1			1			1		
NCDIF	**0.99**	1		**0.99**	1		**0.99**	1		**0.99**	1	
SIBTEST	0.98	0.98	1	0.98	0.98	1	0.95	0.96	1	0.95	0.96	1

	23			24			25			26		
MH	1			1			1			1		
NCDIF	**0.99**	1		**0.99**	1		**1**	1		**0.99**	1	
SIBTEST	1	0.99	1	0.99	0.99	1	1	0.99	1	0.99	0.99	1

	27			28			29			30		
MH	1			1			1			1		
NCDIF	**0.98**	1		**0.98**	1		**0.99**	1		**0.99**	1	
SIBTEST	0.99	0.99	1	0.98	0.99	1	0.97	0.97	1	0.97	0.96	1

	34			35			36			37		
MH	1			1			1			1		
NCDIF	**0.99**	1		**0.99**	1		**0.99**	1		**0.99**	1	
SIBTEST	0.99	0.99	1	0.99	0.98	1	0.99	0.98	1	0.97	0.98	1

	38			39			40			41		
MH	1			1			1			1		
NCDIF	**0.99**	1		**0.99**	1		**0.99**	1		**0.99**	1	
SIBTEST	0.98	0.99	1	0.97	0.98	1	0.97	0.97	1	0.97	0.97	1

	45			46			47			48		
MH	1			1			1			1		
NCDIF	**0.98**	1		**0.99**	1		**0.99**	1		**0.99**	1	
SIBTEST	0.98	0.99	1	0.98	0.97	1	0.98	0.98	1	0.95	0.96	1

	49			50			51			52		
MH	1			1			1			1		
NCDIF	**0.99**	1		**0.99**	1		**0.99**	1		**0.99**	1	
SIBTEST	0.97	0.98	1	0.94	0.97	1	0.98	0.97	1	0.92	0.94	1

	56			57			58			59		
MH	1			1			1			1		
NCDIF	**0.97**	1		**0.99**	1		**0.99**	1		**0.99**	1	
SIBTEST	0.97	0.99	1	0.96	0.96	1	0.97	0.96	1	0.92	0.94	1

	60			61			62			63		
MH	1			1			1			1		
NCDIF	**0.99**	1		**0.99**	1		**0.99**	1		**0.99**	1	
SIBTEST	0.94	0.95	1	0.91	0.95	1	0.94	0.95	1	0.89	0.92	1

	67			68			69			70		
MH	1			1			1			1		
NCDIF	**0.97**	1		**0.99**	1		**0.99**	1		**0.99**	1	
SIBTEST	0.97	0.98	1	0.95	0.95	1	0.96	0.95	1	0.9	0.92	1

	71			72			73			74		
MH	1			1			1			1		
NCDIF	**0.99**	1		**0.99**	1		**0.99**	1		**0.99**	1	
SIBTEST	0.92	0.93	1	0.86	0.91	1	0.95	0.95	1	0.83	0.83	1

	78			79			80			81		
MH	1			1			1			1		
NCDIF	**0.95**	1		**0.99**	1		**0.99**	1		**0.99**	1	
SIBTEST	0.96	0.98	1	0.96	0.94	1	0.94	0.94	1	0.88	0.9	1

	82			83		
MH	1			1		
NCDIF	**0.99**	1		**0.98**	1	
SIBTEST	0.89	0.9	1	0.82	0.88	1

	89			90			91			92		
MH	1			1			1			1		
NCDIF	**0.94**	1		**0.98**	1		**0.99**	1		**0.99**	1	
SIBTEST	0.94	0.97	1	0.97	0.93	1	0.93	0.93	1	0.86	0.88	1

	93			94		
MH	1			1		
NCDIF	**0.99**	1		**0.98**	1	
SIBTEST	0.85	0.87	1	0.81	0.87	1

	100			101			102			103		
MH	1			1			1			1		
NCDIF	**0.99**	1		**0.99**	1		**0.99**	1		**0.99**	1	
SIBTEST	1	0.99	1	1	0.99	1	1	0.99	1	1	0.99	1

	104			105		
MH	1			1		
NCDIF	**0.99**	1		**0.99**	1	
SIBTEST	1	0.99	1	0.99	0.99	1

	111			112			113			114		
MH	1			1			1			1		
NCDIF	**0.98**	1		**0.99**	1		**0.99**	1		**0.99**	1	
SIBTEST	1	0.99	1	1	0.99	1	1	0.99	1	1	0.99	1

	115			116		
MH	1			1		
NCDIF	**0.99**	1		**0.99**	1	
SIBTEST	1	0.99	1	1	0.99	1

	122			123			124			125		
MH	1			1			1			1		
NCDIF	**0.96**	1		**0.98**	1		**0.99**	1		**0.99**	1	
SIBTEST	0.99	0.98	1	1	0.97	1	1	0.98	1	1	0.98	1

	126			127		
MH	1			1		
NCDIF	**0.99**	1		**0.99**	1	
SIBTEST	1	0.98	1	0.99	0.98	1

	133			134			135			136		
MH	1			1			1			1		
NCDIF	**0.94**	1		**0.96**	1		**0.98**	1		**0.98**	1	
SIBTEST	0.99	0.97	1	1	0.96	1	1	0.96	1	0.99	0.96	1

	137			138		
MH	1			1		
NCDIF	**0.99**	1		**0.99**	1	
SIBTEST	0.99	0.97	1	0.99	0.97	1

	144			145			146			147		
MH	1			1			1			1		
NCDIF	**0.91**	1		**0.93**	1		**0.96**	1		**0.97**	1	
SIBTEST	0.93	0.99	1	0.99	0.96	1	1	0.94	1	0.99	0.93	1

	148			149		
MH	1			1		
NCDIF	**0.98**	1		**0.99**	1	
SIBTEST	0.99	0.95	1	0.98	0.95	1

	155			156			157			158		
MH	1			1			1			1		
NCDIF	**0.89**	1		**0.91**	1		**0.95**	1		**0.96**	1	
SIBTEST	0.9	0.98	1	0.98	0.95	1	1	0.92	1	0.99	0.91	1

	159			160		
MH	1			1		
NCDIF	**0.98**	1		**0.99**	1	
SIBTEST	0.98	0.93	1	0.97	0.93	1

	166			167			168			169		
MH	1			1			1			1		
NCDIF	**0.88**	1		**0.91**	1		**0.93**	1		**0.95**	1	
SIBTEST	0.97	0.95	1	1	0.91	1	1	0.91	1	0.99	0.89	1

	170			171		
MH	1			1		
NCDIF	**0.97**	1		**0.98**	1	
SIBTEST	0.98	0.92	1	0.95	0.9	1

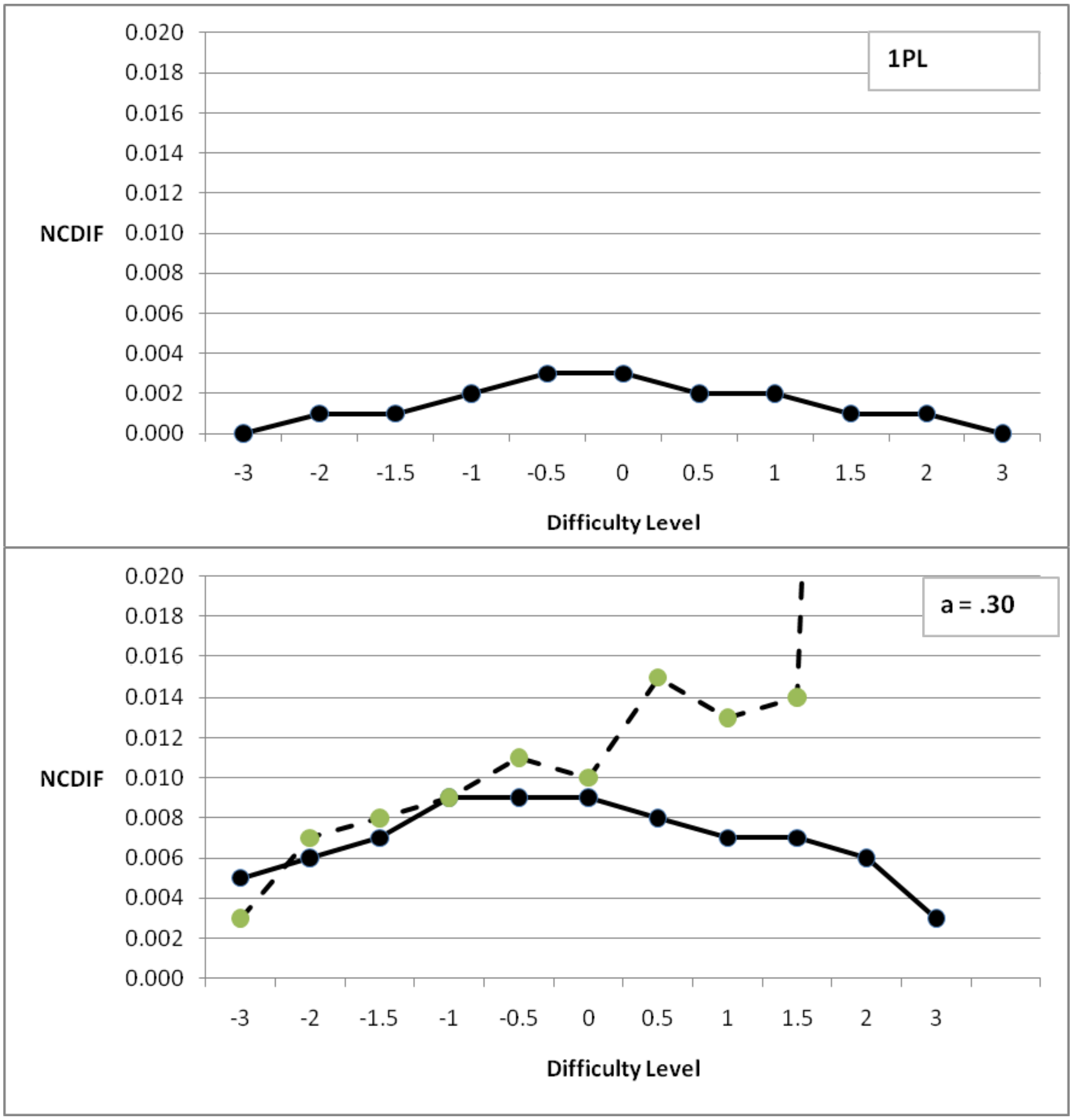
	177			178			179			180		
MH	1			1			1			1		
NCDIF	**0.87**	1		**0.9**	1		**0.93**	1		**0.94**	1	
SIBTEST	0.96	0.94	1	0.99	0.9	1	1	0.9	1	0.98	0.87	1

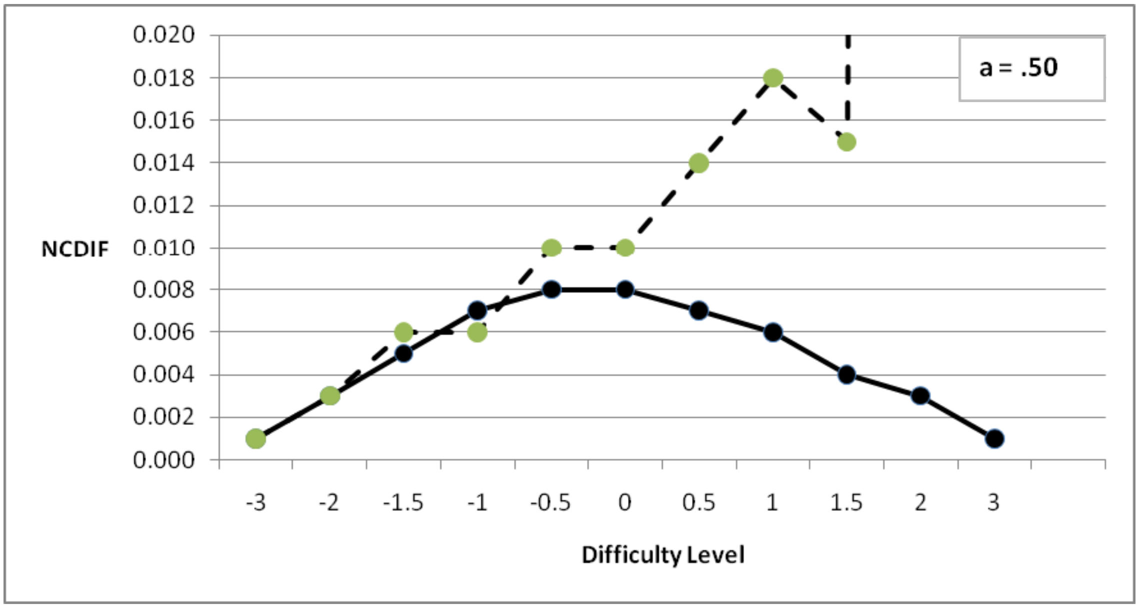
	181			182		
MH	1			1		
NCDIF	**0.97**	1		**0.98**	1	
SIBTEST	0.97	0.89	1	0.94	0.87	1

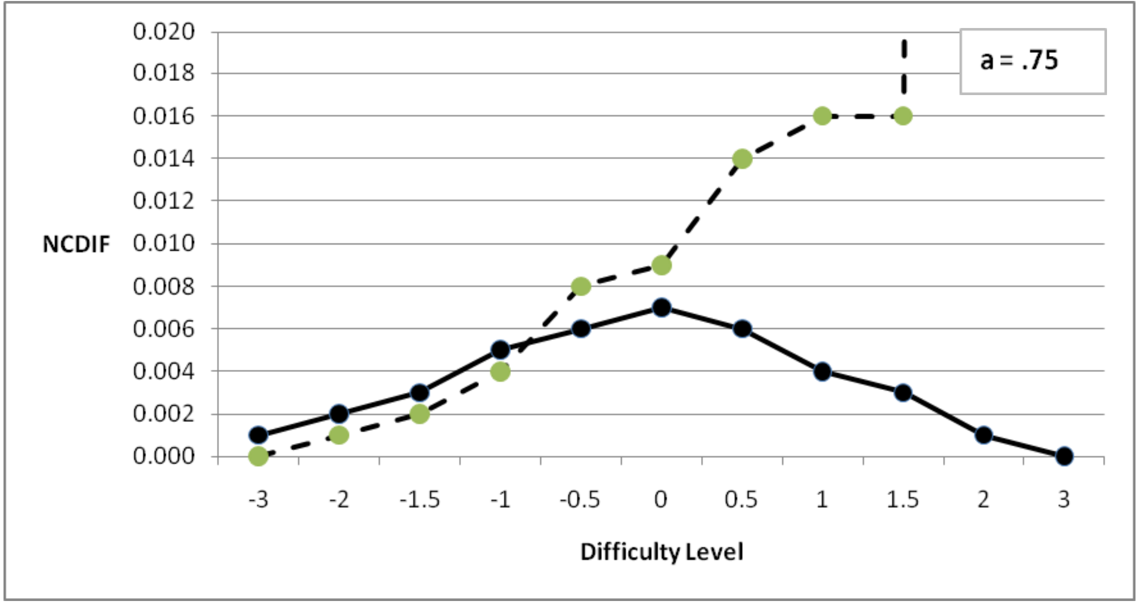
APPENDIX E

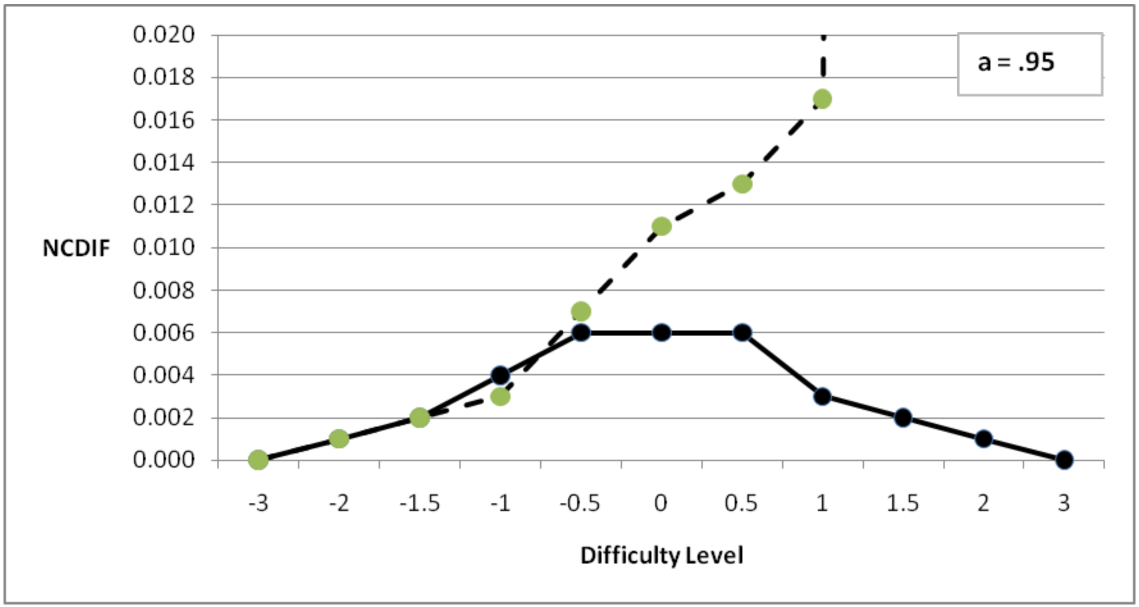
Graphical Relationship Relating– Model, Discrimination Parameter, Difficulty Level and Model Associated with NCDIF Moderate DIF (Category B)

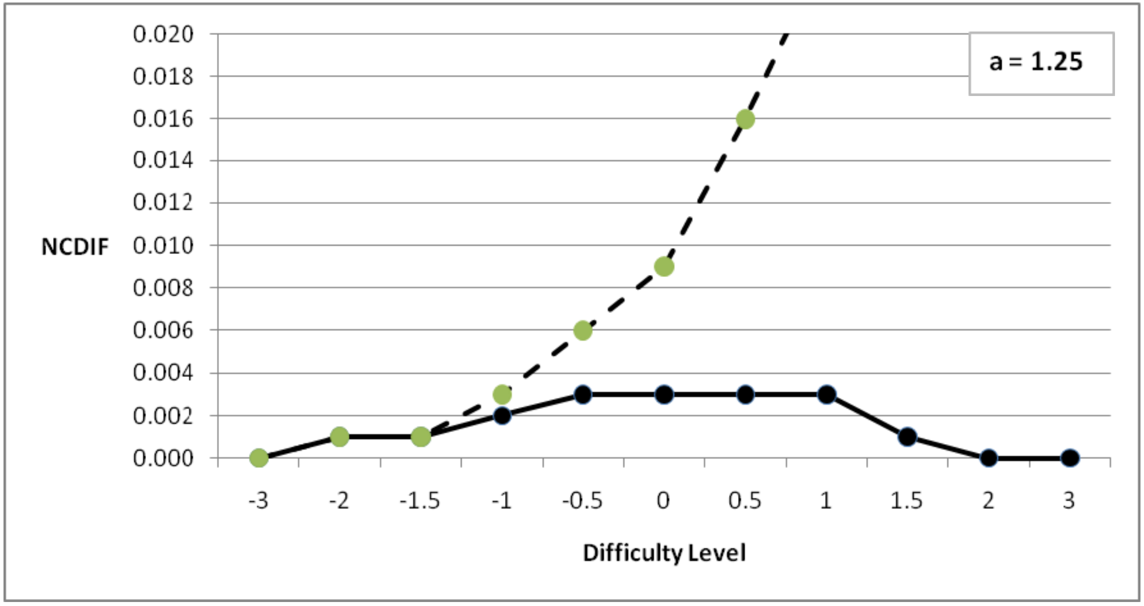
2PL Model – Solid Line, 3PL Model – Dash Line

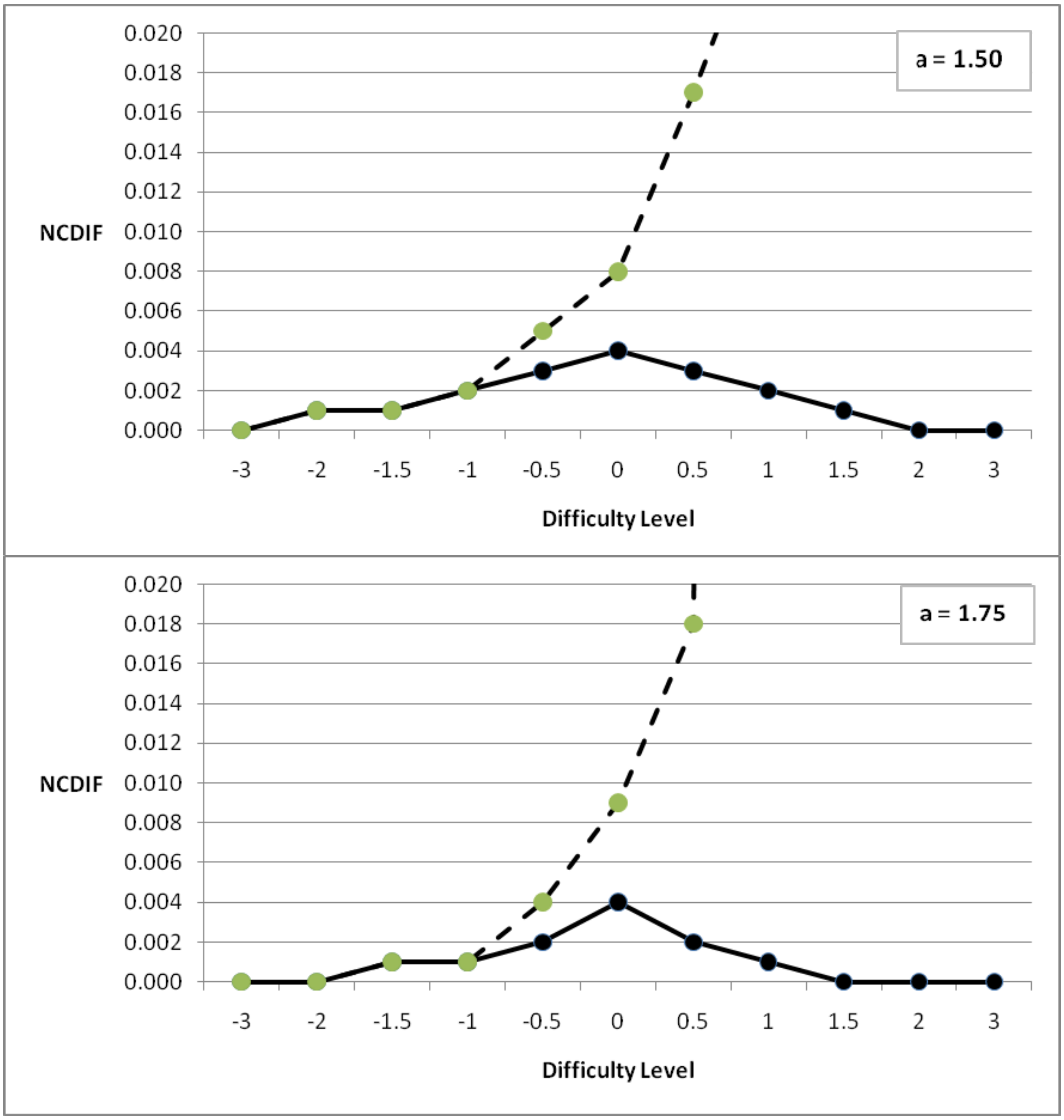


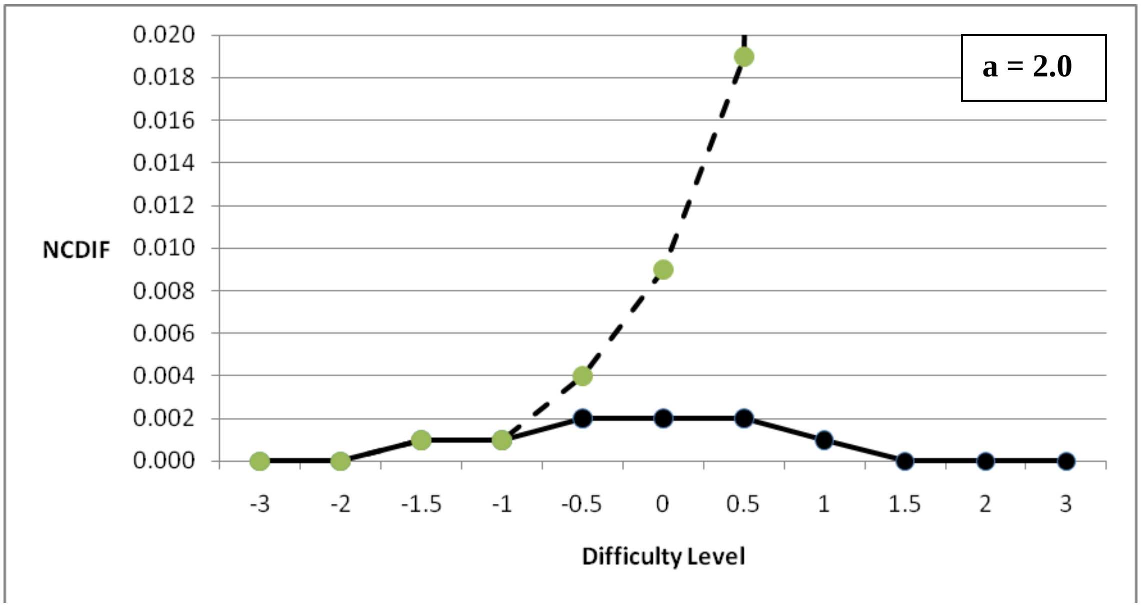

a = .50
NCDIF
0.020
0.018
0.016
0.014
0.012
0.010
0.008
0.006
0.004
0.002
0.000
-3
-2
-1.5
-1
-0.5
0
0.5
1
1.5
2
3
Difficulty Level


a = .75
NCDIF
0.020
0.018
0.016
0.014
0.012
0.010
0.008
0.006
0.004
0.002
0.000
-3
-2
-1.5
-1
-0.5
0
0.5
1
1.5
2
3
Difficulty Level


a = .95
NCDIF
0.020
0.018
0.016
0.014
0.012
0.010
0.008
0.006
0.004
0.002
0.000
-3
-2
-1.5
-1
-0.5
0
0.5
1
1.5
2
3
Difficulty Level


a = 1.25
NCDIF
0.020
0.018
0.016
0.014
0.012
0.010
0.008
0.006
0.004
0.002
0.000
-3
-2
-1.5
-1
-0.5
0
0.5
1
1.5
2
3
Difficulty Level


a = 1.50
NCDIF
0.020
0.018
0.016
0.014
0.012
0.010
0.008
0.006
0.004
0.002
0.000
-3
-2
-1.5
-1
-0.5
0
0.5
1
1.5
2
3
Difficulty Level
a = 1.75
NCDIF
0.020
0.018
0.016
0.014
0.012
0.010
0.008
0.006
0.004
0.002
0.000
-3
-2
-1.5
-1
-0.5
0
0.5
1
1.5
2
3
Difficulty Level


a = 2.0
NCDIF
0.020
0.018
0.016
0.014
0.012
0.010
0.008
0.006
0.004
0.002
0.000
-3
-2
-1.5
-1
-0.5
0
0.5
1
1.5
2
3
Difficulty Level

APPENDIX F

NCDIF Values –Seven Decimal Places

MH parameter for these 29 conditions corresponded to moderate DIF (Category B), whereas the NCDIF parameter value for these conditions is very small, < .001.

