2007-09-12

Differential Item Functioning Analysis of the Herrmann Brain Dominance Instrument

Jared Andrew Lees
Brigham Young University - Provo

DIFFERENTIAL ITEM FUNCTIONING ANALYSIS OF THE HERRMANN BRAIN DOMINANCE INSTRUMENT

by

Jared A. Lees

A master's project submitted to the faculty of

Brigham Young University

in partial fulfillment of the requirements for the degree of

Master of Science

Department of Instructional Psychology and Technology

Brigham Young University

July 2007

BRIGHAM YOUNG UNIVERSITY

GRADUATE COMMITTEE APPROVAL

of a masters project submitted by

Jared A. Lees

This project has been read by each member of the following graduate committee and by majority vote has been found to be satisfactory.

Date Richard R Sudweeks, Chair

Date Charles R. Graham

Date C. Victor Bunderson

Date Diane Strong-Krause

BRIGHAM YOUNG UNIVERSITY

As chair of the candidate's graduate committee, I have read the selected project of Jared A. Lees in its final form and have found that (1) its format, citations, and bibliographical style are consistent and acceptable to fulfill university and department style requirements; (2) its illustrative materials including figures, tables, and charts are in place; and (3) the final manuscript is satisfactory to the graduate committee and is ready for submission to the university library.

Date

Richard R Sudweeks
Chair, Graduate Committee

Accepted for the Department

Andrew S. Gibbons
Department Chair

Accepted for the College

K. Richard Young
Dean, David O. McKay School of Education

ABSTRACT

DIFFERENTIAL ITEM FUNCTIONING ANALYSIS OF THE HERRMANN BRAIN DOMINANCE INSTRUMENT

Jared A. Lees

Department of Instructional Psychology and Technology

Master of Science

Differential item functioning (DIF) is present when examinees who have the same level of a trait have a different probability of correctly answering a test item intended to measure that trait (Shepard & Averill, 1981). The following study is a DIF analysis of the Herrmann Brain Dominance Instrument (HBDI), a preference profiling instrument developed by Herrmann International to help individuals identify their dominant preferences and then classify their level of dominance into four preference quadrants. Examinees who completed the American English version of the instrument were classified as the reference group and examinees of the International English version were classified as the focal group. Out of 105 items, 11 were manifesting a large amount of DIF and were flagged for further review. The POLYSIBTEST procedure was used to carry out the DIF analysis. POLYSIBTEST is an extension of the SIBTEST procedure, which is a conceptually simple method for analyzing DIF that uses a latent trait measure

rather than an observed total score. The latent trait measure helps detect both uniform and nonuniform DIF and the POLYSIBTEST procedure is used for both dichotomous and polytomous items. Each of the four preference quadrants were analyzed separately to reduce incorrect findings as a result of ipsative scoring. The process used to complete the DIF analysis was documented so that additional language groups may be analyzed by Herrmann International.

ACKNOWLEDGEMENTS

I am indeed thankful to those who have helped me complete this project. The knowledge, expertise, and guidance of my committee members along with their willingness to help have led me to completion. I wish to especially thank Dr. Sudweeks and Dr. Bunderson for their encouragement to further my understanding of concepts and theories I once thought were beyond my grasp. I also wish to thank Dr. Mark J. Gierl of the Centre for Research in Applied Measurement and Evaluation in Alberta, Canada for starting me off in the right direction. Gratitude is also extended to Applied Psychological Measurement Inc. for awarding me a grant to purchase the necessary software in order to complete the analysis for this study.

I am most grateful for the support of my wife and daughter. Their patience, example, and love gave me the encouragement and confidence needed to press forward.

TABLE OF CONTENTS

Chapter 1: Introduction..1

Herrmann International..1

Rationale for Study..3

Statement of Purpose 4

Chapter 2: Literature Review 6

Differential Item Functioning 8

Detecting DIF..11

Mantel-Haenszel 11

SIBTEST 12

Reducing DIF..14

Judgmental Reviews 14

Back-translation 16

Summary..17

Chapter 3: Method..18

Instrumentation 18

Participants..19

Analysis 20

Procedure 21

Chapter 4: Results..23

Research Objective 1 23

Research Objective 2 23

Research Objective 3 25

Chapter 5: Discussion 27

Conclusion 27

Item Review 27

Recommendations 30

Closing Remarks 31

Schedule 32

Budget 32

References..35

Appendix A : DIF Analysis Procedure..39

LIST OF TABLES

Table Page

1. Proportion of HBDI Items That Show Evidence of DIF 23

2. Classification and Magnitude of DIF Items that Favor the Reference Group 24

3. Classification and Magnitude of DIF Items that Favor the Focal Group 25

4. Schedule 33

5. Budget 34

LIST OF FIGURES

Figure Page

1. HBDI profile examples from *The Creative Brian* 3

2. ICCs for a test item with no DIF between groups 9

3. ICCs for a test item with DIF between groups 10

4. Items 101 to 105 of the HBDI.. 19

Chapter 1: Introduction

Measurement instruments, such as tests and surveys, are in continuous use throughout the world. Since multiple organizations administer and design measurement instruments, it is important for these organizations to monitor the validity of said instruments to ensure ethical administration and accurate reporting of measurement items. The following study was designed to help Herrmann International detect possible sources of measurement invalidity of their instrument items.

Herrmann International

For over 25 years, Herrmann International has conducted research on how profiling thinking preferences of company officers and employees can benefit businesses. Through their research they have developed the Herrmann Brain Dominance Instrument™ (HBDI), a preference profiling instrument designed to help individuals identify their dominant preferences. According to Herrmann International, individuals begin developing preferences at infancy. For example, a child will reach with one hand as a preferred way to take hold of objects. This preferred hand then develops into the hand used for writing and becomes the dominant hand. As life progresses for the infant, other physical and mental decisions are made. As similar decisions are repeated, preferences are formed that help identify the type of person the infant has become (Herrmann, 1994). Since life offers limited resources, different individuals develop different preferences as decisions are made.

Preference identification may help organizations improve productivity, creativity, and results by leveraging the differences in the thinking styles of individuals. For example, an individual who prefers creative thinking may design an innovative product that satisfies the customer’s needs, but is over time and over budget as a result of focusing more effort on the creative aspect of the product instead of the logistics of completing the product. An individual who prefers logistical thinking will ensure the product is on time and on budget, but may lack the creative ability to develop a product that satisfies the customer’s needs. When the creative person and logistical person are teamed together, the result will be an innovative product that is on time and on budget and fulfills the customer’s requirements. The HBDI helps define how preferences affect behaviors in planning, teamwork, and communication. Through understanding your own preferences, as well as preferences of others in your organization, communication, efficiency, and productivity will increase (Herrmann, 2007).