Condition	*a*	*b*	*c*	NCDIF MODERATE CATEGORY B	NCDIF LARGE CATEGORY C
1	N/A	-3	N/A	0.0003753	0.0008445
11	N/A	3	N/A	0.0002015	0.0004534
44	0.75	3	N/A	0.0002423	0.0005452
45	0.95	-3	N/A	0.0001157	0.0002604
55	0.95	3	N/A	0.0000935	0.0002104
56	1.25	-3	N/A	0.0000495	0.0001115
65	1.25	2	N/A	0.0003307	0.0007441
66	1.25	3	N/A	0.0000221	0.0000497
67	1.5	-3	N/A	0.0000276	0.0000622
76	1.5	2	N/A	0.0002183	0.0004912
77	1.5	3	N/A	0.0000105	0.0000236
78	1.75	-3	N/A	0.0000180	0.0000406
79	1.75	-2	N/A	0.0002665	0.0005997
86	1.75	1.5	N/A	0.0004790	0.0010778
87	1.75	2	N/A	0.0001394	0.0003137
88	1.75	3	N/A	0.0000052	0.0000117
89	2	-3	N/A	0.0000035	0.0000080
90	2	-2	N/A	0.0002349	0.0005287
97	2	1.5	N/A	0.0003256	0.0007326
98	2	2	N/A	0.0000934	0.0002102
99	2	3	N/A	0.0000029	0.0000065
122	0.75	-3	0.2	0.0002606	0.0005864
133	0.95	-3	0.2	0.0000740	0.0001666
144	1.25	-3	0.2	0.0000317	0.0000714
155	1.5	-3	0.2	0.0000177	0.0000398
166	1.75	-3	0.2	0.0000115	0.0000259
167	1.75	-2	0.2	0.0001706	0.0003838
177	2	-3	0.2	0.0000023	0.0000051
178	2	-2	0.2	0.0001503	0.0003383

Appendix G

NO IMPACT CONSTANTS

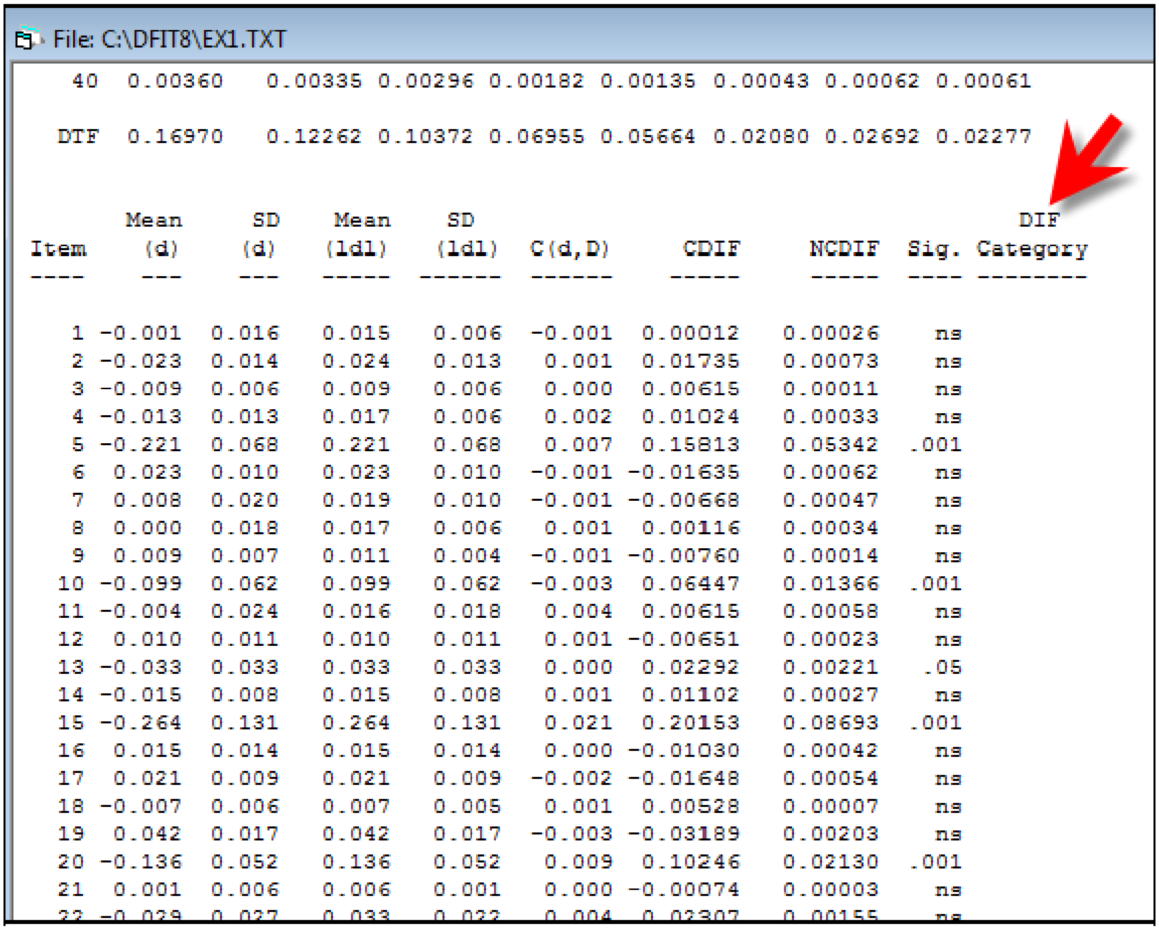
Linear Constants Relating - SIBTEST = (MH / K)

Model	1PL	2PL	2PL	2PL	2PL	2PL	2PL	2PL	2PL
Discrimination		0.30	0.50	0.75	0.95	1.25	1.50	1.75	2.00
Difficulty									
-3	19	11	13	16	19	24	27	30	31
-2	16	10	11	13	15	19	22	24	26
-1.5	14	10	12	12	14	15	18	20	23
-1	13	10	11	12	13	16	18	20	22
-0.5	13	10	11	11	13	15	16	18	20
0	12	10	12	14	16	19	22	24	27
0.5	12	10	12	15	16	21	23	N/A	N/A
1	12	12	13	20	26	36	45	N/A	N/A

Model	3PL	3PL	3PL	3PL	3PL	3PL	3PL	3PL
Discrimination	0.30	0.50	0.75	0.95	1.25	1.50	1.75	2.00
Difficulty								
-3	12	15	16	18	21	23	23	24
-2	10	11	12	13	13	11	15	15
-1.5	10	11	11	12	12	13	13	13
-1	10	10	11	11	11	11	11	11
-0.5	10	10	10	11	11	11	11	11
0	10	10	11	11	12	12	12	12

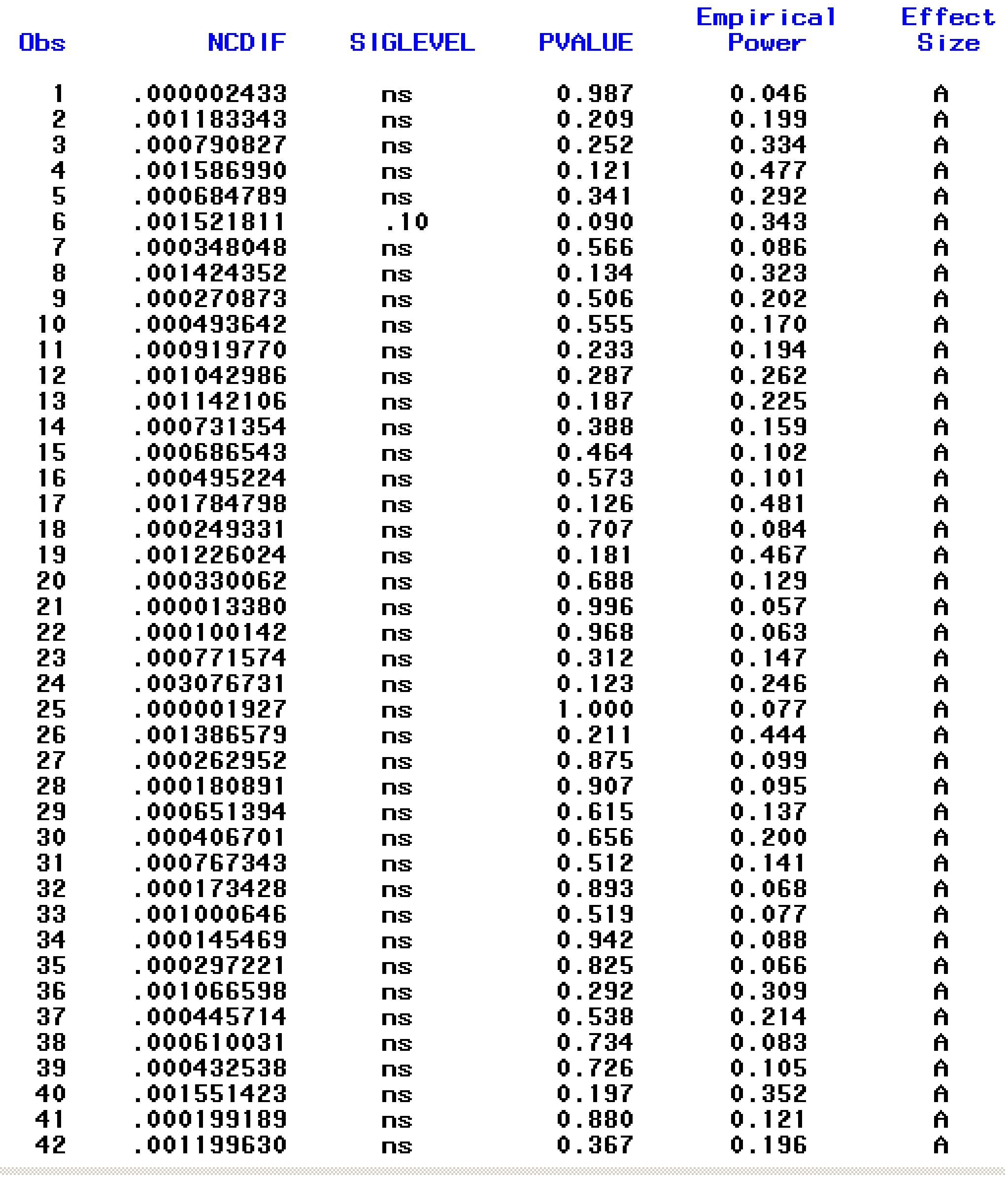
APPENDIX H

Example – DFIT8 Output without a DIF Category or Power

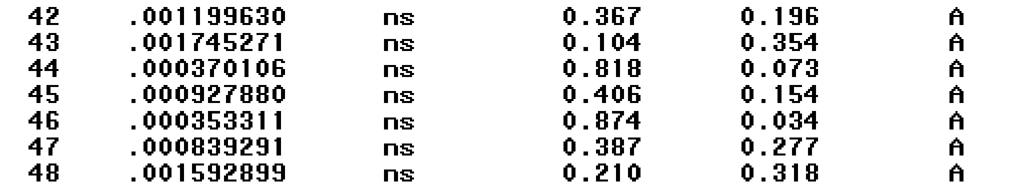


```
   40  0.00360    0.00335 0.00296 0.00182 0.00135 0.00043 0.00062 0.00061

  DTF  0.16970    0.12262 0.10372 0.06955 0.05664 0.02080 0.02692 0.02277


         Mean      SD     Mean     SD                                            DIF
 Item     (d)      (d)    (|d|)   (|d|)   C(d,D)      CDIF       NCDIF   Sig.  Category
 ----     ---      ---    -----   ------  ------      -----      -----   ----  --------


    1 -0.001   0.016   0.015   0.006  -0.001  0.00012   0.00026    ns
    2 -0.023   0.014   0.024   0.013   0.001  0.01735   0.00073    ns
    3 -0.009   0.006   0.009   0.006   0.000  0.00615   0.00011    ns
    4 -0.013   0.013   0.017   0.006   0.002  0.01024   0.00033    ns
    5 -0.221   0.068   0.221   0.068   0.007  0.15813   0.05342  .001
    6  0.023   0.010   0.023   0.010  -0.001 -0.01635   0.00062    ns
    7  0.008   0.020   0.019   0.010  -0.001 -0.00668   0.00047    ns
    8  0.000   0.018   0.017   0.006   0.001  0.00116   0.00034    ns
    9  0.009   0.007   0.011   0.004  -0.001 -0.00760   0.00014    ns
   10 -0.099   0.062   0.099   0.062  -0.003  0.06447   0.01366  .001
   11 -0.004   0.024   0.016   0.018   0.004  0.00615   0.00058    ns
   12  0.010   0.011   0.010   0.011   0.001 -0.00651   0.00023    ns
   13 -0.033   0.033   0.033   0.033   0.000  0.02292   0.00221   .05
   14 -0.015   0.008   0.015   0.008   0.001  0.01102   0.00027    ns
   15 -0.264   0.131   0.264   0.131   0.021  0.20153   0.08693  .001
   16  0.015   0.014   0.015   0.014   0.000 -0.01030   0.00042    ns
   17  0.021   0.009   0.021   0.009  -0.002 -0.01648   0.00054    ns
   18 -0.007   0.006   0.007   0.005   0.001  0.00528   0.00007    ns
   19  0.042   0.017   0.042   0.017  -0.003 -0.03189   0.00203    ns
   20 -0.136   0.052   0.136   0.052   0.009  0.10246   0.02130  .001
   21  0.001   0.006   0.006   0.001   0.000 -0.00074   0.00003    ns
   22 -0.029   0.027   0.033   0.022   0.004  0.02307   0.00155    ns
```

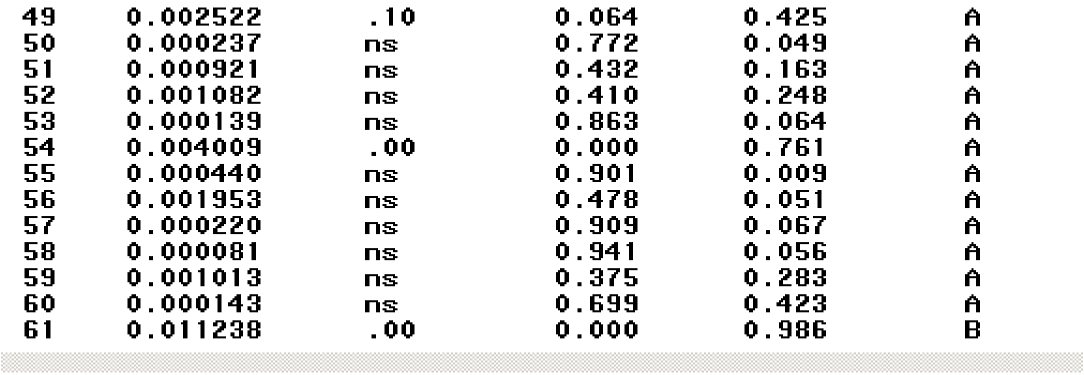
Example – DFIT Output with a DIF Category and Power



```
                                                       Empirical     Effect
Obs          NCDIF       SIGLEVEL       PVALUE           Power         Size

  1       .000002433        ns           0.987           0.046          A
  2       .001183343        ns           0.209           0.199          A
  3       .000790827        ns           0.252           0.334          A
  4       .001586990        ns           0.121           0.477          A
  5       .000684789        ns           0.341           0.292          A
  6       .001521811        .10          0.090           0.343          A
  7       .000348048        ns           0.566           0.086          A
  8       .001424352        ns           0.134           0.323          A
  9       .000270873        ns           0.506           0.202          A
 10       .000493642        ns           0.555           0.170          A
 11       .000919770        ns           0.233           0.194          A
 12       .001042986        ns           0.287           0.262          A
 13       .001142106        ns           0.187           0.225          A
 14       .000731354        ns           0.388           0.159          A
 15       .000686543        ns           0.464           0.102          A
 16       .000495224        ns           0.573           0.101          A
 17       .001784798        ns           0.126           0.481          A
 18       .000249331        ns           0.707           0.084          A
 19       .001226024        ns           0.181           0.467          A
 20       .000330062        ns           0.688           0.129          A
 21       .000013380        ns           0.996           0.057          A
 22       .000100142        ns           0.968           0.063          A
 23       .000771574        ns           0.312           0.147          A
 24       .003076731        ns           0.123           0.246          A
 25       .000001927        ns           1.000           0.077          A
 26       .001386579        ns           0.211           0.444          A
 27       .000262952        ns           0.875           0.099          A
 28       .000180891        ns           0.907           0.095          A
 29       .000651394        ns           0.615           0.137          A
 30       .000406701        ns           0.656           0.200          A
 31       .000767343        ns           0.512           0.141          A
 32       .000173428        ns           0.893           0.068          A
 33       .001000646        ns           0.519           0.077          A
 34       .000145469        ns           0.942           0.088          A
 35       .000297221        ns           0.825           0.066          A
 36       .001066598        ns           0.292           0.309          A
 37       .000445714        ns           0.538           0.214          A
 38       .000610031        ns           0.734           0.083          A
 39       .000432538        ns           0.726           0.105          A
 40       .001551423        ns           0.197           0.352          A
 41       .000199189        ns           0.880           0.121          A
 42       .001199630        ns           0.367           0.196          A
```



```
 42      .001199630        ns            0.367           0.196          A
 43      .001745271        ns            0.104           0.354          A
 44      .000370106        ns            0.818           0.073          A
 45      .000927880        ns            0.406           0.154          A
 46      .000353311        ns            0.874           0.034          A
 47      .000839291        ns            0.387           0.277          A
 48      .001592899        ns            0.210           0.318          A
```



49	0.002522	.10	0.064	0.425	A
50	0.000237	ns	0.772	0.049	A
51	0.000921	ns	0.432	0.163	A
52	0.001082	ns	0.410	0.248	A
53	0.000139	ns	0.863	0.064	A
54	0.004009	.00	0.000	0.761	A
55	0.000440	ns	0.901	0.009	A
56	0.001953	ns	0.478	0.051	A
57	0.000220	ns	0.909	0.067	A
58	0.000081	ns	0.941	0.056	A
59	0.001013	ns	0.375	0.283	A
60	0.000143	ns	0.699	0.423	A
61	0.011238	.00	0.000	0.986	B

APPENDIX I

Monte Carlo Simulation - SAS Programs

```
/**********************************************************************
 * Programmer: Keith D. Wright
 * Date: 3/16/2011
 * Georgia State University
 * Dissertation: Improvements For Differential Functioning of Item
 * & Tests (DFIT): Investigating The Addition of Reporting An Effect
 * Size Measure and Power
 *
 * This is the main program which automates the Monte Carlo simulation
 * study.
 * This program was part of IRTGEN, Whittaker, Fitzpatrick, Williams,
 * and Dodd (2003), with significant modifications for this study.
 *
 * The program reads in several reference group files containing item
 * parameters for 61 test items.  For each file, IRTGEN is invoked with
 * the file, where random response data (1s & Os) are created for 1000
 * examinees.
 *
 * The program then reads in focal group files containing item
 * parameters for the 61 test items.  For each file, IRTGEN is
 * invoked with the file, where random response data (1s & Os) are
 * created for 1000 examinees.
 *
 * IRTGEN is invoked again with the merge flag set to 1, which will
 * cause IRTGEN to merge the response data file for the reference
 * group examinees and focal group examinees into one file. This will
 * result into numerous Mantel-Haenszel files being created for
 * analysis purposes.
 *
/**********************************************************************
%macro simtimes(simnum);
/* Used to control the number of replications for the Monte Carlo
study. */
%DO s=1 %to &simnum;
options nonotes nosource nosource2 errors=0;

%macro reffactors;
      FILENAME IO 'C:\Documents andSettings\SPR2011\Dissertation_Sftw';
      %INCLUDE IO(IRTGEN);
      %do i=80 %to 110;
      %IF (&i=80)or(&i=82)or(&i=84)or(&i=86)or(&i=88)%THEN %DO;
            %do j=1 %to 40;
            %if &s=1 %then %do;
            DATA L1&i&j;
            INFILE IO(ref&i&j);
            INPUT A B C;
            %IRTGEN(DATA=L1&i&j, OUT=REFOUT&i&j, NI=61, NE=1000, GRP=0,
            MERGE=0, thetaflag1=&s, thetaflag2=1);
            %end;
            %else %if &s^=1 %then %do;
            %IRTGEN(DATA=L1&i&j, OUT=REFOUT&i&j, NI=61, NE=1000, GRP=0,
            MERGE=0, thetaflag1=&s, thetaflag2=1);
            %end;
            %end;
      %END;
      %ELSE %IF (&i=81)or(&i=85)or(&i=89)or(&i=93)or(&i=97)or
      (&i=100)or(&i=103)or(&i=106)or(&i=109)%THEN %DO;
            %do j=1 %to 60;
                  %if &s=1 %then %do;
                  DATA L1&i&j;
                  INFILE IO(ref&i&j);
                  INPUT A B C;
                  %IRTGEN(DATA=L1&i&j, OUT=REFOUT&i&j, NI=61, NE=1000,
                  GRP=0, MERGE=0, thetaflag1=&s, thetaflag2=1);
                  %end;
                  %else %if &s^=1 %then %do;
                  %IRTGEN(DATA=L1&i&j, OUT=REFOUT&i&j, NI=61, NE=1000,
                  GRP=0, MERGE=0, thetaflag1=&s, thetaflag2=1);
                  %end;
            %end;
      %END;
      %ELSE %IF (&i=83)or(&i=87)or(&i=91)or(&i=95)%THEN %DO;
            %do j=1 %to 20;
                  %if &s=1 %then %do;
                  DATA L1&i&j;
                  INFILE IO(ref&i&j);
                  INPUT A B C;
                  %IRTGEN(DATA=L1&i&j, OUT=REFOUT&i&j, NI=61, NE=1000,
                  GRP=0, MERGE=0, thetaflag1=&s, thetaflag2=1);
                  %end;
                  %else %if &s^=1 %then %do;
                  %IRTGEN(DATA=L1&i&j, OUT=REFOUT&i&j, NI=61, NE=1000,
                  GRP=0, MERGE=0, thetaflag1=&s, thetaflag2=1);
                  %end;
            %end;
      %END;
%end;

%mend reffactors; /* Ending Macro reffactors */
%reffactors; /* Invoke the Macro reffactors */


/* Macro for creating the focal group random response data */
%macro focalfiles;
FILENAME IO 'C:\Documents and Settings\SPR2011\Dissertation_Sftw';
%INCLUDE IO(IRTGEN);
%do i=80 %to 110;
%IF
(&i=80)or(&i=82)or(&i=84)or(&i=86)or(&i=88)or(&i=90)or(&i=92)or(&i=94)o
r(&i=96)or(&i=98)or(&i=99)or(&i=101)or(&i=102)or(&i=104)or(&i=105)or(&i
=107)or(&i=108)or(&i=110)%THEN %DO;
      %do j=1 %to 40;
      %if &s=1 %then %do;
      DATA L2&i&j;
      INFILE IO(focal&i&j);
      INPUT A B C;
      %IRTGEN(DATA=L2&i&j, OUT=OUT&i&j, NI=61, NE=1000, GRP=1, MERGE=0,
      thetaflag1=&s, thetaflag2=1);
      %end;
      %else %if &s^=1 %then %do;
      %IRTGEN(DATA=L2&i&j, OUT=OUT&i&j, NI=61, NE=1000, GRP=1, MERGE=0,
      thetaflag1=&s, thetaflag2=1);
      %end;
%end;
%END;
%ELSE %IF
(&i=81)or(&i=85)or(&i=89)or(&i=93)or(&i=97)or(&i=100)or(&i=103)or(&i=10
6)or(&i=109)%THEN %DO;
      %do j=1 %to 60;
      %if &s=1 %then %do;
      DATA L2&i&j;
      INFILE IO(focal&i&j);
      INPUT A B C;
      %IRTGEN(DATA=L2&i&j, OUT=OUT&i&j, NI=61, NE=1000, GRP=1, MERGE=0,
      thetaflag1=&s, thetaflag2=1);
      %end;
      %else %if &s^=1 %then %do;
      %IRTGEN(DATA=L2&i&j, OUT=OUT&i&j, NI=61, NE=1000, GRP=1, MERGE=0,
      thetaflag1=&s, thetaflag2=1);
      %end;
%end;
%END;
%ELSE %IF (&i=83)or(&i=87)or(&i=91)or(&i=95)%THEN %DO;
      %do j=1 %to 20;
      %if &s=1 %then %do;
      DATA L2&i&j;
      INFILE IO(focal&i&j);
      INPUT A B C;
      %IRTGEN(DATA=L2&i&j, OUT=OUT&i&j, NI=61, NE=1000, GRP=1, MERGE=0,
      thetaflag1=&s, thetaflag2=1);
      %end;
      %else %if &s^=1 %then %do;
      %IRTGEN(DATA=L2&i&j, OUT=OUT&i&j, NI=61, NE=1000, GRP=1, MERGE=0,
      thetaflag1=&s, thetaflag2=1);
      %end;
%end;
%END;
%end;
/* This statement invokes IRTGEN so that the reference group */
/* and focal group response data s merged together for the Mantel-
Haenszel*/ /* analysis */
%IRTGEN(DATA=L28040, OUT=OUT200, NI=61, NE=1000, GRP=1, MERGE=1,
thetaflag1=&s, thetaflag2=1);


%mend focalfiles; /* Ending Macro focalfiles */
%focalfiles; /* Invoke the Macro focalfiles */


/* The next section of code is for the Mantel-Haenszel analysis */
%macro mhdif(num);

%do i=80 %to 110;
%IF
(&i=80)or(&i=82)or(&i=84)or(&i=86)or(&i=88)or(&i=90)or(&i=92)or(&i=94)o
```

```
r(&i=96)or(&i=98)or(&i=99)or(&i=101)or(&i=102)or(&i=104)or(&i=105)or(&i
=107)or(&i=108)or(&i=110)%THEN %DO;
      %do j=1 %to 40;
      %do k=61 %to &num;
      DATA look4dif;
      infile "C:\Documents and Settings\mh&i&j..dat";
      input group item1-item61;
      score = sum(of item1-item61);
      RUN;


      PROC RANK data=look4dif out=Ability_Groups groups=5;
      var score;
      ranks stratum;


      PROC FREQ Data=Ability_Groups noprint;
      Tables stratum*group*item&k/CMH norow nocol nopercent;

      %IF &k = 61 %THEN %DO;
      output out= out&i&j&k CMH; /*Creating the DIF tables */
      %END;

      RUN;

      %IF (&k = 61) and (&s = 1 or &s = 25 or &s = 50 or &s = 100 or &s
      = 150 or &s = 200 or &s = 250 or &s = 400 or &s = 500) %THEN %DO;
      %put &i&j&k&s; /* Only used to track the place in the simulation
      study */
      %END;
      %end;
      %end;
      %END;
      %ELSE %IF
      (&i=81)or(&i=85)or(&i=89)or(&i=93)or(&i=97)or(&i=100)or(&i=103)or
      (&i=106)or(&i=109)%THEN %DO;
      %do j=1 %to 60;
      %do k=61 %to &num;
      DATA look4dif;
      infile "C:\Documents and Settings\mh&i&j..dat";
      input group item1-item61;
      score = sum(of item1-item61);
      RUN;


      PROC RANK data=look4dif out=Ability_Groups groups=5;
      var score;
      ranks stratum;


      PROC FREQ Data=Ability_Groups noprint;
      Tables stratum*group*item&k/CMH norow nocol nopercent;

      %IF &k = 61 %THEN %DO;
      output out= out&i&j&k CMH; /*Creating the MH DIF tables */
      %END;
      RUN;

      %IF (&k = 61) and (&s = 1 or &s = 25 or &s = 50 or &s = 100 or &s
      = 150 or &s = 200 or &s = 250 or &s = 400 or &s = 500) %THEN %DO;
      %put &i&j&k&s; /* Only used to track the place in the simulation
      study */
      %END;
      %end;
%end;
%END;
%ELSE %IF (&i=83)or(&i=87)or(&i=91)or(&i=95)%THEN %DO;
      %do j=1 %to 20;
      %do k=61 %to &num;
      DATA look4dif;
      infile "C:\Documents and Settings\mh&i&j..dat";
      input group item1-item61;
      score = sum(of item1-item61);
      RUN;


      PROC RANK data=look4dif out=Ability_Groups groups=5;
      var score;
      ranks stratum;


      PROC FREQ Data=Ability_Groups noprint;
      Tables stratum*group*item&k/CMH norow nocol nopercent;

      %IF &k = 61 %THEN %DO;
      output out= out&i&j&k CMH; /*Creating the DIF tables */
      %END;

      RUN;

      %IF (&k = 61) and (&s = 1 or &s = 25 or &s = 50 or &s = 100 or &s
      = 150 or &s = 200 or &s = 250 or &s = 400 or &s = 500) %THEN %DO;
      %put &i&j&k&s; /* Only used to track the place in the simulation
      study */
      %END;
      %end;
      %end;
      %END;
      %end;

/* Formatting the DIF output for analysis purposes */
data all&s (RENAME=(_MHOR_=md P_CMHRMS=mh_pvalue));
set
%do i = 80 %to 110;
%IF
(&i=80)or(&i=82)or(&i=84)or(&i=86)or(&i=88)or(&i=90)or(&i=92)or(&i=94)o
r(&i=96)or(&i=98)or(&i=99)or(&i=101)or(&i=102)or(&i=104)or(&i=105)or(&i
=107)or(&i=108)or(&i=110)%THEN %DO;
%do j = 1 %to 40;
out&i&j&num
%end;
%END;
```

```
%ELSE %IF
(&i=81)or(&i=85)or(&i=89)or(&i=93)or(&i=97)or(&i=100)or(&i=103)or(&i=10
6)or(&i=109)%THEN %DO;
%do j = 1 %to 60;
out&i&j&num
%end;
%END;
%ELSE %IF (&i=83)or(&i=87)or(&i=91)or(&i=95)%THEN %DO;
%do j = 1 %to 20;
out&i&j&num
%end;
%END;
%end;
;
run;


/* Determine the size of DIF based on Mantel-Haenszel */
/* effect size guidelines */
DATA final&s; set all&s;
const = -2.3529;
mhd=const*(log(md));

/* Used for Type I and Type II analysis */
if mh_pvalue > .05 then pvalue=0;else pvalue = 1;


/* Specifying which variables to keep from the MH analysis */
keep md const mhd pvalue;
RUN; /* End determining the size of DIF */

/* Capture the results */
%IF &s = 100 %THEN %DO;
      DATA results;
      merge
      %do n = 1 %to 100;
      final&n(rename = (mhd = run&n) drop=md const pvalue)
      %end;
      ;
      RUN;
      data _null_; set results;
      file 'C:\Documents and Settings\output\resultsout100.txt';
      put run1-run100;
      RUN;

      /* First Set of Files */
      DATA resultsout1; set results;
      file "C:\Documents and Settings\output\output\resultsout100.dat";
      put run1-run100;
      RUN;
      DATA resultsout1;
      INFILE "C:\Documents and
      Settings\output\output\resultsout100.dat";
      INPUT run1-run100;
      RUN;
      PROC EXPORT DATA=resultsout1
      OUTFILE="C:\Documents and
      Settings\output\output\resultsout100.xls";
      RUN;


      DATA power;
      merge
      %do n = 1 %to 100;
      final&n(rename = (pvalue = run&n) drop=md const mhd)
      %end;
      ;
      RUN;
      data _null_; set power;
      file 'C:\Documents and Settings\output\output\powerout100.txt';
      put run1-run100;
      RUN;

      /* First Set of Files */
      DATA powerout1; set power;
      file "C:\Documents and Settings\output\output\power100.dat";
      put run1-run100;
      RUN;
      DATA powerout1;
      INFILE "C:\Documents and Settings\output\output\power100.dat";
      INPUT run1-run100;
      RUN;
      PROC EXPORT DATA=powerout1
      OUTFILE="C:\Documents and Settings\output\output\power100.xls";
      RUN;
%END;


%mend; /* Ending Macro mhdif */
%mhdif(61); /* Invoke the Macro mhdif */


%END; /* Ending TOP do loop, where the number of replications is
running */
%mend;/* Ending Macro simtimes */

/* Used to control the number of replications for the Monte Carlo
study. */
%simtimes(100);
```

```
/**********************************************************************
 * Programmer: Keith D. Wright
 * Date: 3/16/2011
 * Georgia State University
 *
 * The majority of this program was taken from IRTGEN and modified
 * for the purposes of this dissertation
 * (Whittaker, Fitzpatrick, Williams, & Dodd (2003).
/**********************************************************************


%LET DIST='NORMAL';
%LET SEED=34561;


%MACRO IRTGEN(DATA=_LAST_, OUT=GEN, NI=, NE=, GRP=, MERGE=,
thetaflag1=, thetaflag2=);

    %MACRO L3GEN;
            GROUP = 0;
/* Used to control reference versus focal group files */
              %IF &GRP = 1 %THEN %do;
/* Reference group ID will be 0 and focal group ID 1 */
               GROUP = 1;
/* This ordering is necessary for accurate MH analysis */
              %end;

              P=C+(1-C)*(1/(1+exp(-1.7*A*(THETA-B))));

/* The next four lines are key for an accurate MH analysis.
/* MH analysis wants the correct response to be in the first
/* column of PROC FREQ, therefore, a lower number will be assign,
/* (i.e. 7) for a correct response. If the traditional coding of 0 for
/* incorrect and 1 for correct is used, the MH analysis will be
/* backwards generating DIF favoring focal versus reference. The number
/* 9 is used to represent an incorrect response, typically 0 is used.

              IF P GE RANUNI(-1) THEN R(J)=7;
            /* Results into a correct response if probability is */
            /* greater */
              ELSE R(J)=9;
            /* than a randomly generated probability else incorrect*/

    %MEND L3GEN;

    %LET FLAG=0;
    %LET MDL=L3GEN;
      %IF %LENGTH(&NI)=0 OR &NI=0 %THEN %DO;
         %PUT;
         %PUT ***** ERROR ***** YOU MUST SPECIFY NUMBER OF ITEMS *****;
         %PUT;
         %LET FLAG=1;
      %END;
    %IF %LENGTH(&NE)=0 OR &NE=0 %THEN %DO;
         %PUT;
         %PUT ***** ERROR ***** YOU MUST SPECIFY NUMBER OF EXAMINEES
*****;
```

```
         %PUT;
         %LET FLAG=1;
      %END;

    %IF &FLAG=0 %THEN %DO;
        DATA THETA;
               KEEP THETA;
               CALL STREAMINIT(&SEED+&thetaflag1);
           DO I=1 TO &NE;
              IF &GRP=1 THEN
                  THETA=RAND(&DIST);
                 /*THETA=-1+1*RAND(&DIST);*/ /* Impact Case */
              ELSE THETA=RAND(&DIST);
                   OUTPUT;
         END;
        RUN;

            DATA &OUT;
            KEEP GROUP THETA R1-R&NI;
            ARRAY R(*) R1-R&NI;
            SET THETA;
            DO J=1 TO &NI;
                  SET &DATA POINT=J; %&MDL;
            END;
             RUN;

/* These next statements are for merging the response data of the */
/* reference and focal group */
         %IF &MERGE = 1 %THEN %DO;
              %do i=80 %to 110;
                  %IF
(&i=80)or(&i=82)or(&i=84)or(&i=86)or(&i=88)or(&i=90)or(&i=92)or(&i=94)o
r(&i=96)or(&i=98)or(&i=99)or(&i=101)or(&i=102)or(&i=104)or(&i=105)or(&i
=107)or(&i=108)or(&i=110)%THEN %DO;
                        %do j=1 %to 40;
                        DATA merge&i&j; set refout&i&j out&i&j;
                        RUN;
                  /* Output the merge files for the MH Analysis */
                        DATA mh&i&j; SET merge&i&j;
                        file "C:\Documents and Settings\mh&i&j..dat";
                        put GROUP R1-R61;
                        RUN;
                        %end; /* End 1 - 40 loop */
                  %END;
                  %ELSE %IF
(&i=81)or(&i=85)or(&i=89)or(&i=93)or(&i=97)or(&i=100)or(&i=103)or(&i=10
6)or(&i=109)%THEN %DO;
                        %do j=1 %to 60;
                        DATA merge&i&j; set refout&i&j out&i&j;
                        RUN;
                  /* Output the merge files for the MH Analysis */
                        DATA mh&i&j; SET merge&i&j;
                        file "C:\Documents and Settings\mh&i&j..dat";
                        put GROUP R1-R61;
                        RUN;
                        %end; /* End 1 - 60 loop */
                  %END;
                  %ELSE %IF (&i=83)or(&i=87)or(&i=91)or(&i=95)%THEN
%DO;
                        %do j=1 %to 20;
                        DATA merge&i&j; set refout&i&j out&i&j;
                        RUN;
                  /* Output the merge files for the MH Analysis */
                        DATA mh&i&j; SET merge&i&j;
                        file "C:\Documents and Settings\mh&i&j..dat";
                        put GROUP R1-R61;
                        RUN;
                        %end; /* End 1 - 20 loop */
                  %END;

            %END; /* End 80 - 96 loop */

            /* Code necessary to get Thetas for NCDIF calculations */
                  %IF (&thetaflag1 = 85) AND (&thetaflag2 = 1) %THEN
%DO;
                        DATA getthetas;
                              merge
                              %do i=80 %to 80;
                                    out&i (drop=GROUP R1-R61);
                              %end;
                        RUN;

                        PROC EXPORT DATA=getthetas
                        OUTFILE="C:\output\thetasout.xls";
                        RUN;

                  %END;

        %END; /* End MERGE */

    %END; /* End IF FLAG = 0 */

%MEND IRTGEN;
```

APPENDIX J

Monte Carlo Simulation - SAS Program Integrating SIBTEST

```
/**********************************************************************
 * Programmer: Keith D. Wright
 * Date: 3/16/2011
 * Georgia State University
 *
 * Only a portion of the SIBTEST code is included.  This code was added
 * specifically for the Monte Carlo study for SIBTEST’s estimation.
 * This code demonstrates how to automate SIBTEST.
/**********************************************************************
%IF(&i=80)or(&i=82)or(&i=84)or(&i=86)or(&i=88) %THEN %DO;
                  %do j=1 %to 40;
                        proc iml;
                        FILENAME OUT 'C:\Program Files\sibtest\sib.in';
                        FILE OUT;
                        PUT @1 '61'/
                        @1 "C:\SIBTEST\refresp&i..dat"/
                        @1 "C:\SIBTEST\focresp&i&j..dat"/
                        @1 '1'/
                        @1 '"C:\Documents and Settings\SIB.txt"'/
                        @1 '20'/
                        @1 '1'/
                        @1 '0'/
                        @1 '1' //
                        @1 '1'/
                        @1 '61'/
                        @1 '''f'''/
                        @1 '60'/
                        @1 '1  2  3  4  5'/
                        @1 '6  7  8  9  10'/
                        @1 '11  12  13  14  15'/
                        @1 '16  17  18  19  20'/
                        @1 '21  22  23  24  25'/
                        @1 '26  27  28  29  30'/
                        @1 '31  32  33  34  35'/
                        @1 '36  37  38  39  40'/
                        @1 '41  42  43  44  45'/
                        @1 '46  47  48  49  50'/
                        @1 '51  52  53  54  55'/
                        @1 '56  57  58  59  60'/
                        @1 '0.2';
                        CLOSEFILE OUT;
                        start system(command);
                              call push(" x '",command,"'; resume;");
                              pause;
                              finish;
                              run system('c:\SIBTEST\auto_commands');
                        quit;

                        data sib&i&j;
                        INFILE 'C:\Documents and
                        Settings\Desktop\SIB.txt';
                        *Move 81 lines to retrieve sibtest statistic;
                        INPUT
                  /////////////////////////////////////////////////
                        ///////////////////////////////
                        sibtest 31-36  pvalue 47-52;
                        if pvalue > .05 then pvalue=0;else pvalue = 1;
                        RUN;
                        %end;
%END;
```

APPENDIX K

DIFCUT – Added Effect Size, Power, P-VAL., and Modified Output (see Appendix H)

```
/*******************************************************************/
 *  DIFCUT: A Program to determine NCDIF and DTF cutoff scores
 *
 *  Nanda, A. O., Oshima, T. C., & Gagné, P. (2006).
 *
 *  Modified 2/27/2011 – K. D. Wright
 * (Added Effect Size, Power, P-VALUE, and Modified Output)
 *
 *  DIFCUT: A SAS/IML Program for Conducting Significance Tests for
 *          Differential Functioning of Items and Tests (DFIT)
 *          [Computer software].
 *          Atlanta, GA: Georgia State University.
 *
 *  4 input files (focal.cov, focal.sco, reference.cov, link.lin)
 *  Default number of replications = 1000
/*********************************************************************/

options formdlim=' ';
FILENAME IO 'C:\powerdissertation';
data cov;
/*In parentheses below, user must enter the name of their focal group
file with the .cov extension*/
      INFILE IO (focal.cov) missover firstobs=3;
      input id 1-5
                  item $ 6-13
                  test $ 14-20
                  group 21
                  a
                  b
                  c
                  avar
                  abcov
                  /
                  bvar
                  accov
                  bccov
                  cvar;

                  asd=sqrt(avar);
                  bsd=sqrt(bvar);
                  csd=sqrt(cvar);

data sco;
/*In parentheses below, user must enter the name of their focal group
file with the .sco extension*/
      INFILE IO (focal.sco) missover firstobs=3;
      input group
                  id $
                  /
                  resp 1-6
                  calib 7-7
                  subtest $ 8-15
                  attempt 16-20
                  correct 21-25
                  percent 26-35
                  theta 36-47
                  stderr 48-59
                  stdunest 60-60
                  grpprob 61-70
                  margprob 71-80;

data covref;
/*In parentheses below, user must enter the name of their reference
group file with the .cov extension*/
      INFILE IO (reference.cov) missover firstobs=3;
      input id 1-5
                  item $ 6-13
                  test $ 14-20
                  group 21
                  a
                  b
                  c
                  avar
                  abcov
                  /
                  bvar
                  accov
                  bccov
                  cvar;

                  asd=sqrt(avar);
                  bsd=sqrt(bvar);
                  csd=sqrt(cvar);
data iplink;
      INFILE IO (dissertationpwr.lin) missover;
      input /
                  variable
                  alpha
                  beta;

proc print data=iplink noobs;
      var alpha beta;
title3 'Linking Coefficients from TCC Method';




/*Creating data sets to call into IML*/
data orig (keep = a b c abcov accov bccov asd bsd csd); set cov;
data theta (keep = theta stderr); set sco;
data ref (keep = a b c abcov accov bccov asd bsd csd); set covref;
data link (keep = alpha beta); set iplink;

proc iml;
**Creates a matrix with original focus group item parameter
information**;
use orig;
read all into matorig;
```

```
**Creates a matrix with original focus group theta values and standard
error**;
use theta;
read all into mattheta;
**Creates a matrix with original reference group item parameter
information**;
use ref;
read all into matref;
**Creates a matrix with alpha and beta linking coefficients**;
use link;
read all into matlink;

**Values/Matrices to be used later**;
seeds={123456 234567 345678 456789 567890 678901};
items=nrow(matorig);
n=nrow(mattheta);
reps=1000;
ncdifmat=repeat(0,reps,items);
dtfmat=repeat(0,reps,1);
fnor=repeat(0,3,items);
rnor=repeat(0,3,items);
fnort=repeat(0,3,items);
rnort=repeat(0,3,items);
foc=repeat(0,3,items);
ref=repeat(0,3,items);
pfoc=repeat(0,n,items);
pref=repeat(0,n,items);
T=repeat(0,3,3);
r=repeat(1,3,3);


/***** POWER DECLARATIONS – K. D. Wright*****/
pwr_ncdifmat=repeat(0,reps,items);
pwr_ref=repeat(0,3,items);
pwr_pref=repeat(0,n,items);
pwr_rnort=repeat(0,3,items);
pwr_T=repeat(0,3,3);
pwr_r=repeat(1,3,3);
```

```
/****** EFFECT SIZES – K. D. Wright*********/
OnePLB=  {.000, .001, .001, .002, .003, .003, .002, .002, .001, .001, .000};

OnePLC=  {.001, .002, .002, .005, .007, .007, .005, .005, .002, .002, .001};

         /* -3, -2, -1.5  -1  -.5   0   .5    1    1.5   2   3.0 */
TwoPLB=  {.005 .006 .007 .009 .009 .009 .008 .007 .007 .006 .003, /*  .30 */
          .001 .003 .005 .007 .008 .008 .007 .006 .004 .003 .001, /*  .50 */
           .001 .002 .003 .005 .006 .007 .006 .004 .003 .001 .000, /*  .75 */
           .000 .001 .002 .004 .006 .006 .006 .003 .002 .001 .000, /*  .95 */
           .000 .001 .001 .002 .003 .003 .003 .003 .001 .000 .000, /* 1.25 */
           .000 .001 .001 .002 .003 .004 .003 .002 .001 .000 .000, /* 1.50 */
           .000 .000 .001 .001 .002 .004 .002 .001 .000 .000 .000, /* 1.75 */
           .000 .000 .001 .001 .002 .002 .002 .001 .000 .000 .000};/* 2.00 */

TwoPLC=  {.011 .014 .016 .020 .020 .020 .018 .016 .016 .014 .007, /*  .30 */
          .002 .007 .011 .016 .018 .018 .016 .014 .009 .007 .002, /*  .50 */
           .002 .005 .007 .011 .014 .016 .014 .009 .007 .002 .001, /*  .75 */
           .000 .002 .005 .009 .014 .014 .014 .007 .005 .002 .000, /*  .95 */
           .000 .002 .002 .005 .007 .007 .007 .007 .002 .001 .000, /* 1.25 */
           .000 .002 .002 .005 .007 .009 .007 .005 .002 .000 .000, /* 1.50 */
           .000 .001 .002 .002 .005 .009 .005 .002 .001 .000 .000, /* 1.75 */
           .000 .001 .002 .002 .005 .005 .005 .002 .001 .000 .000};/* 2.00 */

           /* -3, -2, -1.5  -1  -.5    0   .5   1    1.5    2  3.0  */
ThreePLB=  {.003 .007 .008 .009 .011 .010 .010 .010 .010 .010 .999, /*  .30 */
            .001 .003 .006 .006 .010 .010 .014 .014 .014 .999 .999, /*  .50 */
              .000 .001 .002 .004 .008 .009 .014 .014 .014 .999 .999, /*  .75 */
              .000 .001 .002 .003 .007 .011 .013 .013 .999 .999 .999, /*  .95 */
              .000 .001 .001 .003 .006 .009 .016 .016 .999 .999 .999, /* 1.25 */
              .000 .001 .001 .002 .005 .008 .017 .017 .999 .999 .999, /* 1.50 */
              .000 .000 .001 .001 .004 .009 .018 .999 .999 .999 .999, /* 1.75 */
              .000 .000 .001 .001 .004 .009 .019 .999 .999 .999 .999};/* 2.00 */




ThreePLC=  {.007 .016 .018 .020 .025 .023 .023 .023 .023 .023 .999, /*  .30 */
            .002 .007 .014 .014 .023 .023 .032 .032 .032 .999 .999, /*  .50 */
              .001 .002 .005 .009 .018 .020 .032 .032 .032 .999 .999, /*  .75 */
              .000 .002 .005 .007 .016 .025 .029 .029 .999 .999 .999, /*  .95 */
              .000 .002 .002 .007 .014 .020 .036 .036 .999 .999 .999, /* 1.25 */
              .000 .002 .002 .005 .011 .018 .038 .038 .999 .999 .999, /* 1.50 */
              .000 .000 .002 .002 .009 .020 .041 .999 .999 .999 .999, /* 1.75 */
              .000 .000 .002 .002 .009 .020 .043 .999 .999 .999 .999};/* 2.00 */

              /* The below are the actual values simulated to produce the effect
sizes */
              /* Cutoffs are programmed as associated with the BParam */
          /* -3,  -2,    -1.5   -1    -.5     0    .5    1    1.5   2  3.0 */
BParam=    {-2.5, -1.75, -1.25, -.75, -0.25, .25, .75, 1.25, 1.75, 2.5, 3};


              /* The below are the actual values simulated to produce the effect
sizes */
              /* Cutoffs are programmed as associated with the BParam */
          /*.30, .50,  .75,  .95, 1.25, 1.50,  1.75, 2.00) */
AParam=    {.40, .625, .85, 1.10, 1.38, 1.625, 1.88, 2.00};
```

```
** 1 Parameter Model*************************************************;
if (matorig[:,9]=0 & matorig[:,7]=0) then do;
      print '1-PARAMETER MODEL';
      do rep=1 to reps;
            do i=1 to items;
                  do param=1 to 3;
**Creates random normally distributed item parameters for focal and
reference groups**;
                        fnor[param,i]=normal(seeds[1,param]*i+rep);
                        rnor[param,i]=normal(seeds[1,3+param]*i+rep);
                  end;
            end;

            do i=1 to items;
                  do param=1 to 3;
**Changes normal matrices to have same means and standard deviations as
originals**;
**These will be the final simulated item parameters used to calculate
p**;

      foc[param,i]=matorig[i,param]+(matorig[i,6+param]*fnor[param,i]);

ref[param,i]=matorig[i,param]+(matorig[i,6+param]*rnor[param,i]);
                        /* Keith's Dissertation */

      pwr_ref[param,i]=matref[i,param]+(matref[i,6+param]*rnor[param,i]
);
                s end;
            end;

            do theta=1 to n;
                  do i=1 to items;
**Calculates p for each set of item parameters using thetas from
BILOG**;
                      pfoc[theta,i]=foc[3,i]+(1-foc[3,i])*
                              ((EXP(1.7*foc[1,i]*(mattheta[theta,1]-
foc[2,i])))/
                              (1+EXP(1.7*foc[1,i]*(mattheta[theta,1]-
foc[2,i]))));
                      pref[theta,i]=ref[3,i]+(1-ref[3,i])*
                              ((EXP(1.7*ref[1,i]*(mattheta[theta,1]-
ref[2,i])))/
                              (1+EXP(1.7*ref[1,i]*(mattheta[theta,1]-
ref[2,i]))));
                        /* Keith's Dissertation */
                        pwr_pref[theta,i]=pwr_ref[3,i]+(1-
pwr_ref[3,i])*

      ((EXP(1.7*pwr_ref[1,i]*(mattheta[theta,1]-pwr_ref[2,i])))/

      (1+EXP(1.7*pwr_ref[1,i]*(mattheta[theta,1]-pwr_ref[2,i]))));
               end;
            end;

            **Calculates d used in NCDIF equation**;
            d=pfoc-pref;
            /* Keith's Dissertatin */
            pwr_d=pfoc-pwr_pref;

            **Calculates NCDIF**;
            do i = 1 to items;
                  ncdifmat[rep,i]=((sum(d[##,i])-
(((d[+,i])**2)/(n)))/(n))+((d[:,i])**2);
                  /* Keith's Dissertation */
                  pwr_ncdifmat[rep,i]=((sum(pwr_d[##,i])-
(((pwr_d[+,i])**2)/(n)))/(n))+((pwr_d[:,i])**2);
            end;
      end;
end;


**Two Parameter Model and Three Parameter Model with a Fixed c**;
else if (matorig[:,9]=0 & matorig[:,7]<>0) then do;
/*else if (matorig[:,9]<>0 & matorig[:,7]<>0) then do;*/
      print '2 or 3-PARAMETER MODEL';
      do rep=1 to reps;
            do i=1 to items;
      **Fills r then makes T if the r matrix is positive definite**;
                  r[1,2]=matorig[i,4]/(matorig[i,7]*matorig[i,8]);
                  r[2,1]=matorig[i,4]/(matorig[i,7]*matorig[i,8]);
                  r[1,3]=0;
                  r[3,1]=0;
                  r[2,3]=0;
                  r[3,2]=0;
                  T=half(r);


                  /* Keith's Dissertation */
                  pwr_r[1,2]=matref[i,4]/(matref[i,7]*matref[i,8]);
                  pwr_r[2,1]=matref[i,4]/(matref[i,7]*matref[i,8]);
                  pwr_r[1,3]=0;
                  pwr_r[3,1]=0;
                  pwr_r[2,3]=0;
                  pwr_r[3,2]=0;
                  pwr_T=half(pwr_r);

                  do param=1 to 3;
**Creates random normally distributed item parameters for focal and
reference groups**;
                        fnor[param,i]=normal(seeds[1,param]*i+rep);
                        rnor[param,i]=normal(seeds[1,3+param]*i+rep);
                  end;

**Transforms simulated item parameters to have same covariances as
originals**;
                  fnort[,i]=T`*fnor[,i];
                  rnort[,i]=T`*rnor[,i];
                  pwr_rnort[,i]=pwr_T`*rnor[,i];
            end;

            do i=1 to items;
                  do param=1 to 3;
```

```
**Changes normal matrices to have same means and standard deviations as
originals**;
**These will be the final simulated item parameters used to calculate
p**;

foc[param,i]=matorig[i,param]+(matorig[i,6+param]*fnort[param,i]);

ref[param,i]=matorig[i,param]+(matorig[i,6+param]*rnort[param,i]);

                        /* Keith's Dissertation */

      pwr_ref[param,i]=matref[i,param]+(matref[i,6+param]*pwr_rnort[par
am,i]);
                end;
          end;

            do theta=1 to n;
                  do i=1 to items;
**Calculates p for each set of item parameters using thetas from
BILOG**;
                      pfoc[theta,i]=foc[3,i]+(1-foc[3,i])*
                              ((EXP(1.7*foc[1,i]*(mattheta[theta,1]-
foc[2,i])))/
                              (1+EXP(1.7*foc[1,i]*(mattheta[theta,1]-
foc[2,i]))));
                      pref[theta,i]=ref[3,i]+(1-ref[3,i])*
                              ((EXP(1.7*ref[1,i]*(mattheta[theta,1]-
ref[2,i])))/
                              (1+EXP(1.7*ref[1,i]*(mattheta[theta,1]-
ref[2,i]))));

                        /* Keith's Dissertation */
                        pwr_pref[theta,i]=pwr_ref[3,i]+(1-
pwr_ref[3,i])*

      ((EXP(1.7*pwr_ref[1,i]*(mattheta[theta,1]-pwr_ref[2,i])))/

      (1+EXP(1.7*pwr_ref[1,i]*(mattheta[theta,1]-pwr_ref[2,i]))));
               end;
            end;

            **Calculates d used in NCDIF equation**;
            d=pfoc-pref;

            /* Keith's Dissertatin */
            pwr_d=pfoc-pwr_pref;


            **Calculates NCDIF**;
            do i = 1 to items;
                  ncdifmat[rep,i]=((sum(d[##,i])-
(((d[+,i])**2)/(n)))/(n))+((d[:,i])**2);

                  /* Keith's Dissertation */
                  pwr_ncdifmat[rep,i]=((sum(pwr_d[##,i])-
(((pwr_d[+,i])**2)/(n)))/(n))+((pwr_d[:,i])**2);
            end;
      end;
end;


**********************************************************************;
**********************************************************************;
**Three Parameter Model without Fixed c**;
else if (matorig[:,9]<>0 & matorig[:,7]<>0) then do;
/*else if (matorig[:,9]=0 & matorig[:,7]<>0) then do;*/
      problem_c=repeat('         ',1,items);
      print '3-PARAMETER MODEL';
      do rep=1 to reps;
            do i=1 to items;
      **Fills r then makes T if the r matrix is positive definite**;
                  r[1,2]=matorig[i,4]/(matorig[i,7]*matorig[i,8]);
                  r[2,1]=matorig[i,4]/(matorig[i,7]*matorig[i,8]);
                  r[1,3]=matorig[i,5]/(matorig[i,7]*matorig[i,9]);
                  r[3,1]=matorig[i,5]/(matorig[i,7]*matorig[i,9]);
                  r[2,3]=matorig[i,6]/(matorig[i,8]*matorig[i,9]);
                  r[3,2]=matorig[i,6]/(matorig[i,8]*matorig[i,9]);

                  if det(r)>0 then do;
                        T=half(r);
                  end;

                  if det(r)<=0 then do;
                        problem_c[1,i]='x        ';

      r[1,2]=matorig[i,4]/(matorig[i,7]*matorig[i,8]);

      r[2,1]=matorig[i,4]/(matorig[i,7]*matorig[i,8]);
                        r[1,3]=0;
                        r[3,1]=0;
                        r[2,3]=0;
                        r[3,2]=0;
                        T=half(r);
                  end;

                  /* Keith's Dissertation */
                  pwr_r[1,2]=matref[i,4]/(matref[i,7]*matref[i,8]);
                  pwr_r[2,1]=matref[i,4]/(matref[i,7]*matref[i,8]);
                  pwr_r[1,3]=matref[i,5]/(matref[i,7]*matref[i,9]);
                  pwr_r[3,1]=matref[i,5]/(matref[i,7]*matref[i,9]);
                  pwr_r[2,3]=matref[i,6]/(matref[i,8]*matref[i,9]);
                  pwr_r[3,2]=matref[i,6]/(matref[i,8]*matref[i,9]);
                  if det(pwr_r)>0 then do;
                        pwr_T=half(pwr_r);
                  end;
                  if det(pwr_r)<=0 then do;
                        problem_c[1,i]='x        ';

      pwr_r[1,2]=matref[i,4]/(matref[i,7]*matref[i,8]);

      pwr_r[2,1]=matref[i,4]/(matref[i,7]*matref[i,8]);
                        pwr_r[1,3]=0;
                        pwr_r[3,1]=0;
                        pwr_r[2,3]=0;
                        pwr_r[3,2]=0;
                        pwr_T=half(pwr_r);
                  end;


                  do param=1 to 3;
**Creates random normally distributed item parameters for focal and
reference groups**;
                        fnor[param,i]=normal(seeds[1,param]*i+rep);
                        rnor[param,i]=normal(seeds[1,3+param]*i+rep);
                  end;

**Transforms simulated item parameters to have same covariances as
originals**;
                  fnort[,i]=T`*fnor[,i];
                  rnort[,i]=T`*rnor[,i];
                  pwr_rnort[,i]=pwr_T`*rnor[,i];
            end;


            do i=1 to items;
                  do param=1 to 3;
**Changes normal matrices to have same means and standard deviations as
originals**;
**These will be the final simulated item parameters used to calculate
p**;

foc[param,i]=matorig[i,param]+(matorig[i,6+param]*fnort[param,i]);

ref[param,i]=matorig[i,param]+(matorig[i,6+param]*rnort[param,i]);
                        /* Keith's Dissertation */

      pwr_ref[param,i]=matref[i,param]+(matref[i,6+param]*pwr_rnort[par
am,i]);
                end;
          end;


            do theta=1 to n;
                  do i=1 to items;
**Calculates p for each set of item parameters using thetas from
BILOG**;
                      pfoc[theta,i]=foc[3,i]+(1-foc[3,i])*
                              ((EXP(1.7*foc[1,i]*(mattheta[theta,1]-
foc[2,i])))/
                              (1+EXP(1.7*foc[1,i]*(mattheta[theta,1]-
foc[2,i]))));
                      pref[theta,i]=ref[3,i]+(1-ref[3,i])*
                              ((EXP(1.7*ref[1,i]*(mattheta[theta,1]-
ref[2,i])))/
                              (1+EXP(1.7*ref[1,i]*(mattheta[theta,1]-
ref[2,i]))));

                        /* Keith's Dissertation */
                        pwr_pref[theta,i]=pwr_ref[3,i]+(1-
pwr_ref[3,i])*
```

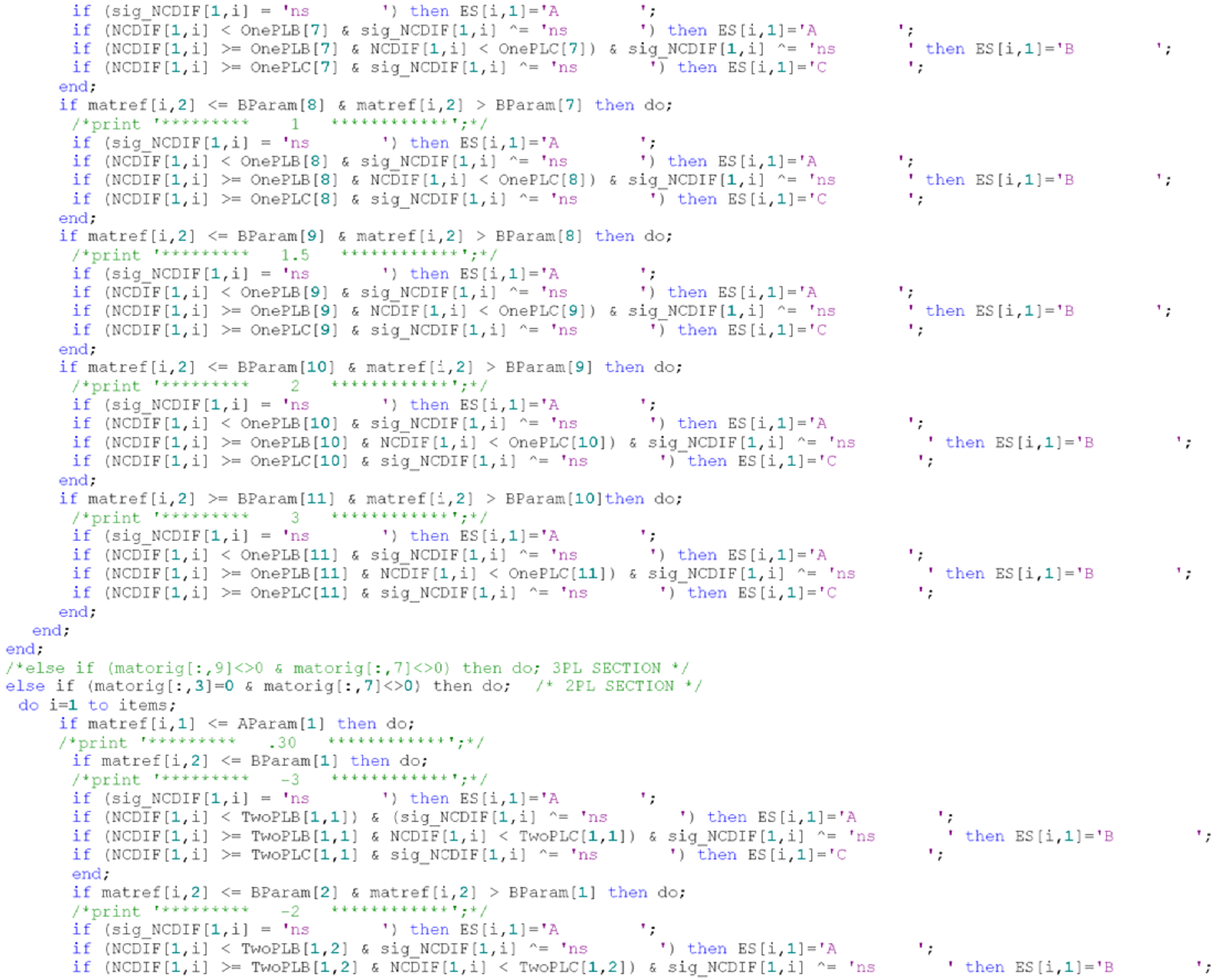
```
      ((EXP(1.7*pwr_ref[1,i]*(mattheta[theta,1]-pwr_ref[2,i])))/

      (1+EXP(1.7*pwr_ref[1,i]*(mattheta[theta,1]-pwr_ref[2,i]))));
               end;
            end;

            **Calculates d used in NCDIF equation**;
            d=pfoc-pref;

            /* Keith's Dissertatin */
            pwr_d=pfoc-pwr_pref;

            **Calculates NCDIF**;
            do i = 1 to items;
                  ncdifmat[rep,i]=((sum(d[##,i])-
(((d[+,i])**2)/(n)))/(n))+((d[:,i])**2);

                    /* Keith's Dissertation */
                    pwr_ncdifmat[rep,i]=((sum(pwr_d[##,i])-
(((pwr_d[+,i])**2)/(n)))/(n))+((pwr_d[:,i])**2);
            end;
      end;

      title3 ' ';
      print 'Columns marked with x are items with simulated c-
parameters not related to a and b' problem_c;

end;

**********************************************************************;
**********************************************************************;




**Creates an itemrank matrix with ncdif values for each item in
ascending order**;
itemrank=repeat(0,reps,items);
do i=1 to items;
      k=repeat(0,reps,1);
      k=ncdifmat[,i];
      f=k;
      k[rank(k),]=f; ;
      itemrank[,i]=k;
end;

/* Keith's Dissertation */
pwr_itemrank=repeat(0,reps,items);
do i=1 to items;
      k=repeat(0,reps,1);
      k=pwr_ncdifmat[,i];
      f=k;
      k[rank(k),]=f; ;
      pwr_itemrank[,i]=k;
end;
```

```
title3 ' ';
cutoffnames={'Cutoff .10', 'Cutoff .05', 'Cutoff .01', 'Cutoff .001'};
NCDIF_ITEM_CUTOFFS=repeat(0,4,items);
NCDIF_ITEM_CUTOFFS[1,]=itemrank[ceil(.90*reps),];
NCDIF_ITEM_CUTOFFS[2,]=itemrank[ceil(.95*reps),];
NCDIF_ITEM_CUTOFFS[3,]=itemrank[ceil(.99*reps),];
NCDIF_ITEM_CUTOFFS[4,]=itemrank[ceil(.999*reps),];
print NCDIF_ITEM_CUTOFFS [r=cutoffnames];



**Creates an empty column matrix that will be filled with NCDIF & POWER
values**;
ncdifcol=repeat(0,reps*items,1);
pwr_ncdif=repeat(0,reps*items,1);
EMPIRICAL_POWER=repeat(0,items,1);
NCDIF95=repeat(0,items,1);
PVALUE2=repeat(0,items,1);
item_num=repeat(0,items,1);


**Reads NCDIF values 1 column**;
do i=1 to items;
      do r=1 to reps;
            ncdifcol[r+(i-1)*reps,1]=ncdifmat[r,i];
      end;
      item_num[i,1]=i;
end;


/* Keith's Dissertation */
do i=1 to items;
      power = 0;
      do r=1 to reps;
            pwr_ncdif[r+(i-1)*reps,1]=pwr_ncdifmat[r,i];
            if pwr_ncdifmat[r,i] >= NCDIF_ITEM_CUTOFFS[2,i] then
power=power+1;
      end;
      EMPIRICAL_POWER[i,1]=(power/1000);
      NCDIF95[i,1]=NCDIF_ITEM_CUTOFFS[2,i];
end;




***********************************************************************
**Puts the reference group on the same scale as the focal group**;
newref=repeat(0,items,3);
do i=1 to items;
      newref[i,1]=(1/matlink[1,1])*matref[i,1];
      newref[i,2]=matlink[1,1]*matref[i,2]+matlink[1,2];
      newref[i,3]=matref[i,3];
end;
```

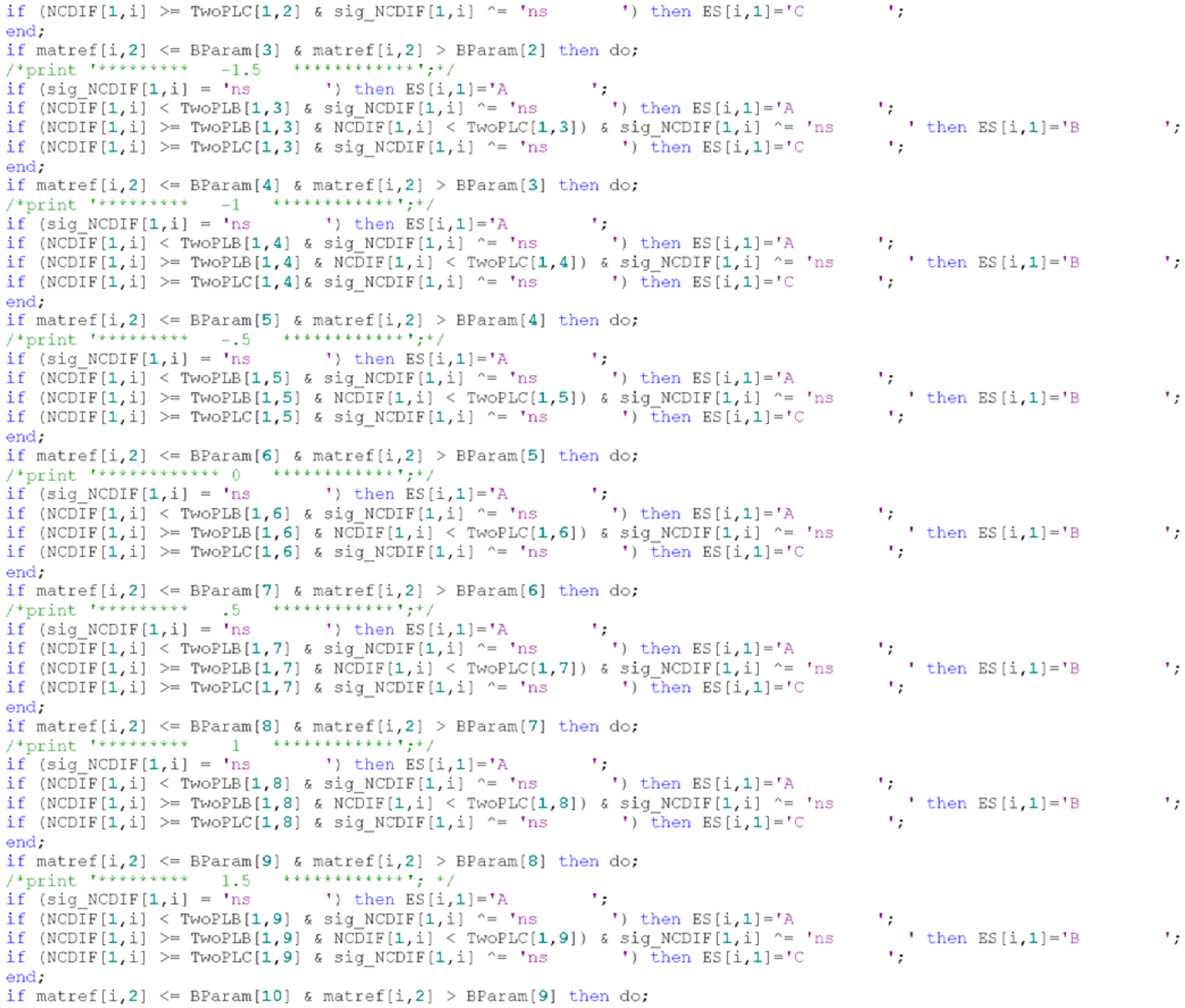
```
**Calculates p for the focal group and linked reference group**;
pf=repeat(0,n,items);
pr=repeat(0,n,items);
NCDIF=repeat(0,1,items);
NCDIF2=repeat(0,items,1);
      do theta=1 to n;
            do i=1 to items;
**Calculates p for each set of item parameters using thetas from
BILOG**;
                pf[theta,i]=matorig[i,3]+(1-matorig[i,3])*
                        ((EXP(1.7*matorig[i,1]*(mattheta[theta,1]-
matorig[i,2])))/
                        (1+EXP(1.7*matorig[i,1]*(mattheta[theta,1]-
matorig[i,2]))));

                pr[theta,i]=newref[i,3]+(1-newref[i,3])*
                        ((EXP(1.7*newref[i,1]*(mattheta[theta,1]-
newref[i,2])))/
                        (1+EXP(1.7*newref[i,1]*(mattheta[theta,1]-
newref[i,2]))));
         end;
      end;

**Calculates d used in NCDIF equation**;
d=pf-pr;

**Calculates NCDIF**;
do i = 1 to items;
    NCDIF[1,i]=((sum(d[##,i])-(((d[+,i])**2)/(n)))/(n))+((d[:,i])**2);
      NCDIF2[i,1]=((sum(d[##,i])-
(((d[+,i])**2)/(n)))/(n))+((d[:,i])**2);
end;



**Flags significant NCDIF**;
sig_NCDIF=repeat('         ',1,items);
do i=1 to items;
if NCDIF[1,i]>=NCDIF_ITEM_CUTOFFS[1,i] then sig_NCDIF[1,i]='*        ';
if NCDIF[1,i]>=NCDIF_ITEM_CUTOFFS[2,i] then sig_NCDIF[1,i]='**       ';
if NCDIF[1,i]>=NCDIF_ITEM_CUTOFFS[3,i] then sig_NCDIF[1,i]='***      ';
if NCDIF[1,i]>=NCDIF_ITEM_CUTOFFS[4,i] then sig_NCDIF[1,i]='****     ';
if NCDIF[1,i]<NCDIF_ITEM_CUTOFFS[1,i] then sig_NCDIF[1,i]='ns       ';
end;


/* ONLY FOR OUTPUT - Modified by K. D. Wright*/
sig_NCDIF2=repeat('   ',items,1);
do i=1 to items;
if NCDIF[1,i]>=NCDIF_ITEM_CUTOFFS[1,i] then sig_NCDIF2[i,1]='.10';
if NCDIF[1,i]>=NCDIF_ITEM_CUTOFFS[2,i] then sig_NCDIF2[i,1]='.05';
if NCDIF[1,i]>=NCDIF_ITEM_CUTOFFS[3,i] then sig_NCDIF2[i,1]='.01';
if NCDIF[1,i]>=NCDIF_ITEM_CUTOFFS[4,i] then sig_NCDIF2[i,1]='.001';
if NCDIF[1,i]<NCDIF_ITEM_CUTOFFS[1,i] then sig_NCDIF2[i,1]='ns';
end;
```

```
/* Keith's Dissertation P-VALUE CODE */
do i=1 to items;
            pvalue=0;
            do r=1 to reps;
                  if NCDIF[1,i]>=ncdifmat[r,i]then pvalue=pvalue+1;
            end;
            PVALUE2[i,1]=1-(pvalue/1000);
end;
```

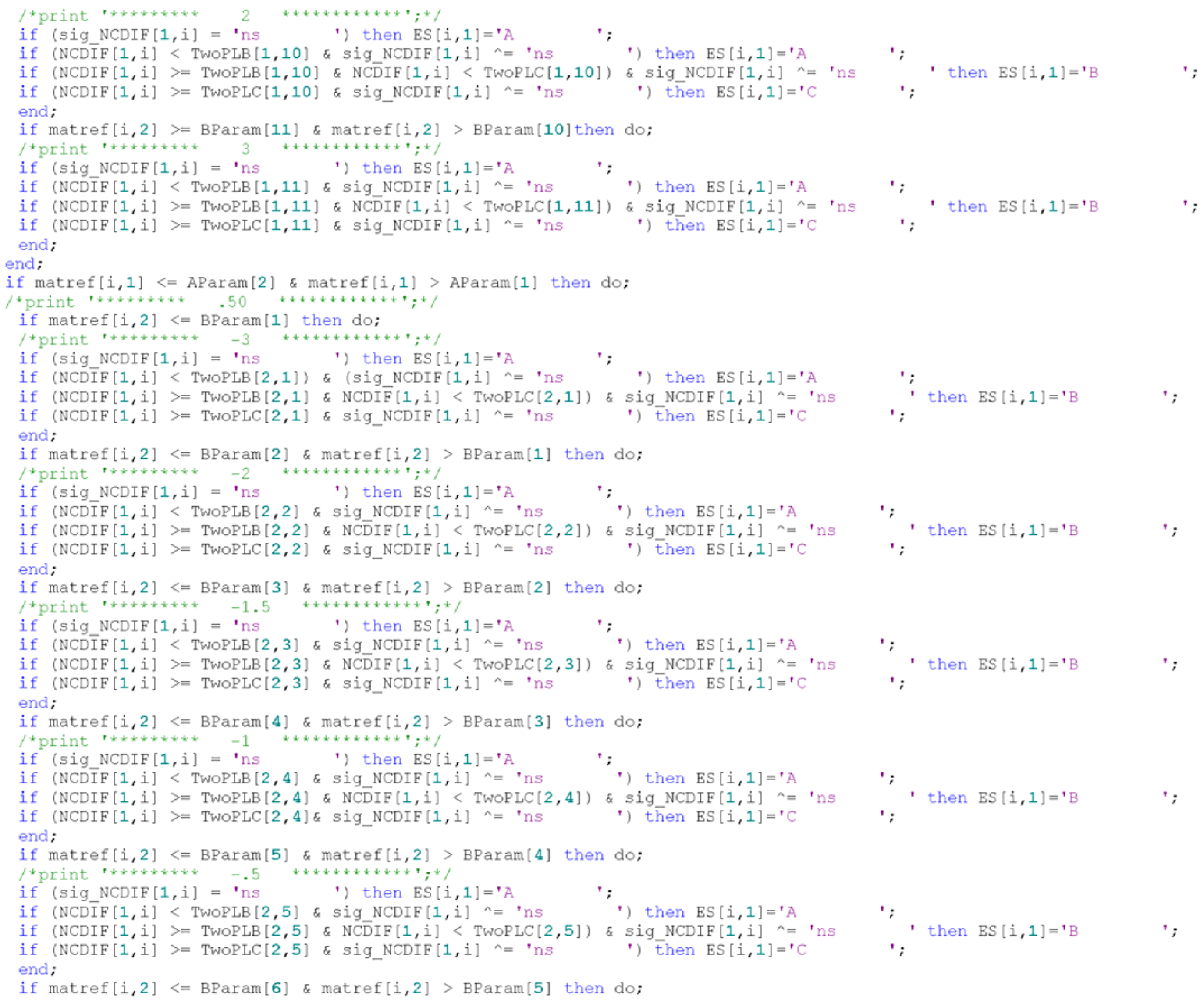
```
/**********************   EFFECT SIZE SECTION OF IPR - K. D. Wright Dissertation **********************/
/* Note - To conserve space only the 1PL and 2PL code is presented.  The 3PL code is similar to the 2PL **/
/* code with the exception of the the first line associated with the "if statement"

/**EFFECT SIZE DECLARATION**/
ES=repeat('         ',items,1);
if (matorig[:,9]=0 & matorig[:,7]=0) then do; /* 1PL SECTION */
  do i=1 to items;
     if matref[i,2] <= BParam[1] then do;
      /*print '*********   -3   ************';*/
      if (sig_NCDIF[1,i] = 'ns        ') then ES[i,1]='A        ';
      if (NCDIF[1,i] < OnePLB[1]) & (sig_NCDIF[1,i] ^= 'ns        ') then ES[i,1]='A        ';
      if (NCDIF[1,i] >= OnePLB[1] & NCDIF[1,i] < OnePLC[1]) & sig_NCDIF[1,i] ^= 'ns        ' then ES[i,1]='B        ';
      if (NCDIF[1,i] >= OnePLC[1] & sig_NCDIF[1,i] ^= 'ns        ') then ES[i,1]='C        ';
     end;
     if matref[i,2] <= BParam[2] & matref[i,2] > BParam[1] then do;
      /*print '*********   -2   ************';*/
      if (sig_NCDIF[1,i] = 'ns        ') then ES[i,1]='A        ';
      if (NCDIF[1,i] < OnePLB[2] & sig_NCDIF[1,i] ^= 'ns        ') then ES[i,1]='A        ';
      if (NCDIF[1,i] >= OnePLB[2] & NCDIF[1,i] < OnePLC[2]) & sig_NCDIF[1,i] ^= 'ns        ' then ES[i,1]='B        ';
      if (NCDIF[1,i] >= OnePLC[2] & sig_NCDIF[1,i] ^= 'ns        ') then ES[i,1]='C        ';
     end;
     if matref[i,2] <= BParam[3] & matref[i,2] > BParam[2] then do;
      /*print '*********   -1.5   ************';*/
      if (sig_NCDIF[1,i] = 'ns        ') then ES[i,1]='A        ';
      if (NCDIF[1,i] < OnePLB[3] & sig_NCDIF[1,i] ^= 'ns        ') then ES[i,1]='A        ';
      if (NCDIF[1,i] >= OnePLB[3] & NCDIF[1,i] < OnePLC[3]) & sig_NCDIF[1,i] ^= 'ns        ' then ES[i,1]='B        ';
      if (NCDIF[1,i] >= OnePLC[3] & sig_NCDIF[1,i] ^= 'ns        ') then ES[i,1]='C        ';
     end;
     if matref[i,2] <= BParam[4] & matref[i,2] > BParam[3] then do;
      /*print '*********   -1   ************';*/
      if (sig_NCDIF[1,i] = 'ns        ') then ES[i,1]='A        ';
      if (NCDIF[1,i] < OnePLB[4] & sig_NCDIF[1,i] ^= 'ns        ') then ES[i,1]='A        ';
      if (NCDIF[1,i] >= OnePLB[4] & NCDIF[1,i] < OnePLC[4]) & sig_NCDIF[1,i] ^= 'ns        ' then ES[i,1]='B        ';
      if (NCDIF[1,i] >= OnePLC[4]& sig_NCDIF[1,i] ^= 'ns        ') then ES[i,1]='C        ';
     end;
     if matref[i,2] <= BParam[5] & matref[i,2] > BParam[4] then do;
      /*print '*********   -.5   ************';*/
      if (sig_NCDIF[1,i] = 'ns        ') then ES[i,1]='A        ';
      if (NCDIF[1,i] < OnePLB[5] & sig_NCDIF[1,i] ^= 'ns        ') then ES[i,1]='A        ';
      if (NCDIF[1,i] >= OnePLB[5] & NCDIF[1,i] < OnePLC[5]) & sig_NCDIF[1,i] ^= 'ns        ' then ES[i,1]='B        ';
      if (NCDIF[1,i] >= OnePLC[5] & sig_NCDIF[1,i] ^= 'ns        ') then ES[i,1]='C        ';
     end;
     if matref[i,2] <= BParam[6] & matref[i,2] > BParam[5] then do;
      /*print '************ 0   ************';*/
      if (sig_NCDIF[1,i] = 'ns        ') then ES[i,1]='A        ';
      if (NCDIF[1,i] < OnePLB[6] & sig_NCDIF[1,i] ^= 'ns        ') then ES[i,1]='A        ';
      if (NCDIF[1,i] >= OnePLB[6] & NCDIF[1,i] < OnePLC[6]) & sig_NCDIF[1,i] ^= 'ns        ' then ES[i,1]='B        ';
      if (NCDIF[1,i] >= OnePLC[6] & sig_NCDIF[1,i] ^= 'ns        ') then ES[i,1]='C        ';
     end;
     if matref[i,2] <= BParam[7] & matref[i,2] > BParam[6] then do;
      /*print '*********   .5   ************';*/
```



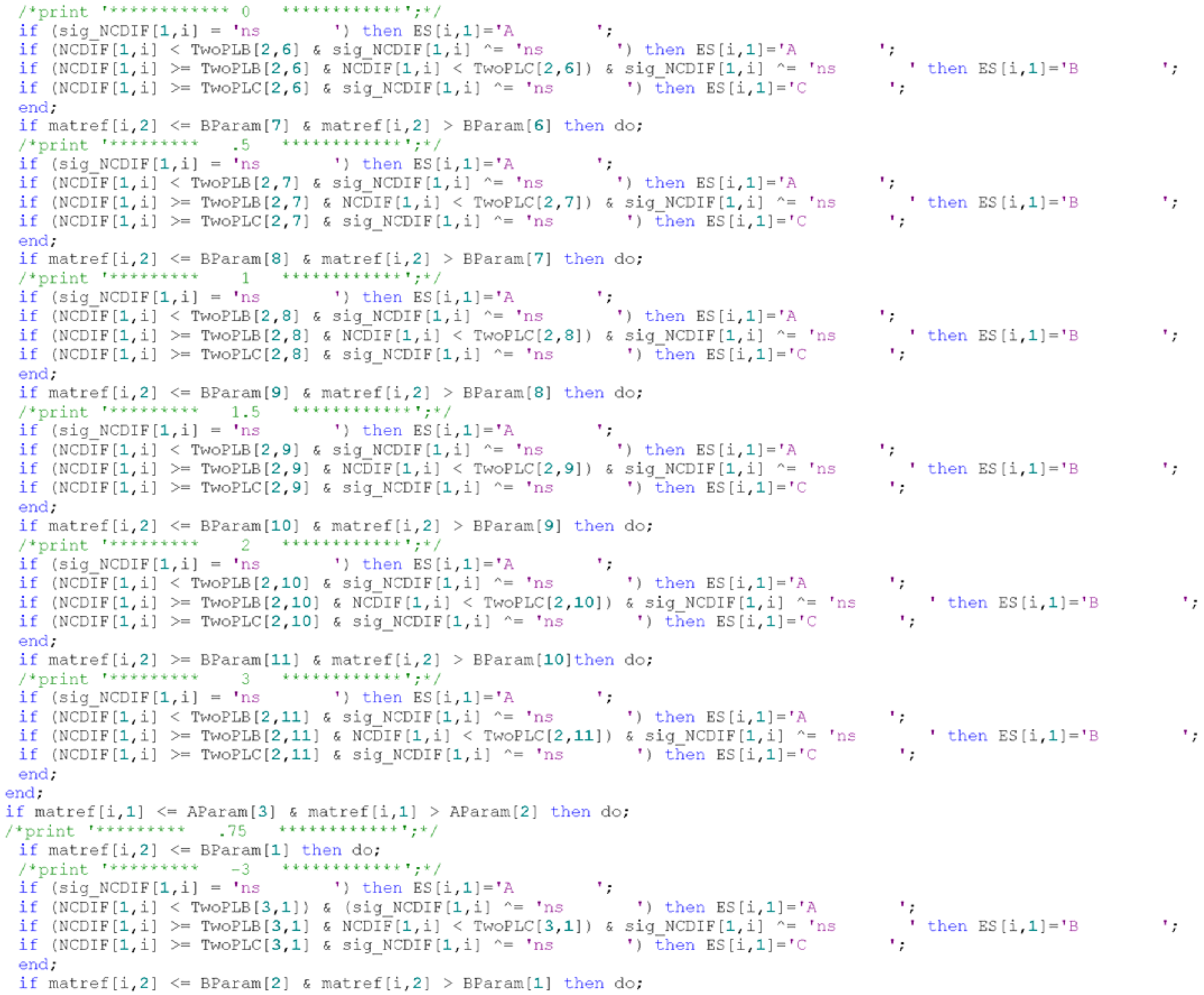
```
      if (sig_NCDIF[1,i] = 'ns        ') then ES[i,1]='A         ';
      if (NCDIF[1,i] < OnePLB[7] & sig_NCDIF[1,i] ^= 'ns        ') then ES[i,1]='A         ';
      if (NCDIF[1,i] >= OnePLB[7] & NCDIF[1,i] < OnePLC[7]) & sig_NCDIF[1,i] ^= 'ns        ' then ES[i,1]='B         ';
      if (NCDIF[1,i] >= OnePLC[7] & sig_NCDIF[1,i] ^= 'ns        ') then ES[i,1]='C         ';
     end;
     if matref[i,2] <= BParam[8] & matref[i,2] > BParam[7] then do;
      /*print '*********    1    *************';*/
      if (sig_NCDIF[1,i] = 'ns        ') then ES[i,1]='A         ';
      if (NCDIF[1,i] < OnePLB[8] & sig_NCDIF[1,i] ^= 'ns        ') then ES[i,1]='A         ';
      if (NCDIF[1,i] >= OnePLB[8] & NCDIF[1,i] < OnePLC[8]) & sig_NCDIF[1,i] ^= 'ns        ' then ES[i,1]='B         ';
      if (NCDIF[1,i] >= OnePLC[8] & sig_NCDIF[1,i] ^= 'ns        ') then ES[i,1]='C         ';
     end;
     if matref[i,2] <= BParam[9] & matref[i,2] > BParam[8] then do;
      /*print '*********   1.5   *************';*/
      if (sig_NCDIF[1,i] = 'ns        ') then ES[i,1]='A         ';
      if (NCDIF[1,i] < OnePLB[9] & sig_NCDIF[1,i] ^= 'ns        ') then ES[i,1]='A         ';
      if (NCDIF[1,i] >= OnePLB[9] & NCDIF[1,i] < OnePLC[9]) & sig_NCDIF[1,i] ^= 'ns        ' then ES[i,1]='B         ';
      if (NCDIF[1,i] >= OnePLC[9] & sig_NCDIF[1,i] ^= 'ns        ') then ES[i,1]='C         ';
     end;
     if matref[i,2] <= BParam[10] & matref[i,2] > BParam[9] then do;
      /*print '*********    2    *************';*/
      if (sig_NCDIF[1,i] = 'ns        ') then ES[i,1]='A         ';
      if (NCDIF[1,i] < OnePLB[10] & sig_NCDIF[1,i] ^= 'ns        ') then ES[i,1]='A         ';
      if (NCDIF[1,i] >= OnePLB[10] & NCDIF[1,i] < OnePLC[10]) & sig_NCDIF[1,i] ^= 'ns        ' then ES[i,1]='B         ';
      if (NCDIF[1,i] >= OnePLC[10] & sig_NCDIF[1,i] ^= 'ns        ') then ES[i,1]='C         ';
     end;
     if matref[i,2] >= BParam[11] & matref[i,2] > BParam[10]then do;
      /*print '*********    3    *************';*/
      if (sig_NCDIF[1,i] = 'ns        ') then ES[i,1]='A         ';
      if (NCDIF[1,i] < OnePLB[11] & sig_NCDIF[1,i] ^= 'ns        ') then ES[i,1]='A         ';
      if (NCDIF[1,i] >= OnePLB[11] & NCDIF[1,i] < OnePLC[11]) & sig_NCDIF[1,i] ^= 'ns        ' then ES[i,1]='B         ';
      if (NCDIF[1,i] >= OnePLC[11] & sig_NCDIF[1,i] ^= 'ns        ') then ES[i,1]='C         ';
     end;
   end;
end;
/*else if (matorig[:,9]<>0 & matorig[:,7]<>0) then do; 3PL SECTION */
else if (matorig[:,3]=0 & matorig[:,7]<>0) then do;  /* 2PL SECTION */
 do i=1 to items;
     if matref[i,1] <= AParam[1] then do;
     /*print '*********   .30   *************';*/
      if matref[i,2] <= BParam[1] then do;
      /*print '*********   -3   *************';*/
      if (sig_NCDIF[1,i] = 'ns        ') then ES[i,1]='A         ';
      if (NCDIF[1,i] < TwoPLB[1,1]) & (sig_NCDIF[1,i] ^= 'ns        ') then ES[i,1]='A         ';
      if (NCDIF[1,i] >= TwoPLB[1,1] & NCDIF[1,i] < TwoPLC[1,1]) & sig_NCDIF[1,i] ^= 'ns        ' then ES[i,1]='B         ';
      if (NCDIF[1,i] >= TwoPLC[1,1] & sig_NCDIF[1,i] ^= 'ns        ') then ES[i,1]='C         ';
      end;
      if matref[i,2] <= BParam[2] & matref[i,2] > BParam[1] then do;
      /*print '*********   -2   *************';*/
      if (sig_NCDIF[1,i] = 'ns        ') then ES[i,1]='A         ';
      if (NCDIF[1,i] < TwoPLB[1,2] & sig_NCDIF[1,i] ^= 'ns        ') then ES[i,1]='A         ';
      if (NCDIF[1,i] >= TwoPLB[1,2] & NCDIF[1,i] < TwoPLC[1,2]) & sig_NCDIF[1,i] ^= 'ns        ' then ES[i,1]='B         ';
```



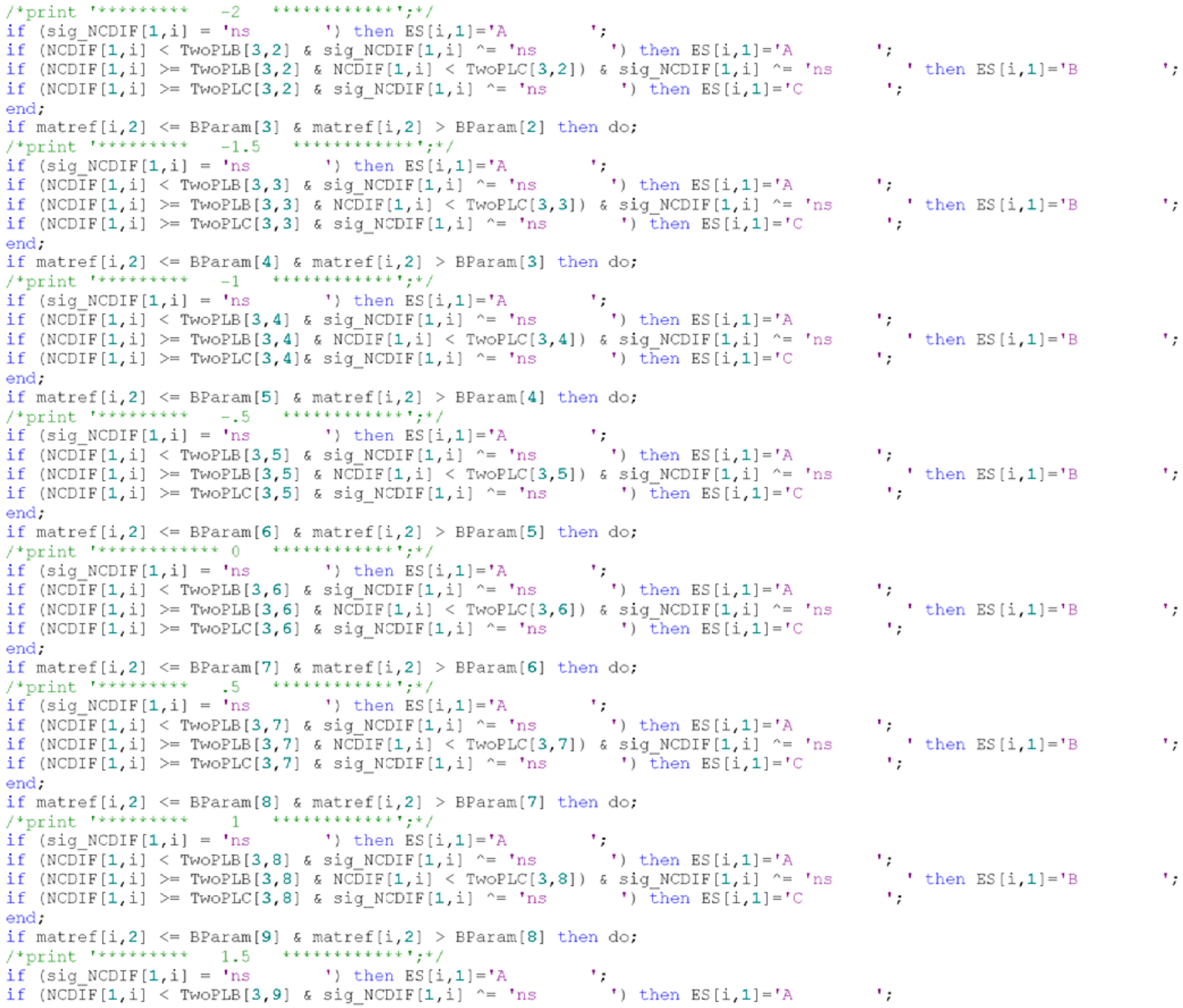
```
if (NCDIF[1,i] >= TwoPLC[1,2] & sig_NCDIF[1,i] ^= 'ns        ') then ES[i,1]='C        ';
end;
if matref[i,2] <= BParam[3] & matref[i,2] > BParam[2] then do;
/*print '*********   -1.5   ************';*/
if (sig_NCDIF[1,i] = 'ns        ') then ES[i,1]='A        ';
if (NCDIF[1,i] < TwoPLB[1,3] & sig_NCDIF[1,i] ^= 'ns        ') then ES[i,1]='A        ';
if (NCDIF[1,i] >= TwoPLB[1,3] & NCDIF[1,i] < TwoPLC[1,3]) & sig_NCDIF[1,i] ^= 'ns        ' then ES[i,1]='B        ';
if (NCDIF[1,i] >= TwoPLC[1,3] & sig_NCDIF[1,i] ^= 'ns        ') then ES[i,1]='C        ';
end;
if matref[i,2] <= BParam[4] & matref[i,2] > BParam[3] then do;
/*print '*********   -1   ************';*/
if (sig_NCDIF[1,i] = 'ns        ') then ES[i,1]='A        ';
if (NCDIF[1,i] < TwoPLB[1,4] & sig_NCDIF[1,i] ^= 'ns        ') then ES[i,1]='A        ';
if (NCDIF[1,i] >= TwoPLB[1,4] & NCDIF[1,i] < TwoPLC[1,4]) & sig_NCDIF[1,i] ^= 'ns        ' then ES[i,1]='B        ';
if (NCDIF[1,i] >= TwoPLC[1,4]& sig_NCDIF[1,i] ^= 'ns        ') then ES[i,1]='C        ';
end;
if matref[i,2] <= BParam[5] & matref[i,2] > BParam[4] then do;
/*print '*********   -.5   ************';*/
if (sig_NCDIF[1,i] = 'ns        ') then ES[i,1]='A        ';
if (NCDIF[1,i] < TwoPLB[1,5] & sig_NCDIF[1,i] ^= 'ns        ') then ES[i,1]='A        ';
if (NCDIF[1,i] >= TwoPLB[1,5] & NCDIF[1,i] < TwoPLC[1,5]) & sig_NCDIF[1,i] ^= 'ns        ' then ES[i,1]='B        ';
if (NCDIF[1,i] >= TwoPLC[1,5] & sig_NCDIF[1,i] ^= 'ns        ') then ES[i,1]='C        ';
end;
if matref[i,2] <= BParam[6] & matref[i,2] > BParam[5] then do;
/*print '************ 0   ************';*/
if (sig_NCDIF[1,i] = 'ns        ') then ES[i,1]='A        ';
if (NCDIF[1,i] < TwoPLB[1,6] & sig_NCDIF[1,i] ^= 'ns        ') then ES[i,1]='A        ';
if (NCDIF[1,i] >= TwoPLB[1,6] & NCDIF[1,i] < TwoPLC[1,6]) & sig_NCDIF[1,i] ^= 'ns        ' then ES[i,1]='B        ';
if (NCDIF[1,i] >= TwoPLC[1,6] & sig_NCDIF[1,i] ^= 'ns        ') then ES[i,1]='C        ';
end;
if matref[i,2] <= BParam[7] & matref[i,2] > BParam[6] then do;
/*print '*********   .5   ************';*/
if (sig_NCDIF[1,i] = 'ns        ') then ES[i,1]='A        ';
if (NCDIF[1,i] < TwoPLB[1,7] & sig_NCDIF[1,i] ^= 'ns        ') then ES[i,1]='A        ';
if (NCDIF[1,i] >= TwoPLB[1,7] & NCDIF[1,i] < TwoPLC[1,7]) & sig_NCDIF[1,i] ^= 'ns        ' then ES[i,1]='B        ';
if (NCDIF[1,i] >= TwoPLC[1,7] & sig_NCDIF[1,i] ^= 'ns        ') then ES[i,1]='C        ';
end;
if matref[i,2] <= BParam[8] & matref[i,2] > BParam[7] then do;
/*print '*********    1   ************';*/
if (sig_NCDIF[1,i] = 'ns        ') then ES[i,1]='A        ';
if (NCDIF[1,i] < TwoPLB[1,8] & sig_NCDIF[1,i] ^= 'ns        ') then ES[i,1]='A        ';
if (NCDIF[1,i] >= TwoPLB[1,8] & NCDIF[1,i] < TwoPLC[1,8]) & sig_NCDIF[1,i] ^= 'ns        ' then ES[i,1]='B        ';
if (NCDIF[1,i] >= TwoPLC[1,8] & sig_NCDIF[1,i] ^= 'ns        ') then ES[i,1]='C        ';
end;
if matref[i,2] <= BParam[9] & matref[i,2] > BParam[8] then do;
/*print '*********   1.5   ************'; */
if (sig_NCDIF[1,i] = 'ns        ') then ES[i,1]='A        ';
if (NCDIF[1,i] < TwoPLB[1,9] & sig_NCDIF[1,i] ^= 'ns        ') then ES[i,1]='A        ';
if (NCDIF[1,i] >= TwoPLB[1,9] & NCDIF[1,i] < TwoPLC[1,9]) & sig_NCDIF[1,i] ^= 'ns        ' then ES[i,1]='B        ';
if (NCDIF[1,i] >= TwoPLC[1,9] & sig_NCDIF[1,i] ^= 'ns        ') then ES[i,1]='C        ';
end;
if matref[i,2] <= BParam[10] & matref[i,2] > BParam[9] then do;
```



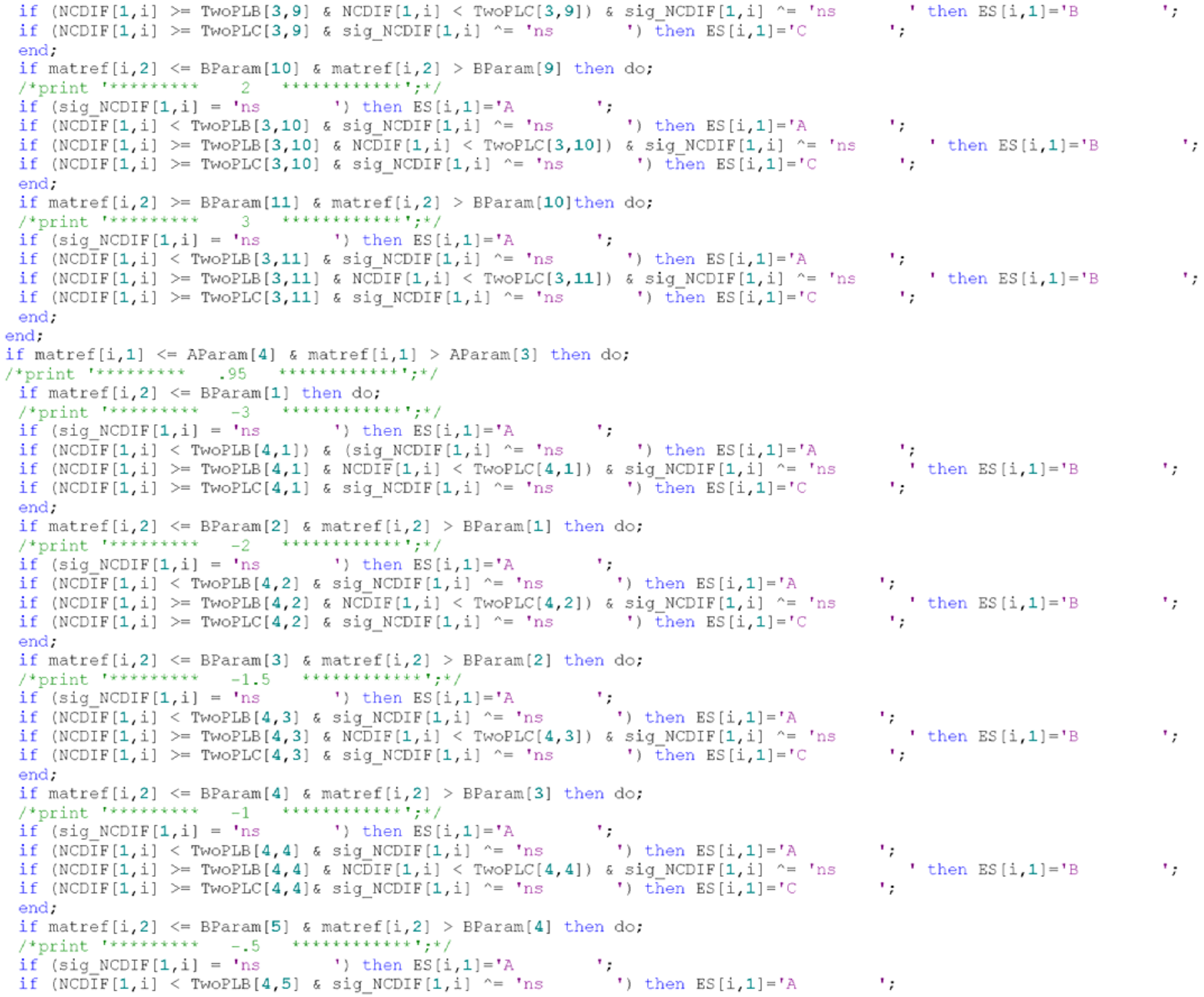
```
 /*print '*********    2   *************';*/
 if (sig_NCDIF[1,i] = 'ns        ') then ES[i,1]='A         ';
 if (NCDIF[1,i] < TwoPLB[1,10] & sig_NCDIF[1,i] ^= 'ns        ') then ES[i,1]='A         ';
 if (NCDIF[1,i] >= TwoPLB[1,10] & NCDIF[1,i] < TwoPLC[1,10]) & sig_NCDIF[1,i] ^= 'ns        ' then ES[i,1]='B         ';
 if (NCDIF[1,i] >= TwoPLC[1,10] & sig_NCDIF[1,i] ^= 'ns        ') then ES[i,1]='C         ';
 end;
 if matref[i,2] >= BParam[11] & matref[i,2] > BParam[10]then do;
 /*print '*********    3   *************';*/
 if (sig_NCDIF[1,i] = 'ns        ') then ES[i,1]='A         ';
 if (NCDIF[1,i] < TwoPLB[1,11] & sig_NCDIF[1,i] ^= 'ns        ') then ES[i,1]='A         ';
 if (NCDIF[1,i] >= TwoPLB[1,11] & NCDIF[1,i] < TwoPLC[1,11]) & sig_NCDIF[1,i] ^= 'ns        ' then ES[i,1]='B         ';
 if (NCDIF[1,i] >= TwoPLC[1,11] & sig_NCDIF[1,i] ^= 'ns        ') then ES[i,1]='C         ';
 end;
end;
if matref[i,1] <= AParam[2] & matref[i,1] > AParam[1] then do;
/*print '*********   .50   *************';*/
 if matref[i,2] <= BParam[1] then do;
 /*print '*********   -3   *************';*/
 if (sig_NCDIF[1,i] = 'ns        ') then ES[i,1]='A         ';
 if (NCDIF[1,i] < TwoPLB[2,1]) & (sig_NCDIF[1,i] ^= 'ns        ') then ES[i,1]='A         ';
 if (NCDIF[1,i] >= TwoPLB[2,1] & NCDIF[1,i] < TwoPLC[2,1]) & sig_NCDIF[1,i] ^= 'ns        ' then ES[i,1]='B         ';
 if (NCDIF[1,i] >= TwoPLC[2,1] & sig_NCDIF[1,i] ^= 'ns        ') then ES[i,1]='C         ';
 end;
 if matref[i,2] <= BParam[2] & matref[i,2] > BParam[1] then do;
 /*print '*********   -2   *************';*/
 if (sig_NCDIF[1,i] = 'ns        ') then ES[i,1]='A         ';
 if (NCDIF[1,i] < TwoPLB[2,2] & sig_NCDIF[1,i] ^= 'ns        ') then ES[i,1]='A         ';
 if (NCDIF[1,i] >= TwoPLB[2,2] & NCDIF[1,i] < TwoPLC[2,2]) & sig_NCDIF[1,i] ^= 'ns        ' then ES[i,1]='B         ';
 if (NCDIF[1,i] >= TwoPLC[2,2] & sig_NCDIF[1,i] ^= 'ns        ') then ES[i,1]='C         ';
 end;
 if matref[i,2] <= BParam[3] & matref[i,2] > BParam[2] then do;
 /*print '*********   -1.5   *************';*/
 if (sig_NCDIF[1,i] = 'ns        ') then ES[i,1]='A         ';
 if (NCDIF[1,i] < TwoPLB[2,3] & sig_NCDIF[1,i] ^= 'ns        ') then ES[i,1]='A         ';
 if (NCDIF[1,i] >= TwoPLB[2,3] & NCDIF[1,i] < TwoPLC[2,3]) & sig_NCDIF[1,i] ^= 'ns        ' then ES[i,1]='B         ';
 if (NCDIF[1,i] >= TwoPLC[2,3] & sig_NCDIF[1,i] ^= 'ns        ') then ES[i,1]='C         ';
 end;
 if matref[i,2] <= BParam[4] & matref[i,2] > BParam[3] then do;
 /*print '*********   -1   *************';*/
 if (sig_NCDIF[1,i] = 'ns        ') then ES[i,1]='A         ';
 if (NCDIF[1,i] < TwoPLB[2,4] & sig_NCDIF[1,i] ^= 'ns        ') then ES[i,1]='A         ';
 if (NCDIF[1,i] >= TwoPLB[2,4] & NCDIF[1,i] < TwoPLC[2,4]) & sig_NCDIF[1,i] ^= 'ns        ' then ES[i,1]='B         ';
 if (NCDIF[1,i] >= TwoPLC[2,4]& sig_NCDIF[1,i] ^= 'ns        ') then ES[i,1]='C         ';
 end;
 if matref[i,2] <= BParam[5] & matref[i,2] > BParam[4] then do;
 /*print '*********   -.5   *************';*/
 if (sig_NCDIF[1,i] = 'ns        ') then ES[i,1]='A         ';
 if (NCDIF[1,i] < TwoPLB[2,5] & sig_NCDIF[1,i] ^= 'ns        ') then ES[i,1]='A         ';
 if (NCDIF[1,i] >= TwoPLB[2,5] & NCDIF[1,i] < TwoPLC[2,5]) & sig_NCDIF[1,i] ^= 'ns        ' then ES[i,1]='B         ';
 if (NCDIF[1,i] >= TwoPLC[2,5] & sig_NCDIF[1,i] ^= 'ns        ') then ES[i,1]='C         ';
 end;
 if matref[i,2] <= BParam[6] & matref[i,2] > BParam[5] then do;
```



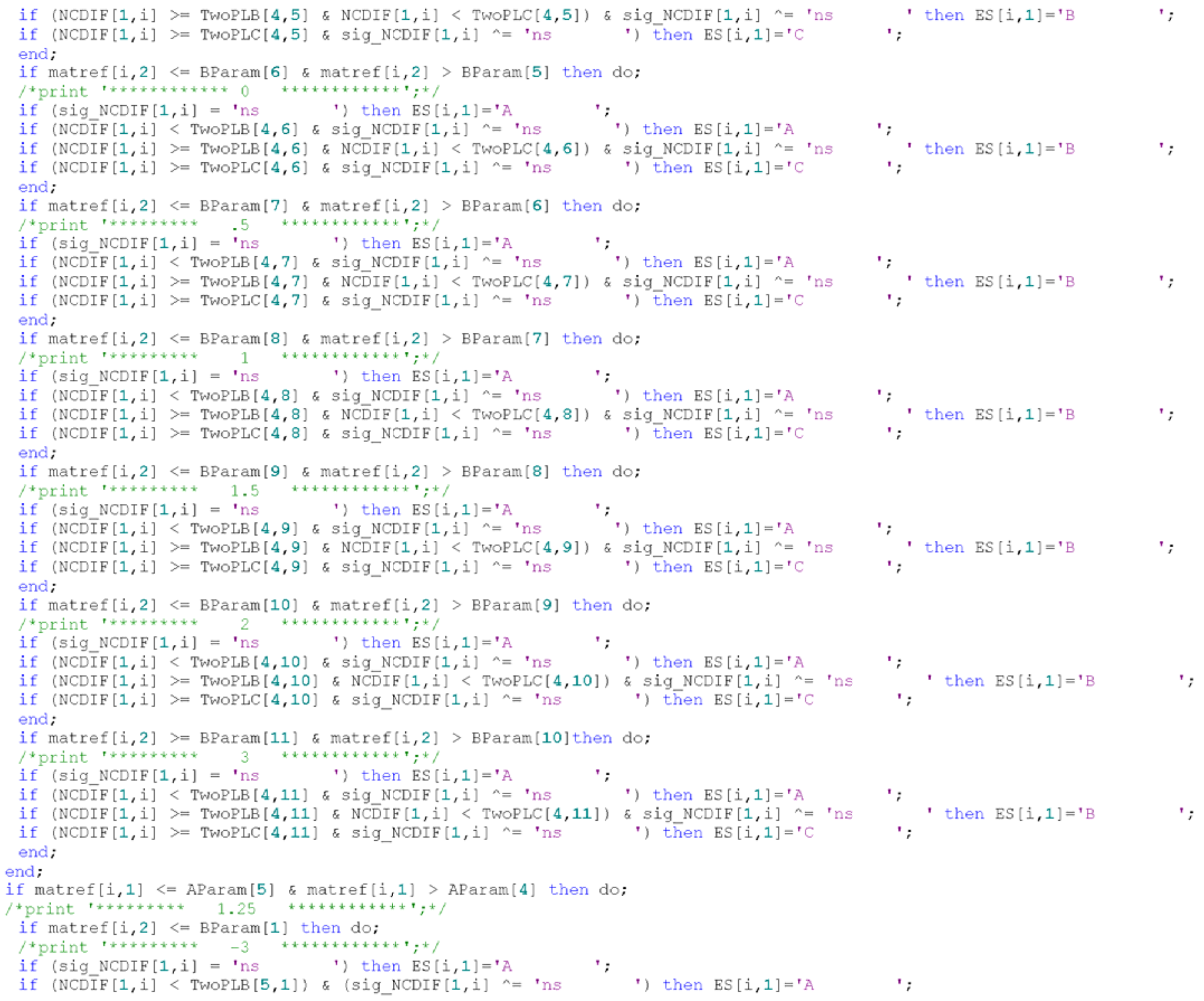
```
 /*print '************ 0   ************';*/
 if (sig_NCDIF[1,i] = 'ns       ') then ES[i,1]='A        ';
 if (NCDIF[1,i] < TwoPLB[2,6] & sig_NCDIF[1,i] ^= 'ns       ') then ES[i,1]='A        ';
 if (NCDIF[1,i] >= TwoPLB[2,6] & NCDIF[1,i] < TwoPLC[2,6]) & sig_NCDIF[1,i] ^= 'ns       ' then ES[i,1]='B        ';
 if (NCDIF[1,i] >= TwoPLC[2,6] & sig_NCDIF[1,i] ^= 'ns       ') then ES[i,1]='C        ';
 end;
 if matref[i,2] <= BParam[7] & matref[i,2] > BParam[6] then do;
 /*print '*********   .5   ************';*/
 if (sig_NCDIF[1,i] = 'ns       ') then ES[i,1]='A        ';
 if (NCDIF[1,i] < TwoPLB[2,7] & sig_NCDIF[1,i] ^= 'ns       ') then ES[i,1]='A        ';
 if (NCDIF[1,i] >= TwoPLB[2,7] & NCDIF[1,i] < TwoPLC[2,7]) & sig_NCDIF[1,i] ^= 'ns       ' then ES[i,1]='B        ';
 if (NCDIF[1,i] >= TwoPLC[2,7] & sig_NCDIF[1,i] ^= 'ns       ') then ES[i,1]='C        ';
 end;
 if matref[i,2] <= BParam[8] & matref[i,2] > BParam[7] then do;
 /*print '*********    1   ************';*/
 if (sig_NCDIF[1,i] = 'ns       ') then ES[i,1]='A        ';
 if (NCDIF[1,i] < TwoPLB[2,8] & sig_NCDIF[1,i] ^= 'ns       ') then ES[i,1]='A        ';
 if (NCDIF[1,i] >= TwoPLB[2,8] & NCDIF[1,i] < TwoPLC[2,8]) & sig_NCDIF[1,i] ^= 'ns       ' then ES[i,1]='B        ';
 if (NCDIF[1,i] >= TwoPLC[2,8] & sig_NCDIF[1,i] ^= 'ns       ') then ES[i,1]='C        ';
 end;
 if matref[i,2] <= BParam[9] & matref[i,2] > BParam[8] then do;
 /*print '*********   1.5   ************';*/
 if (sig_NCDIF[1,i] = 'ns       ') then ES[i,1]='A        ';
 if (NCDIF[1,i] < TwoPLB[2,9] & sig_NCDIF[1,i] ^= 'ns       ') then ES[i,1]='A        ';
 if (NCDIF[1,i] >= TwoPLB[2,9] & NCDIF[1,i] < TwoPLC[2,9]) & sig_NCDIF[1,i] ^= 'ns       ' then ES[i,1]='B        ';
 if (NCDIF[1,i] >= TwoPLC[2,9] & sig_NCDIF[1,i] ^= 'ns       ') then ES[i,1]='C        ';
 end;
 if matref[i,2] <= BParam[10] & matref[i,2] > BParam[9] then do;
 /*print '*********    2   ************';*/
 if (sig_NCDIF[1,i] = 'ns       ') then ES[i,1]='A        ';
 if (NCDIF[1,i] < TwoPLB[2,10] & sig_NCDIF[1,i] ^= 'ns       ') then ES[i,1]='A        ';
 if (NCDIF[1,i] >= TwoPLB[2,10] & NCDIF[1,i] < TwoPLC[2,10]) & sig_NCDIF[1,i] ^= 'ns       ' then ES[i,1]='B        ';
 if (NCDIF[1,i] >= TwoPLC[2,10] & sig_NCDIF[1,i] ^= 'ns       ') then ES[i,1]='C        ';
 end;
 if matref[i,2] >= BParam[11] & matref[i,2] > BParam[10]then do;
 /*print '*********    3   ************';*/
 if (sig_NCDIF[1,i] = 'ns       ') then ES[i,1]='A        ';
 if (NCDIF[1,i] < TwoPLB[2,11] & sig_NCDIF[1,i] ^= 'ns       ') then ES[i,1]='A        ';
 if (NCDIF[1,i] >= TwoPLB[2,11] & NCDIF[1,i] < TwoPLC[2,11]) & sig_NCDIF[1,i] ^= 'ns       ' then ES[i,1]='B        ';
 if (NCDIF[1,i] >= TwoPLC[2,11] & sig_NCDIF[1,i] ^= 'ns       ') then ES[i,1]='C        ';
 end;
end;
if matref[i,1] <= AParam[3] & matref[i,1] > AParam[2] then do;
/*print '*********   .75   ************';*/
 if matref[i,2] <= BParam[1] then do;
 /*print '*********   -3   ************';*/
 if (sig_NCDIF[1,i] = 'ns       ') then ES[i,1]='A        ';
 if (NCDIF[1,i] < TwoPLB[3,1]) & (sig_NCDIF[1,i] ^= 'ns       ') then ES[i,1]='A        ';
 if (NCDIF[1,i] >= TwoPLB[3,1] & NCDIF[1,i] < TwoPLC[3,1]) & sig_NCDIF[1,i] ^= 'ns       ' then ES[i,1]='B        ';
 if (NCDIF[1,i] >= TwoPLC[3,1] & sig_NCDIF[1,i] ^= 'ns       ') then ES[i,1]='C        ';
 end;
 if matref[i,2] <= BParam[2] & matref[i,2] > BParam[1] then do;
```



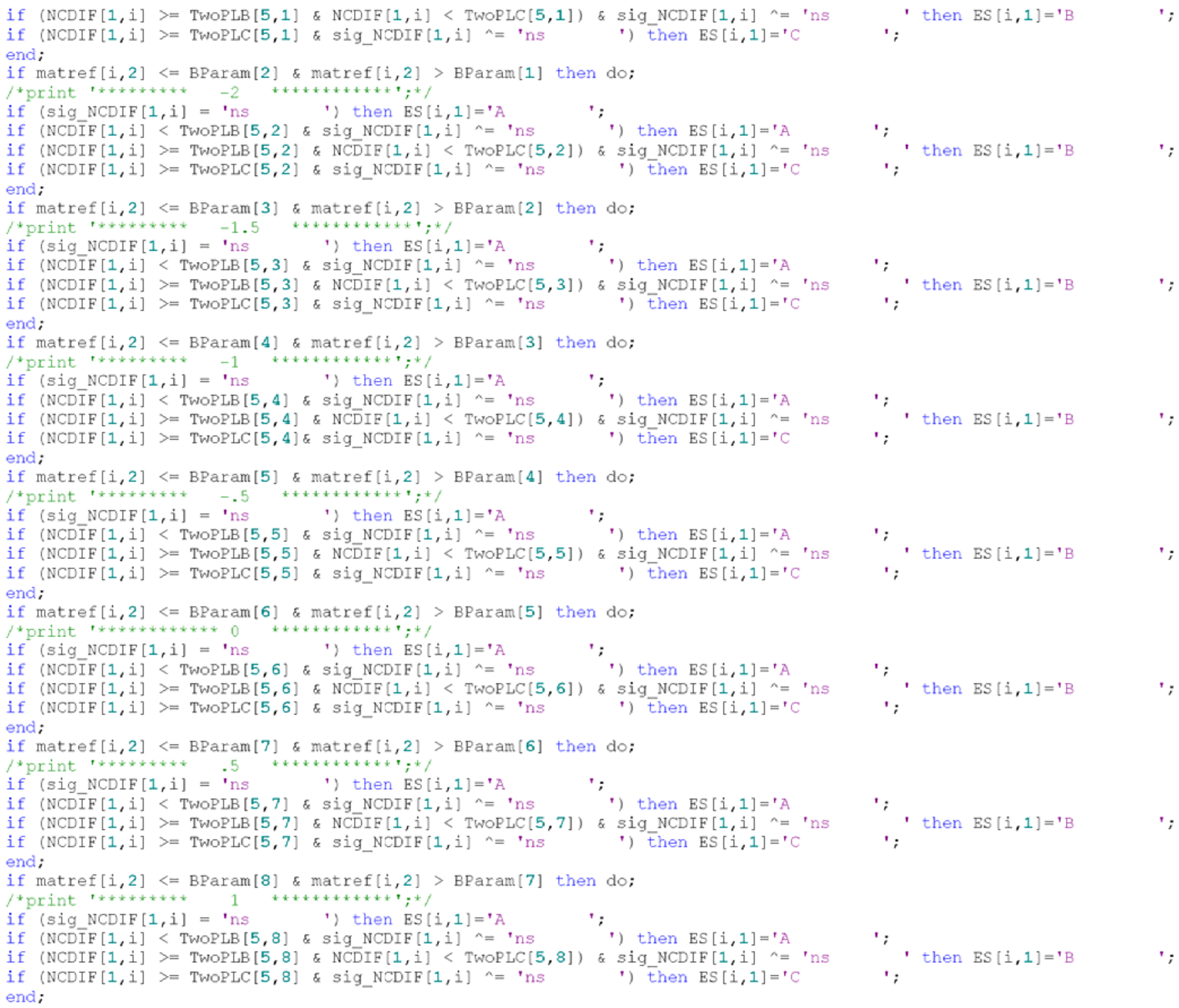
```
/*print '*********    -2    ************';*/
if (sig_NCDIF[1,i] = 'ns        ') then ES[i,1]='A        ';
if (NCDIF[1,i] < TwoPLB[3,2] & sig_NCDIF[1,i] ^= 'ns        ') then ES[i,1]='A        ';
if (NCDIF[1,i] >= TwoPLB[3,2] & NCDIF[1,i] < TwoPLC[3,2]) & sig_NCDIF[1,i] ^= 'ns        ' then ES[i,1]='B        ';
if (NCDIF[1,i] >= TwoPLC[3,2] & sig_NCDIF[1,i] ^= 'ns        ') then ES[i,1]='C        ';
end;
if matref[i,2] <= BParam[3] & matref[i,2] > BParam[2] then do;
/*print '*********    -1.5    ************';*/
if (sig_NCDIF[1,i] = 'ns        ') then ES[i,1]='A        ';
if (NCDIF[1,i] < TwoPLB[3,3] & sig_NCDIF[1,i] ^= 'ns        ') then ES[i,1]='A        ';
if (NCDIF[1,i] >= TwoPLB[3,3] & NCDIF[1,i] < TwoPLC[3,3]) & sig_NCDIF[1,i] ^= 'ns        ' then ES[i,1]='B        ';
if (NCDIF[1,i] >= TwoPLC[3,3] & sig_NCDIF[1,i] ^= 'ns        ') then ES[i,1]='C        ';
end;
if matref[i,2] <= BParam[4] & matref[i,2] > BParam[3] then do;
/*print '*********    -1    ************';*/
if (sig_NCDIF[1,i] = 'ns        ') then ES[i,1]='A        ';
if (NCDIF[1,i] < TwoPLB[3,4] & sig_NCDIF[1,i] ^= 'ns        ') then ES[i,1]='A        ';
if (NCDIF[1,i] >= TwoPLB[3,4] & NCDIF[1,i] < TwoPLC[3,4]) & sig_NCDIF[1,i] ^= 'ns        ' then ES[i,1]='B        ';
if (NCDIF[1,i] >= TwoPLC[3,4]& sig_NCDIF[1,i] ^= 'ns        ') then ES[i,1]='C        ';
end;
if matref[i,2] <= BParam[5] & matref[i,2] > BParam[4] then do;
/*print '*********    -.5    ************';*/
if (sig_NCDIF[1,i] = 'ns        ') then ES[i,1]='A        ';
if (NCDIF[1,i] < TwoPLB[3,5] & sig_NCDIF[1,i] ^= 'ns        ') then ES[i,1]='A        ';
if (NCDIF[1,i] >= TwoPLB[3,5] & NCDIF[1,i] < TwoPLC[3,5]) & sig_NCDIF[1,i] ^= 'ns        ' then ES[i,1]='B        ';
if (NCDIF[1,i] >= TwoPLC[3,5] & sig_NCDIF[1,i] ^= 'ns        ') then ES[i,1]='C        ';
end;
if matref[i,2] <= BParam[6] & matref[i,2] > BParam[5] then do;
/*print '************ 0    ************';*/
if (sig_NCDIF[1,i] = 'ns        ') then ES[i,1]='A        ';
if (NCDIF[1,i] < TwoPLB[3,6] & sig_NCDIF[1,i] ^= 'ns        ') then ES[i,1]='A        ';
if (NCDIF[1,i] >= TwoPLB[3,6] & NCDIF[1,i] < TwoPLC[3,6]) & sig_NCDIF[1,i] ^= 'ns        ' then ES[i,1]='B        ';
if (NCDIF[1,i] >= TwoPLC[3,6] & sig_NCDIF[1,i] ^= 'ns        ') then ES[i,1]='C        ';
end;
if matref[i,2] <= BParam[7] & matref[i,2] > BParam[6] then do;
/*print '*********    .5    ************';*/
if (sig_NCDIF[1,i] = 'ns        ') then ES[i,1]='A        ';
if (NCDIF[1,i] < TwoPLB[3,7] & sig_NCDIF[1,i] ^= 'ns        ') then ES[i,1]='A        ';
if (NCDIF[1,i] >= TwoPLB[3,7] & NCDIF[1,i] < TwoPLC[3,7]) & sig_NCDIF[1,i] ^= 'ns        ' then ES[i,1]='B        ';
if (NCDIF[1,i] >= TwoPLC[3,7] & sig_NCDIF[1,i] ^= 'ns        ') then ES[i,1]='C        ';
end;
if matref[i,2] <= BParam[8] & matref[i,2] > BParam[7] then do;
/*print '*********     1    ************';*/
if (sig_NCDIF[1,i] = 'ns        ') then ES[i,1]='A        ';
if (NCDIF[1,i] < TwoPLB[3,8] & sig_NCDIF[1,i] ^= 'ns        ') then ES[i,1]='A        ';
if (NCDIF[1,i] >= TwoPLB[3,8] & NCDIF[1,i] < TwoPLC[3,8]) & sig_NCDIF[1,i] ^= 'ns        ' then ES[i,1]='B        ';
if (NCDIF[1,i] >= TwoPLC[3,8] & sig_NCDIF[1,i] ^= 'ns        ') then ES[i,1]='C        ';
end;
if matref[i,2] <= BParam[9] & matref[i,2] > BParam[8] then do;
/*print '*********    1.5    ************';*/
if (sig_NCDIF[1,i] = 'ns        ') then ES[i,1]='A        ';
if (NCDIF[1,i] < TwoPLB[3,9] & sig_NCDIF[1,i] ^= 'ns        ') then ES[i,1]='A        ';
```



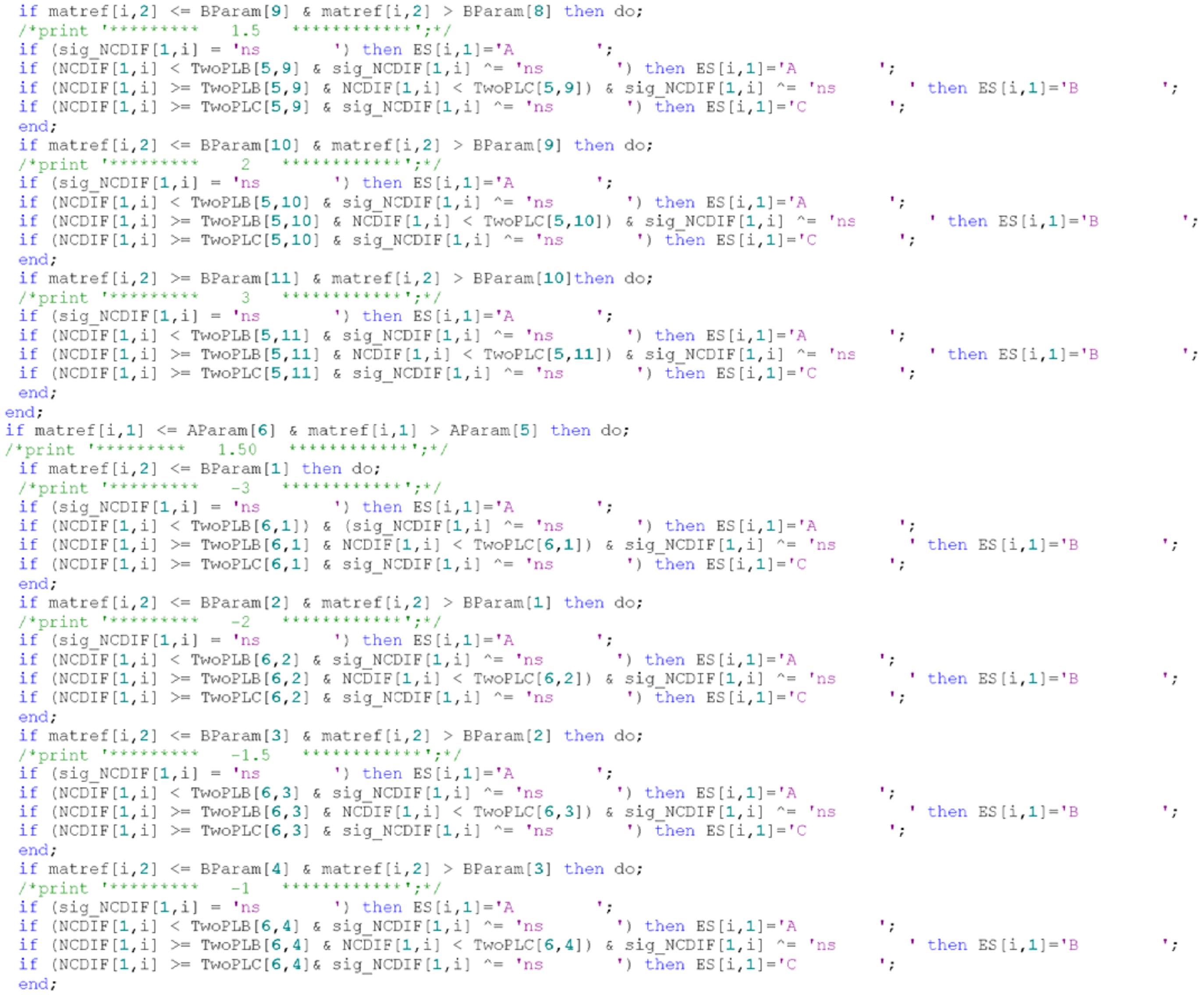
```
  if (NCDIF[1,i] >= TwoPLB[3,9] & NCDIF[1,i] < TwoPLC[3,9]) & sig_NCDIF[1,i] ^= 'ns       ' then ES[i,1]='B        ';
  if (NCDIF[1,i] >= TwoPLC[3,9] & sig_NCDIF[1,i] ^= 'ns       ') then ES[i,1]='C        ';
  end;
  if matref[i,2] <= BParam[10] & matref[i,2] > BParam[9] then do;
  /*print '*********    2   *************';*/
  if (sig_NCDIF[1,i] = 'ns        ') then ES[i,1]='A        ';
  if (NCDIF[1,i] < TwoPLB[3,10] & sig_NCDIF[1,i] ^= 'ns       ') then ES[i,1]='A        ';
  if (NCDIF[1,i] >= TwoPLB[3,10] & NCDIF[1,i] < TwoPLC[3,10]) & sig_NCDIF[1,i] ^= 'ns       ' then ES[i,1]='B        ';
  if (NCDIF[1,i] >= TwoPLC[3,10] & sig_NCDIF[1,i] ^= 'ns       ') then ES[i,1]='C        ';
  end;
  if matref[i,2] >= BParam[11] & matref[i,2] > BParam[10]then do;
  /*print '*********    3   *************';*/
  if (sig_NCDIF[1,i] = 'ns        ') then ES[i,1]='A        ';
  if (NCDIF[1,i] < TwoPLB[3,11] & sig_NCDIF[1,i] ^= 'ns       ') then ES[i,1]='A        ';
  if (NCDIF[1,i] >= TwoPLB[3,11] & NCDIF[1,i] < TwoPLC[3,11]) & sig_NCDIF[1,i] ^= 'ns       ' then ES[i,1]='B        ';
  if (NCDIF[1,i] >= TwoPLC[3,11] & sig_NCDIF[1,i] ^= 'ns       ') then ES[i,1]='C        ';
  end;
end;
if matref[i,1] <= AParam[4] & matref[i,1] > AParam[3] then do;
/*print '*********   .95   *************';*/
  if matref[i,2] <= BParam[1] then do;
  /*print '*********   -3   *************';*/
  if (sig_NCDIF[1,i] = 'ns        ') then ES[i,1]='A        ';
  if (NCDIF[1,i] < TwoPLB[4,1]) & (sig_NCDIF[1,i] ^= 'ns       ') then ES[i,1]='A        ';
  if (NCDIF[1,i] >= TwoPLB[4,1] & NCDIF[1,i] < TwoPLC[4,1]) & sig_NCDIF[1,i] ^= 'ns       ' then ES[i,1]='B        ';
  if (NCDIF[1,i] >= TwoPLC[4,1] & sig_NCDIF[1,i] ^= 'ns       ') then ES[i,1]='C        ';
  end;
  if matref[i,2] <= BParam[2] & matref[i,2] > BParam[1] then do;
  /*print '*********   -2   *************';*/
  if (sig_NCDIF[1,i] = 'ns        ') then ES[i,1]='A        ';
  if (NCDIF[1,i] < TwoPLB[4,2] & sig_NCDIF[1,i] ^= 'ns       ') then ES[i,1]='A        ';
  if (NCDIF[1,i] >= TwoPLB[4,2] & NCDIF[1,i] < TwoPLC[4,2]) & sig_NCDIF[1,i] ^= 'ns       ' then ES[i,1]='B        ';
  if (NCDIF[1,i] >= TwoPLC[4,2] & sig_NCDIF[1,i] ^= 'ns       ') then ES[i,1]='C        ';
  end;
  if matref[i,2] <= BParam[3] & matref[i,2] > BParam[2] then do;
  /*print '*********   -1.5   *************';*/
  if (sig_NCDIF[1,i] = 'ns        ') then ES[i,1]='A        ';
  if (NCDIF[1,i] < TwoPLB[4,3] & sig_NCDIF[1,i] ^= 'ns       ') then ES[i,1]='A        ';
  if (NCDIF[1,i] >= TwoPLB[4,3] & NCDIF[1,i] < TwoPLC[4,3]) & sig_NCDIF[1,i] ^= 'ns       ' then ES[i,1]='B        ';
  if (NCDIF[1,i] >= TwoPLC[4,3] & sig_NCDIF[1,i] ^= 'ns       ') then ES[i,1]='C        ';
  end;
  if matref[i,2] <= BParam[4] & matref[i,2] > BParam[3] then do;
  /*print '*********   -1   *************';*/
  if (sig_NCDIF[1,i] = 'ns        ') then ES[i,1]='A        ';
  if (NCDIF[1,i] < TwoPLB[4,4] & sig_NCDIF[1,i] ^= 'ns       ') then ES[i,1]='A        ';
  if (NCDIF[1,i] >= TwoPLB[4,4] & NCDIF[1,i] < TwoPLC[4,4]) & sig_NCDIF[1,i] ^= 'ns       ' then ES[i,1]='B        ';
  if (NCDIF[1,i] >= TwoPLC[4,4]& sig_NCDIF[1,i] ^= 'ns       ') then ES[i,1]='C        ';
  end;
  if matref[i,2] <= BParam[5] & matref[i,2] > BParam[4] then do;
  /*print '*********   -.5   *************';*/
  if (sig_NCDIF[1,i] = 'ns        ') then ES[i,1]='A        ';
  if (NCDIF[1,i] < TwoPLB[4,5] & sig_NCDIF[1,i] ^= 'ns       ') then ES[i,1]='A        ';
```



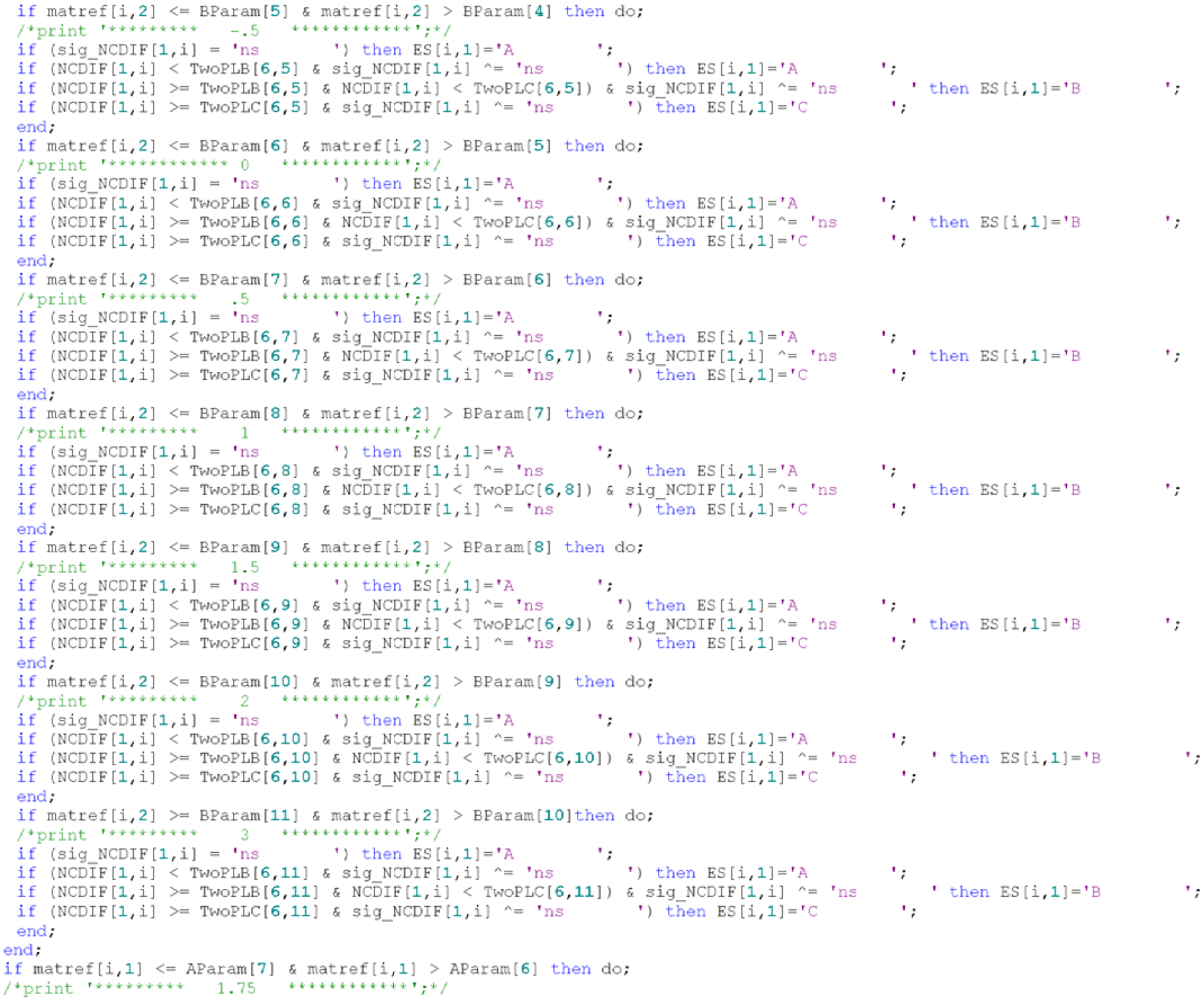
```
 if (NCDIF[1,i] >= TwoPLB[4,5] & NCDIF[1,i] < TwoPLC[4,5]) & sig_NCDIF[1,i] ^= 'ns        ' then ES[i,1]='B         ';
 if (NCDIF[1,i] >= TwoPLC[4,5] & sig_NCDIF[1,i] ^= 'ns        ') then ES[i,1]='C         ';
 end;
 if matref[i,2] <= BParam[6] & matref[i,2] > BParam[5] then do;
 /*print '************ 0   ************';*/
 if (sig_NCDIF[1,i] = 'ns        ') then ES[i,1]='A         ';
 if (NCDIF[1,i] < TwoPLB[4,6] & sig_NCDIF[1,i] ^= 'ns        ') then ES[i,1]='A         ';
 if (NCDIF[1,i] >= TwoPLB[4,6] & NCDIF[1,i] < TwoPLC[4,6]) & sig_NCDIF[1,i] ^= 'ns        ' then ES[i,1]='B         ';
 if (NCDIF[1,i] >= TwoPLC[4,6] & sig_NCDIF[1,i] ^= 'ns        ') then ES[i,1]='C         ';
 end;
 if matref[i,2] <= BParam[7] & matref[i,2] > BParam[6] then do;
 /*print '*********   .5   ************';*/
 if (sig_NCDIF[1,i] = 'ns        ') then ES[i,1]='A         ';
 if (NCDIF[1,i] < TwoPLB[4,7] & sig_NCDIF[1,i] ^= 'ns        ') then ES[i,1]='A         ';
 if (NCDIF[1,i] >= TwoPLB[4,7] & NCDIF[1,i] < TwoPLC[4,7]) & sig_NCDIF[1,i] ^= 'ns        ' then ES[i,1]='B         ';
 if (NCDIF[1,i] >= TwoPLC[4,7] & sig_NCDIF[1,i] ^= 'ns        ') then ES[i,1]='C         ';
 end;
 if matref[i,2] <= BParam[8] & matref[i,2] > BParam[7] then do;
 /*print '*********    1   ************';*/
 if (sig_NCDIF[1,i] = 'ns        ') then ES[i,1]='A         ';
 if (NCDIF[1,i] < TwoPLB[4,8] & sig_NCDIF[1,i] ^= 'ns        ') then ES[i,1]='A         ';
 if (NCDIF[1,i] >= TwoPLB[4,8] & NCDIF[1,i] < TwoPLC[4,8]) & sig_NCDIF[1,i] ^= 'ns        ' then ES[i,1]='B         ';
 if (NCDIF[1,i] >= TwoPLC[4,8] & sig_NCDIF[1,i] ^= 'ns        ') then ES[i,1]='C         ';
 end;
 if matref[i,2] <= BParam[9] & matref[i,2] > BParam[8] then do;
 /*print '*********   1.5   ************';*/
 if (sig_NCDIF[1,i] = 'ns        ') then ES[i,1]='A         ';
 if (NCDIF[1,i] < TwoPLB[4,9] & sig_NCDIF[1,i] ^= 'ns        ') then ES[i,1]='A         ';
 if (NCDIF[1,i] >= TwoPLB[4,9] & NCDIF[1,i] < TwoPLC[4,9]) & sig_NCDIF[1,i] ^= 'ns        ' then ES[i,1]='B         ';
 if (NCDIF[1,i] >= TwoPLC[4,9] & sig_NCDIF[1,i] ^= 'ns        ') then ES[i,1]='C         ';
 end;
 if matref[i,2] <= BParam[10] & matref[i,2] > BParam[9] then do;
 /*print '*********    2   ************';*/
 if (sig_NCDIF[1,i] = 'ns        ') then ES[i,1]='A         ';
 if (NCDIF[1,i] < TwoPLB[4,10] & sig_NCDIF[1,i] ^= 'ns        ') then ES[i,1]='A         ';
 if (NCDIF[1,i] >= TwoPLB[4,10] & NCDIF[1,i] < TwoPLC[4,10]) & sig_NCDIF[1,i] ^= 'ns        ' then ES[i,1]='B         ';
 if (NCDIF[1,i] >= TwoPLC[4,10] & sig_NCDIF[1,i] ^= 'ns        ') then ES[i,1]='C         ';
 end;
 if matref[i,2] >= BParam[11] & matref[i,2] > BParam[10]then do;
 /*print '*********    3   ************';*/
 if (sig_NCDIF[1,i] = 'ns        ') then ES[i,1]='A         ';
 if (NCDIF[1,i] < TwoPLB[4,11] & sig_NCDIF[1,i] ^= 'ns        ') then ES[i,1]='A         ';
 if (NCDIF[1,i] >= TwoPLB[4,11] & NCDIF[1,i] < TwoPLC[4,11]) & sig_NCDIF[1,i] ^= 'ns        ' then ES[i,1]='B         ';
 if (NCDIF[1,i] >= TwoPLC[4,11] & sig_NCDIF[1,i] ^= 'ns        ') then ES[i,1]='C         ';
 end;
end;
if matref[i,1] <= AParam[5] & matref[i,1] > AParam[4] then do;
/*print '*********   1.25   ************';*/
 if matref[i,2] <= BParam[1] then do;
 /*print '*********   -3   ************';*/
 if (sig_NCDIF[1,i] = 'ns        ') then ES[i,1]='A         ';
 if (NCDIF[1,i] < TwoPLB[5,1]) & (sig_NCDIF[1,i] ^= 'ns        ') then ES[i,1]='A         ';
```



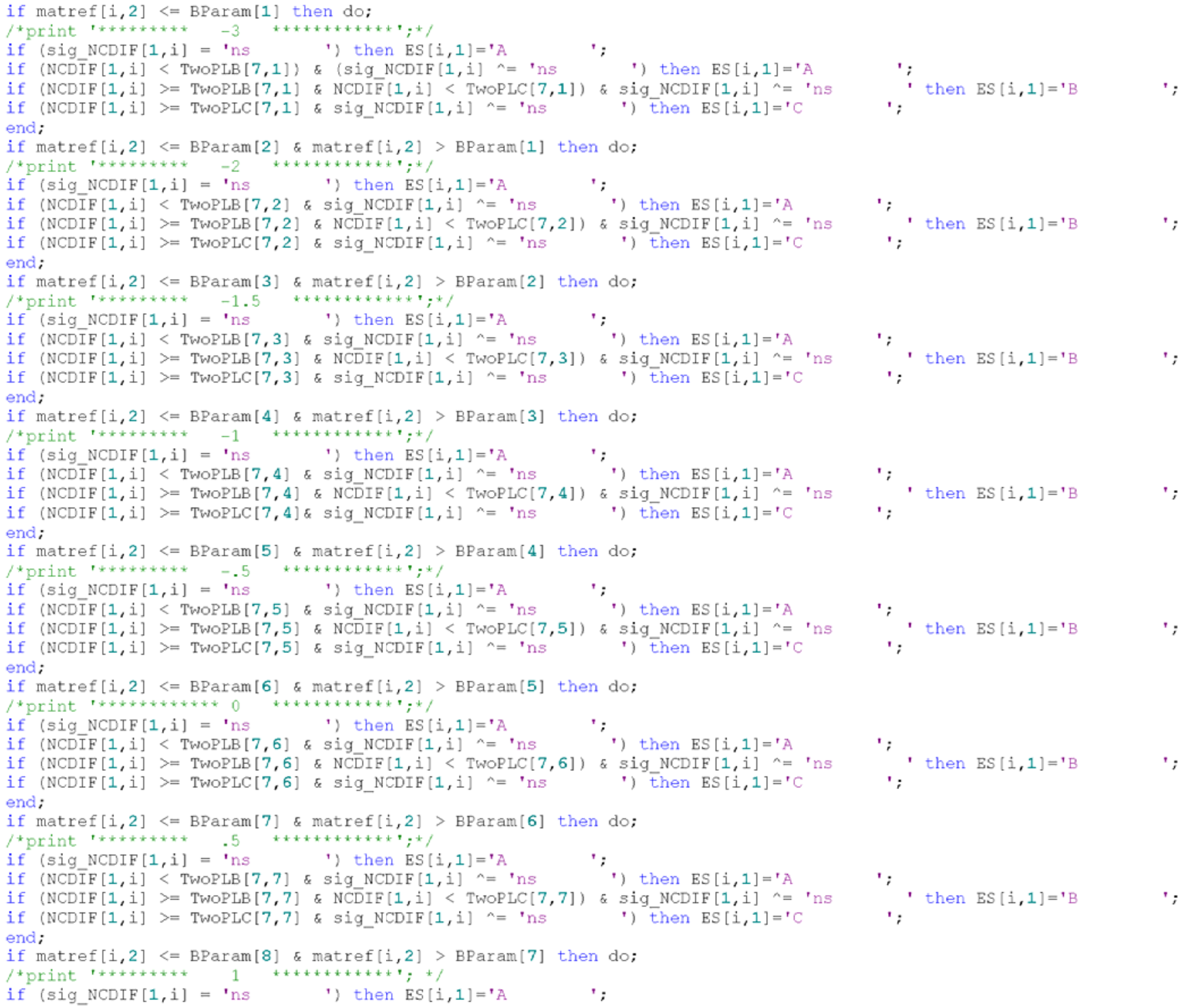
```
if (NCDIF[1,i] >= TwoPLB[5,1] & NCDIF[1,i] < TwoPLC[5,1]) & sig_NCDIF[1,i] ^= 'ns        ' then ES[i,1]='B         ';
if (NCDIF[1,i] >= TwoPLC[5,1] & sig_NCDIF[1,i] ^= 'ns       ') then ES[i,1]='C        ';
end;
if matref[i,2] <= BParam[2] & matref[i,2] > BParam[1] then do;
/*print '*********   -2   *************';*/
if (sig_NCDIF[1,i] = 'ns       ') then ES[i,1]='A        ';
if (NCDIF[1,i] < TwoPLB[5,2] & sig_NCDIF[1,i] ^= 'ns       ') then ES[i,1]='A        ';
if (NCDIF[1,i] >= TwoPLB[5,2] & NCDIF[1,i] < TwoPLC[5,2]) & sig_NCDIF[1,i] ^= 'ns        ' then ES[i,1]='B         ';
if (NCDIF[1,i] >= TwoPLC[5,2] & sig_NCDIF[1,i] ^= 'ns       ') then ES[i,1]='C        ';
end;
if matref[i,2] <= BParam[3] & matref[i,2] > BParam[2] then do;
/*print '*********   -1.5   *************';*/
if (sig_NCDIF[1,i] = 'ns       ') then ES[i,1]='A        ';
if (NCDIF[1,i] < TwoPLB[5,3] & sig_NCDIF[1,i] ^= 'ns       ') then ES[i,1]='A        ';
if (NCDIF[1,i] >= TwoPLB[5,3] & NCDIF[1,i] < TwoPLC[5,3]) & sig_NCDIF[1,i] ^= 'ns        ' then ES[i,1]='B         ';
if (NCDIF[1,i] >= TwoPLC[5,3] & sig_NCDIF[1,i] ^= 'ns       ') then ES[i,1]='C        ';
end;
if matref[i,2] <= BParam[4] & matref[i,2] > BParam[3] then do;
/*print '*********   -1   *************';*/
if (sig_NCDIF[1,i] = 'ns       ') then ES[i,1]='A        ';
if (NCDIF[1,i] < TwoPLB[5,4] & sig_NCDIF[1,i] ^= 'ns       ') then ES[i,1]='A        ';
if (NCDIF[1,i] >= TwoPLB[5,4] & NCDIF[1,i] < TwoPLC[5,4]) & sig_NCDIF[1,i] ^= 'ns        ' then ES[i,1]='B         ';
if (NCDIF[1,i] >= TwoPLC[5,4]& sig_NCDIF[1,i] ^= 'ns       ') then ES[i,1]='C        ';
end;
if matref[i,2] <= BParam[5] & matref[i,2] > BParam[4] then do;
/*print '*********   -.5   *************';*/
if (sig_NCDIF[1,i] = 'ns       ') then ES[i,1]='A        ';
if (NCDIF[1,i] < TwoPLB[5,5] & sig_NCDIF[1,i] ^= 'ns       ') then ES[i,1]='A        ';
if (NCDIF[1,i] >= TwoPLB[5,5] & NCDIF[1,i] < TwoPLC[5,5]) & sig_NCDIF[1,i] ^= 'ns        ' then ES[i,1]='B         ';
if (NCDIF[1,i] >= TwoPLC[5,5] & sig_NCDIF[1,i] ^= 'ns       ') then ES[i,1]='C        ';
end;
if matref[i,2] <= BParam[6] & matref[i,2] > BParam[5] then do;
/*print '************ 0   *************';*/
if (sig_NCDIF[1,i] = 'ns       ') then ES[i,1]='A        ';
if (NCDIF[1,i] < TwoPLB[5,6] & sig_NCDIF[1,i] ^= 'ns       ') then ES[i,1]='A        ';
if (NCDIF[1,i] >= TwoPLB[5,6] & NCDIF[1,i] < TwoPLC[5,6]) & sig_NCDIF[1,i] ^= 'ns        ' then ES[i,1]='B         ';
if (NCDIF[1,i] >= TwoPLC[5,6] & sig_NCDIF[1,i] ^= 'ns       ') then ES[i,1]='C        ';
end;
if matref[i,2] <= BParam[7] & matref[i,2] > BParam[6] then do;
/*print '*********   .5   *************';*/
if (sig_NCDIF[1,i] = 'ns       ') then ES[i,1]='A        ';
if (NCDIF[1,i] < TwoPLB[5,7] & sig_NCDIF[1,i] ^= 'ns       ') then ES[i,1]='A        ';
if (NCDIF[1,i] >= TwoPLB[5,7] & NCDIF[1,i] < TwoPLC[5,7]) & sig_NCDIF[1,i] ^= 'ns        ' then ES[i,1]='B         ';
if (NCDIF[1,i] >= TwoPLC[5,7] & sig_NCDIF[1,i] ^= 'ns       ') then ES[i,1]='C        ';
end;
if matref[i,2] <= BParam[8] & matref[i,2] > BParam[7] then do;
/*print '*********    1   *************';*/
if (sig_NCDIF[1,i] = 'ns       ') then ES[i,1]='A        ';
if (NCDIF[1,i] < TwoPLB[5,8] & sig_NCDIF[1,i] ^= 'ns       ') then ES[i,1]='A        ';
if (NCDIF[1,i] >= TwoPLB[5,8] & NCDIF[1,i] < TwoPLC[5,8]) & sig_NCDIF[1,i] ^= 'ns        ' then ES[i,1]='B         ';
if (NCDIF[1,i] >= TwoPLC[5,8] & sig_NCDIF[1,i] ^= 'ns       ') then ES[i,1]='C        ';
end;
```



```
 if matref[i,2] <= BParam[9] & matref[i,2] > BParam[8] then do;
 /*print '*********   1.5   ************';*/
 if (sig_NCDIF[1,i] = 'ns       ') then ES[i,1]='A        ';
 if (NCDIF[1,i] < TwoPLB[5,9] & sig_NCDIF[1,i] ^= 'ns       ') then ES[i,1]='A        ';
 if (NCDIF[1,i] >= TwoPLB[5,9] & NCDIF[1,i] < TwoPLC[5,9]) & sig_NCDIF[1,i] ^= 'ns        ' then ES[i,1]='B        ';
 if (NCDIF[1,i] >= TwoPLC[5,9] & sig_NCDIF[1,i] ^= 'ns       ') then ES[i,1]='C        ';
 end;
 if matref[i,2] <= BParam[10] & matref[i,2] > BParam[9] then do;
 /*print '*********    2   ************';*/
 if (sig_NCDIF[1,i] = 'ns       ') then ES[i,1]='A        ';
 if (NCDIF[1,i] < TwoPLB[5,10] & sig_NCDIF[1,i] ^= 'ns       ') then ES[i,1]='A        ';
 if (NCDIF[1,i] >= TwoPLB[5,10] & NCDIF[1,i] < TwoPLC[5,10]) & sig_NCDIF[1,i] ^= 'ns        ' then ES[i,1]='B        ';
 if (NCDIF[1,i] >= TwoPLC[5,10] & sig_NCDIF[1,i] ^= 'ns       ') then ES[i,1]='C        ';
 end;
 if matref[i,2] >= BParam[11] & matref[i,2] > BParam[10]then do;
 /*print '*********    3   ************';*/
 if (sig_NCDIF[1,i] = 'ns       ') then ES[i,1]='A        ';
 if (NCDIF[1,i] < TwoPLB[5,11] & sig_NCDIF[1,i] ^= 'ns       ') then ES[i,1]='A        ';
 if (NCDIF[1,i] >= TwoPLB[5,11] & NCDIF[1,i] < TwoPLC[5,11]) & sig_NCDIF[1,i] ^= 'ns        ' then ES[i,1]='B        ';
 if (NCDIF[1,i] >= TwoPLC[5,11] & sig_NCDIF[1,i] ^= 'ns       ') then ES[i,1]='C        ';
 end;
end;
if matref[i,1] <= AParam[6] & matref[i,1] > AParam[5] then do;
/*print '*********   1.50   ************';*/
 if matref[i,2] <= BParam[1] then do;
 /*print '*********   -3   ************';*/
 if (sig_NCDIF[1,i] = 'ns       ') then ES[i,1]='A        ';
 if (NCDIF[1,i] < TwoPLB[6,1]) & (sig_NCDIF[1,i] ^= 'ns       ') then ES[i,1]='A        ';
 if (NCDIF[1,i] >= TwoPLB[6,1] & NCDIF[1,i] < TwoPLC[6,1]) & sig_NCDIF[1,i] ^= 'ns        ' then ES[i,1]='B        ';
 if (NCDIF[1,i] >= TwoPLC[6,1] & sig_NCDIF[1,i] ^= 'ns       ') then ES[i,1]='C        ';
 end;
 if matref[i,2] <= BParam[2] & matref[i,2] > BParam[1] then do;
 /*print '*********   -2   ************';*/
 if (sig_NCDIF[1,i] = 'ns       ') then ES[i,1]='A        ';
 if (NCDIF[1,i] < TwoPLB[6,2] & sig_NCDIF[1,i] ^= 'ns       ') then ES[i,1]='A        ';
 if (NCDIF[1,i] >= TwoPLB[6,2] & NCDIF[1,i] < TwoPLC[6,2]) & sig_NCDIF[1,i] ^= 'ns        ' then ES[i,1]='B        ';
 if (NCDIF[1,i] >= TwoPLC[6,2] & sig_NCDIF[1,i] ^= 'ns       ') then ES[i,1]='C        ';
 end;
 if matref[i,2] <= BParam[3] & matref[i,2] > BParam[2] then do;
 /*print '*********   -1.5   ************';*/
 if (sig_NCDIF[1,i] = 'ns       ') then ES[i,1]='A        ';
 if (NCDIF[1,i] < TwoPLB[6,3] & sig_NCDIF[1,i] ^= 'ns       ') then ES[i,1]='A        ';
 if (NCDIF[1,i] >= TwoPLB[6,3] & NCDIF[1,i] < TwoPLC[6,3]) & sig_NCDIF[1,i] ^= 'ns        ' then ES[i,1]='B        ';
 if (NCDIF[1,i] >= TwoPLC[6,3] & sig_NCDIF[1,i] ^= 'ns       ') then ES[i,1]='C        ';
 end;
 if matref[i,2] <= BParam[4] & matref[i,2] > BParam[3] then do;
 /*print '*********   -1   ************';*/
 if (sig_NCDIF[1,i] = 'ns       ') then ES[i,1]='A        ';
 if (NCDIF[1,i] < TwoPLB[6,4] & sig_NCDIF[1,i] ^= 'ns       ') then ES[i,1]='A        ';
 if (NCDIF[1,i] >= TwoPLB[6,4] & NCDIF[1,i] < TwoPLC[6,4]) & sig_NCDIF[1,i] ^= 'ns        ' then ES[i,1]='B        ';
 if (NCDIF[1,i] >= TwoPLC[6,4]& sig_NCDIF[1,i] ^= 'ns       ') then ES[i,1]='C        ';
 end;
```

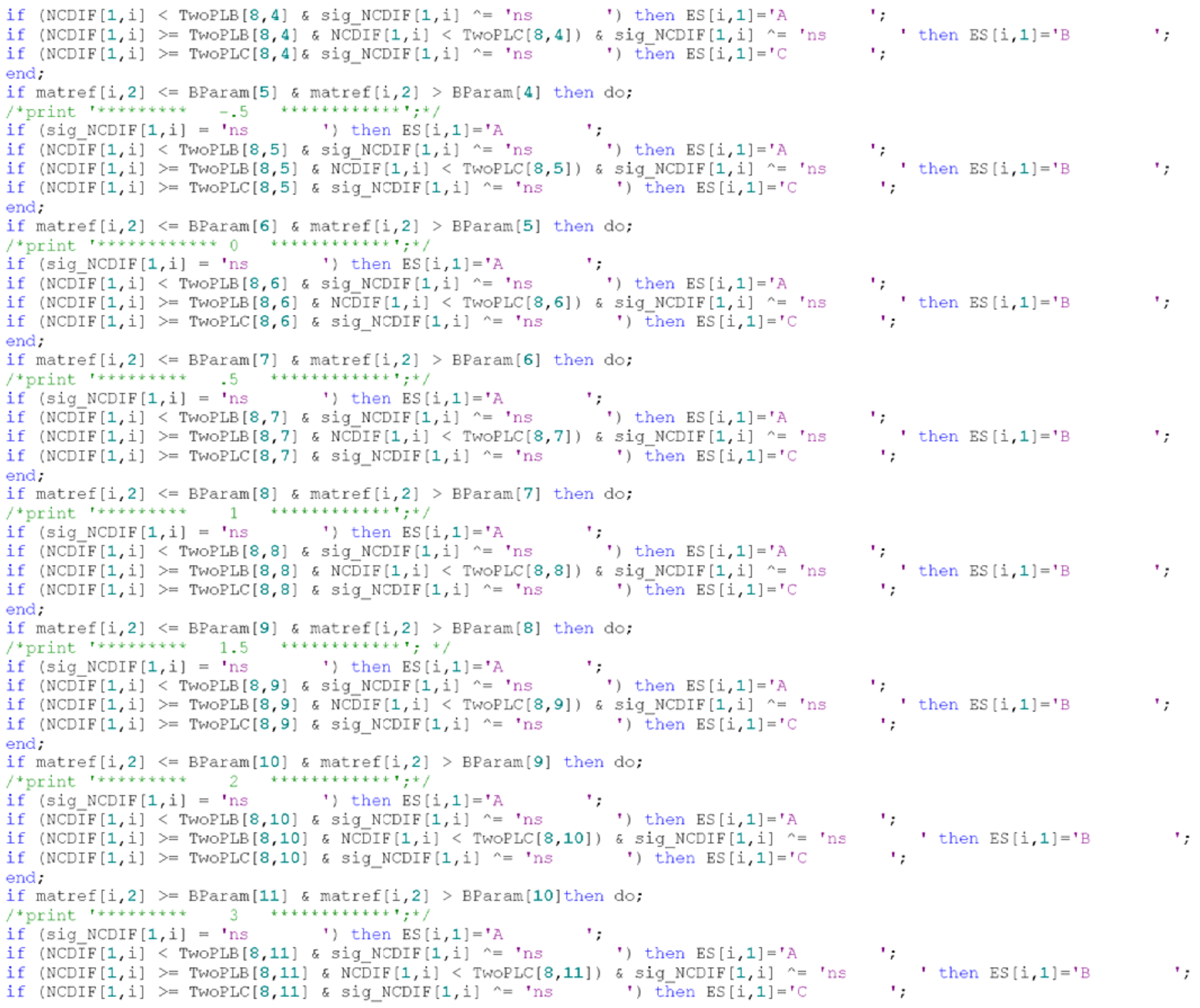


```
 if matref[i,2] <= BParam[5] & matref[i,2] > BParam[4] then do;
 /*print '*********   -.5   ************';*/
 if (sig_NCDIF[1,i] = 'ns        ') then ES[i,1]='A        ';
 if (NCDIF[1,i] < TwoPLB[6,5] & sig_NCDIF[1,i] ^= 'ns        ') then ES[i,1]='A        ';
 if (NCDIF[1,i] >= TwoPLB[6,5] & NCDIF[1,i] < TwoPLC[6,5]) & sig_NCDIF[1,i] ^= 'ns        ' then ES[i,1]='B        ';
 if (NCDIF[1,i] >= TwoPLC[6,5] & sig_NCDIF[1,i] ^= 'ns        ') then ES[i,1]='C        ';
 end;
 if matref[i,2] <= BParam[6] & matref[i,2] > BParam[5] then do;
 /*print '************ 0   ************';*/
 if (sig_NCDIF[1,i] = 'ns        ') then ES[i,1]='A        ';
 if (NCDIF[1,i] < TwoPLB[6,6] & sig_NCDIF[1,i] ^= 'ns        ') then ES[i,1]='A        ';
 if (NCDIF[1,i] >= TwoPLB[6,6] & NCDIF[1,i] < TwoPLC[6,6]) & sig_NCDIF[1,i] ^= 'ns        ' then ES[i,1]='B        ';
 if (NCDIF[1,i] >= TwoPLC[6,6] & sig_NCDIF[1,i] ^= 'ns        ') then ES[i,1]='C        ';
 end;
 if matref[i,2] <= BParam[7] & matref[i,2] > BParam[6] then do;
 /*print '*********   .5   ************';*/
 if (sig_NCDIF[1,i] = 'ns        ') then ES[i,1]='A        ';
 if (NCDIF[1,i] < TwoPLB[6,7] & sig_NCDIF[1,i] ^= 'ns        ') then ES[i,1]='A        ';
 if (NCDIF[1,i] >= TwoPLB[6,7] & NCDIF[1,i] < TwoPLC[6,7]) & sig_NCDIF[1,i] ^= 'ns        ' then ES[i,1]='B        ';
 if (NCDIF[1,i] >= TwoPLC[6,7] & sig_NCDIF[1,i] ^= 'ns        ') then ES[i,1]='C        ';
 end;
 if matref[i,2] <= BParam[8] & matref[i,2] > BParam[7] then do;
 /*print '*********    1   ************';*/
 if (sig_NCDIF[1,i] = 'ns        ') then ES[i,1]='A        ';
 if (NCDIF[1,i] < TwoPLB[6,8] & sig_NCDIF[1,i] ^= 'ns        ') then ES[i,1]='A        ';
 if (NCDIF[1,i] >= TwoPLB[6,8] & NCDIF[1,i] < TwoPLC[6,8]) & sig_NCDIF[1,i] ^= 'ns        ' then ES[i,1]='B        ';
 if (NCDIF[1,i] >= TwoPLC[6,8] & sig_NCDIF[1,i] ^= 'ns        ') then ES[i,1]='C        ';
 end;
 if matref[i,2] <= BParam[9] & matref[i,2] > BParam[8] then do;
 /*print '*********   1.5   ************';*/
 if (sig_NCDIF[1,i] = 'ns        ') then ES[i,1]='A        ';
 if (NCDIF[1,i] < TwoPLB[6,9] & sig_NCDIF[1,i] ^= 'ns        ') then ES[i,1]='A        ';
 if (NCDIF[1,i] >= TwoPLB[6,9] & NCDIF[1,i] < TwoPLC[6,9]) & sig_NCDIF[1,i] ^= 'ns        ' then ES[i,1]='B        ';
 if (NCDIF[1,i] >= TwoPLC[6,9] & sig_NCDIF[1,i] ^= 'ns        ') then ES[i,1]='C        ';
 end;
 if matref[i,2] <= BParam[10] & matref[i,2] > BParam[9] then do;
 /*print '*********    2   ************';*/
 if (sig_NCDIF[1,i] = 'ns        ') then ES[i,1]='A        ';
 if (NCDIF[1,i] < TwoPLB[6,10] & sig_NCDIF[1,i] ^= 'ns        ') then ES[i,1]='A        ';
 if (NCDIF[1,i] >= TwoPLB[6,10] & NCDIF[1,i] < TwoPLC[6,10]) & sig_NCDIF[1,i] ^= 'ns        ' then ES[i,1]='B        ';
 if (NCDIF[1,i] >= TwoPLC[6,10] & sig_NCDIF[1,i] ^= 'ns        ') then ES[i,1]='C        ';
 end;
 if matref[i,2] >= BParam[11] & matref[i,2] > BParam[10]then do;
 /*print '*********    3   ************';*/
 if (sig_NCDIF[1,i] = 'ns        ') then ES[i,1]='A        ';
 if (NCDIF[1,i] < TwoPLB[6,11] & sig_NCDIF[1,i] ^= 'ns        ') then ES[i,1]='A        ';
 if (NCDIF[1,i] >= TwoPLB[6,11] & NCDIF[1,i] < TwoPLC[6,11]) & sig_NCDIF[1,i] ^= 'ns        ' then ES[i,1]='B        ';
 if (NCDIF[1,i] >= TwoPLC[6,11] & sig_NCDIF[1,i] ^= 'ns        ') then ES[i,1]='C        ';
 end;
end;
if matref[i,1] <= AParam[7] & matref[i,1] > AParam[6] then do;
/*print '*********   1.75   ************';*/
```



```
if matref[i,2] <= BParam[1] then do;
/*print '*********   -3   ************';*/
if (sig_NCDIF[1,i] = 'ns        ') then ES[i,1]='A        ';
if (NCDIF[1,i] < TwoPLB[7,1]) & (sig_NCDIF[1,i] ^= 'ns        ') then ES[i,1]='A        ';
if (NCDIF[1,i] >= TwoPLB[7,1] & NCDIF[1,i] < TwoPLC[7,1]) & sig_NCDIF[1,i] ^= 'ns        ' then ES[i,1]='B        ';
if (NCDIF[1,i] >= TwoPLC[7,1] & sig_NCDIF[1,i] ^= 'ns        ') then ES[i,1]='C        ';
end;
if matref[i,2] <= BParam[2] & matref[i,2] > BParam[1] then do;
/*print '*********   -2   ************';*/
if (sig_NCDIF[1,i] = 'ns        ') then ES[i,1]='A        ';
if (NCDIF[1,i] < TwoPLB[7,2] & sig_NCDIF[1,i] ^= 'ns        ') then ES[i,1]='A        ';
if (NCDIF[1,i] >= TwoPLB[7,2] & NCDIF[1,i] < TwoPLC[7,2]) & sig_NCDIF[1,i] ^= 'ns        ' then ES[i,1]='B        ';
if (NCDIF[1,i] >= TwoPLC[7,2] & sig_NCDIF[1,i] ^= 'ns        ') then ES[i,1]='C        ';
end;
if matref[i,2] <= BParam[3] & matref[i,2] > BParam[2] then do;
/*print '*********   -1.5   ************';*/
if (sig_NCDIF[1,i] = 'ns        ') then ES[i,1]='A        ';
if (NCDIF[1,i] < TwoPLB[7,3] & sig_NCDIF[1,i] ^= 'ns        ') then ES[i,1]='A        ';
if (NCDIF[1,i] >= TwoPLB[7,3] & NCDIF[1,i] < TwoPLC[7,3]) & sig_NCDIF[1,i] ^= 'ns        ' then ES[i,1]='B        ';
if (NCDIF[1,i] >= TwoPLC[7,3] & sig_NCDIF[1,i] ^= 'ns        ') then ES[i,1]='C        ';
end;
if matref[i,2] <= BParam[4] & matref[i,2] > BParam[3] then do;
/*print '*********   -1   ************';*/
if (sig_NCDIF[1,i] = 'ns        ') then ES[i,1]='A        ';
if (NCDIF[1,i] < TwoPLB[7,4] & sig_NCDIF[1,i] ^= 'ns        ') then ES[i,1]='A        ';
if (NCDIF[1,i] >= TwoPLB[7,4] & NCDIF[1,i] < TwoPLC[7,4]) & sig_NCDIF[1,i] ^= 'ns        ' then ES[i,1]='B        ';
if (NCDIF[1,i] >= TwoPLC[7,4]& sig_NCDIF[1,i] ^= 'ns        ') then ES[i,1]='C        ';
end;
if matref[i,2] <= BParam[5] & matref[i,2] > BParam[4] then do;
/*print '*********   -.5   ************';*/
if (sig_NCDIF[1,i] = 'ns        ') then ES[i,1]='A        ';
if (NCDIF[1,i] < TwoPLB[7,5] & sig_NCDIF[1,i] ^= 'ns        ') then ES[i,1]='A        ';
if (NCDIF[1,i] >= TwoPLB[7,5] & NCDIF[1,i] < TwoPLC[7,5]) & sig_NCDIF[1,i] ^= 'ns        ' then ES[i,1]='B        ';
if (NCDIF[1,i] >= TwoPLC[7,5] & sig_NCDIF[1,i] ^= 'ns        ') then ES[i,1]='C        ';
end;
if matref[i,2] <= BParam[6] & matref[i,2] > BParam[5] then do;
/*print '************ 0   ************';*/
if (sig_NCDIF[1,i] = 'ns        ') then ES[i,1]='A        ';
if (NCDIF[1,i] < TwoPLB[7,6] & sig_NCDIF[1,i] ^= 'ns        ') then ES[i,1]='A        ';
if (NCDIF[1,i] >= TwoPLB[7,6] & NCDIF[1,i] < TwoPLC[7,6]) & sig_NCDIF[1,i] ^= 'ns        ' then ES[i,1]='B        ';
if (NCDIF[1,i] >= TwoPLC[7,6] & sig_NCDIF[1,i] ^= 'ns        ') then ES[i,1]='C        ';
end;
if matref[i,2] <= BParam[7] & matref[i,2] > BParam[6] then do;
/*print '*********   .5   ************';*/
if (sig_NCDIF[1,i] = 'ns        ') then ES[i,1]='A        ';
if (NCDIF[1,i] < TwoPLB[7,7] & sig_NCDIF[1,i] ^= 'ns        ') then ES[i,1]='A        ';
if (NCDIF[1,i] >= TwoPLB[7,7] & NCDIF[1,i] < TwoPLC[7,7]) & sig_NCDIF[1,i] ^= 'ns        ' then ES[i,1]='B        ';
if (NCDIF[1,i] >= TwoPLC[7,7] & sig_NCDIF[1,i] ^= 'ns        ') then ES[i,1]='C        ';
end;
if matref[i,2] <= BParam[8] & matref[i,2] > BParam[7] then do;
/*print '*********    1   ************'; */
if (sig_NCDIF[1,i] = 'ns        ') then ES[i,1]='A        ';
```

```
 if (NCDIF[1,i] < TwoPLB[7,8] & sig_NCDIF[1,i] ^= 'ns       ') then ES[i,1]='A        ';
 if (NCDIF[1,i] >= TwoPLB[7,8] & NCDIF[1,i] < TwoPLC[7,8]) & sig_NCDIF[1,i] ^= 'ns       ' then ES[i,1]='B        ';
 if (NCDIF[1,i] >= TwoPLC[7,8] & sig_NCDIF[1,i] ^= 'ns       ') then ES[i,1]='C        ';
 end;
 if matref[i,2] <= BParam[9] & matref[i,2] > BParam[8] then do;
 /*print '*********   1.5   ************';*/
 if (sig_NCDIF[1,i] = 'ns       ') then ES[i,1]='A        ';
 if (NCDIF[1,i] < TwoPLB[7,9] & sig_NCDIF[1,i] ^= 'ns       ') then ES[i,1]='A        ';
 if (NCDIF[1,i] >= TwoPLB[7,9] & NCDIF[1,i] < TwoPLC[7,9]) & sig_NCDIF[1,i] ^= 'ns       ' then ES[i,1]='B        ';
 if (NCDIF[1,i] >= TwoPLC[7,9] & sig_NCDIF[1,i] ^= 'ns       ') then ES[i,1]='C        ';
 end;
 if matref[i,2] <= BParam[10] & matref[i,2] > BParam[9] then do;
 /*print '*********    2   ************';*/
 if (sig_NCDIF[1,i] = 'ns       ') then ES[i,1]='A        ';
 if (NCDIF[1,i] < TwoPLB[7,10] & sig_NCDIF[1,i] ^= 'ns       ') then ES[i,1]='A        ';
 if (NCDIF[1,i] >= TwoPLB[7,10] & NCDIF[1,i] < TwoPLC[7,10]) & sig_NCDIF[1,i] ^= 'ns       ' then ES[i,1]='B        ';
 if (NCDIF[1,i] >= TwoPLC[7,10] & sig_NCDIF[1,i] ^= 'ns       ') then ES[i,1]='C        ';
 end;
 if matref[i,2] >= BParam[11] & matref[i,2] > BParam[10]then do;
 /*print '*********    3   ************';*/
 if (sig_NCDIF[1,i] = 'ns       ') then ES[i,1]='A        ';
 if (NCDIF[1,i] < TwoPLB[7,11] & sig_NCDIF[1,i] ^= 'ns       ') then ES[i,1]='A        ';
 if (NCDIF[1,i] >= TwoPLB[7,11] & NCDIF[1,i] < TwoPLC[7,11]) & sig_NCDIF[1,i] ^= 'ns       ' then ES[i,1]='B        ';
 if (NCDIF[1,i] >= TwoPLC[7,11] & sig_NCDIF[1,i] ^= 'ns       ') then ES[i,1]='C        ';
 end;
end;
if matref[i,1] >= AParam[8] & matref[i,1] > AParam[7]then do;
/*print '*********   2.00   ************';*/
 if matref[i,2] <= BParam[1] then do;
 /*print '*********   -3   ************';*/
 if (sig_NCDIF[1,i] = 'ns       ') then ES[i,1]='A        ';
 if (NCDIF[1,i] < TwoPLB[8,1]) & (sig_NCDIF[1,i] ^= 'ns       ') then ES[i,1]='A        ';
 if (NCDIF[1,i] >= TwoPLB[8,1] & NCDIF[1,i] < TwoPLC[8,1]) & sig_NCDIF[1,i] ^= 'ns       ' then ES[i,1]='B        ';
 if (NCDIF[1,i] >= TwoPLC[8,1] & sig_NCDIF[1,i] ^= 'ns       ') then ES[i,1]='C        ';
 end;
 if matref[i,2] <= BParam[2] & matref[i,2] > BParam[1] then do;
 /*print '*********   -2   ************';*/
 if (sig_NCDIF[1,i] = 'ns       ') then ES[i,1]='A        ';
 if (NCDIF[1,i] < TwoPLB[8,2] & sig_NCDIF[1,i] ^= 'ns       ') then ES[i,1]='A        ';
 if (NCDIF[1,i] >= TwoPLB[8,2] & NCDIF[1,i] < TwoPLC[8,2]) & sig_NCDIF[1,i] ^= 'ns       ' then ES[i,1]='B        ';
 if (NCDIF[1,i] >= TwoPLC[8,2] & sig_NCDIF[1,i] ^= 'ns       ') then ES[i,1]='C        ';
 end;
 if matref[i,2] <= BParam[3] & matref[i,2] > BParam[2] then do;
 /*print '*********   -1.5   ************';*/
 if (sig_NCDIF[1,i] = 'ns       ') then ES[i,1]='A        ';
 if (NCDIF[1,i] < TwoPLB[8,3] & sig_NCDIF[1,i] ^= 'ns       ') then ES[i,1]='A        ';
 if (NCDIF[1,i] >= TwoPLB[8,3] & NCDIF[1,i] < TwoPLC[8,3]) & sig_NCDIF[1,i] ^= 'ns       ' then ES[i,1]='B        ';
 if (NCDIF[1,i] >= TwoPLC[8,3] & sig_NCDIF[1,i] ^= 'ns       ') then ES[i,1]='C        ';
 end;
 if matref[i,2] <= BParam[4] & matref[i,2] > BParam[3] then do;
 /*print '*********   -1   ************';*/
 if (sig_NCDIF[1,i] = 'ns       ') then ES[i,1]='A        ';
```



```
if (NCDIF[1,i] < TwoPLB[8,4] & sig_NCDIF[1,i] ^= 'ns        ') then ES[i,1]='A         ';
if (NCDIF[1,i] >= TwoPLB[8,4] & NCDIF[1,i] < TwoPLC[8,4]) & sig_NCDIF[1,i] ^= 'ns        ' then ES[i,1]='B         ';
if (NCDIF[1,i] >= TwoPLC[8,4]& sig_NCDIF[1,i] ^= 'ns        ') then ES[i,1]='C         ';
end;
if matref[i,2] <= BParam[5] & matref[i,2] > BParam[4] then do;
/*print '*********   -.5   ************';*/
if (sig_NCDIF[1,i] = 'ns        ') then ES[i,1]='A         ';
if (NCDIF[1,i] < TwoPLB[8,5] & sig_NCDIF[1,i] ^= 'ns        ') then ES[i,1]='A         ';
if (NCDIF[1,i] >= TwoPLB[8,5] & NCDIF[1,i] < TwoPLC[8,5]) & sig_NCDIF[1,i] ^= 'ns        ' then ES[i,1]='B         ';
if (NCDIF[1,i] >= TwoPLC[8,5] & sig_NCDIF[1,i] ^= 'ns        ') then ES[i,1]='C         ';
end;
if matref[i,2] <= BParam[6] & matref[i,2] > BParam[5] then do;
/*print '************ 0   ************';*/
if (sig_NCDIF[1,i] = 'ns        ') then ES[i,1]='A         ';
if (NCDIF[1,i] < TwoPLB[8,6] & sig_NCDIF[1,i] ^= 'ns        ') then ES[i,1]='A         ';
if (NCDIF[1,i] >= TwoPLB[8,6] & NCDIF[1,i] < TwoPLC[8,6]) & sig_NCDIF[1,i] ^= 'ns        ' then ES[i,1]='B         ';
if (NCDIF[1,i] >= TwoPLC[8,6] & sig_NCDIF[1,i] ^= 'ns        ') then ES[i,1]='C         ';
end;
if matref[i,2] <= BParam[7] & matref[i,2] > BParam[6] then do;
/*print '*********   .5   ************';*/
if (sig_NCDIF[1,i] = 'ns        ') then ES[i,1]='A         ';
if (NCDIF[1,i] < TwoPLB[8,7] & sig_NCDIF[1,i] ^= 'ns        ') then ES[i,1]='A         ';
if (NCDIF[1,i] >= TwoPLB[8,7] & NCDIF[1,i] < TwoPLC[8,7]) & sig_NCDIF[1,i] ^= 'ns        ' then ES[i,1]='B         ';
if (NCDIF[1,i] >= TwoPLC[8,7] & sig_NCDIF[1,i] ^= 'ns        ') then ES[i,1]='C         ';
end;
if matref[i,2] <= BParam[8] & matref[i,2] > BParam[7] then do;
/*print '*********    1   ************';*/
if (sig_NCDIF[1,i] = 'ns        ') then ES[i,1]='A         ';
if (NCDIF[1,i] < TwoPLB[8,8] & sig_NCDIF[1,i] ^= 'ns        ') then ES[i,1]='A         ';
if (NCDIF[1,i] >= TwoPLB[8,8] & NCDIF[1,i] < TwoPLC[8,8]) & sig_NCDIF[1,i] ^= 'ns        ' then ES[i,1]='B         ';
if (NCDIF[1,i] >= TwoPLC[8,8] & sig_NCDIF[1,i] ^= 'ns        ') then ES[i,1]='C         ';
end;
if matref[i,2] <= BParam[9] & matref[i,2] > BParam[8] then do;
/*print '*********   1.5   ************'; */
if (sig_NCDIF[1,i] = 'ns        ') then ES[i,1]='A         ';
if (NCDIF[1,i] < TwoPLB[8,9] & sig_NCDIF[1,i] ^= 'ns        ') then ES[i,1]='A         ';
if (NCDIF[1,i] >= TwoPLB[8,9] & NCDIF[1,i] < TwoPLC[8,9]) & sig_NCDIF[1,i] ^= 'ns        ' then ES[i,1]='B         ';
if (NCDIF[1,i] >= TwoPLC[8,9] & sig_NCDIF[1,i] ^= 'ns        ') then ES[i,1]='C         ';
end;
if matref[i,2] <= BParam[10] & matref[i,2] > BParam[9] then do;
/*print '*********    2   ************';*/
if (sig_NCDIF[1,i] = 'ns        ') then ES[i,1]='A         ';
if (NCDIF[1,i] < TwoPLB[8,10] & sig_NCDIF[1,i] ^= 'ns        ') then ES[i,1]='A         ';
if (NCDIF[1,i] >= TwoPLB[8,10] & NCDIF[1,i] < TwoPLC[8,10]) & sig_NCDIF[1,i] ^= 'ns        ' then ES[i,1]='B         ';
if (NCDIF[1,i] >= TwoPLC[8,10] & sig_NCDIF[1,i] ^= 'ns        ') then ES[i,1]='C         ';
end;
if matref[i,2] >= BParam[11] & matref[i,2] > BParam[10]then do;
/*print '*********    3   ************';*/
if (sig_NCDIF[1,i] = 'ns        ') then ES[i,1]='A         ';
if (NCDIF[1,i] < TwoPLB[8,11] & sig_NCDIF[1,i] ^= 'ns        ') then ES[i,1]='A         ';
if (NCDIF[1,i] >= TwoPLB[8,11] & NCDIF[1,i] < TwoPLC[8,11]) & sig_NCDIF[1,i] ^= 'ns        ' then ES[i,1]='B         ';
if (NCDIF[1,i] >= TwoPLC[8,11] & sig_NCDIF[1,i] ^= 'ns        ') then ES[i,1]='C         ';
```

```
       end;
     end;
   end;
end;
/******************************************* END OF IPR EFFECT SIZE CODE ****************************/
```

```
/************ FORMATTING OUTPUT – K. D. Wright *****************/
cov1=char(NCDIF2);
cov2=char(PVALUE2);
cov3=char(EMPIRICAL_POWER);

out1=cov1;
out2=out1||sig_NCDIF2;
out3=out2||cov2;
out4=out3||cov3;
DIF_ANALYSIS=out4||ES;

names={NCDIF, SIGLEVEL, PVALUE, POWER, EffectSize};
print DIF_ANALYSIS [rowname="" colname=names];

/**** MORE OUTPUT CODE ****/
create out5 FROM DIF_ANALYSIS [colname={NCDIF, SIGLEVEL, PVLAUE, POWER,
EffectSize}];
append from DIF_ANALYSIS;


quit;

run;


PROC EXPORT DATA=out5
OUTFILE="C:\powerdissertation\output\out5.csv";
RUN;


DATA newout;
      FILENAME IO 'C:\powerdissertation\output';
      INFILE IO(out5.csv) dlm='2C0D'x dsd missover lrecl=10000
firstobs=2;
      INPUT NCDIF SIGLEVEL $ PVALUE EmpiricalPower EffectSize $;
RUN;

PROC PRINT data=newout;
RUN;
/************ END FORMATTING OUTPUT – K. D. Wright ****************/
```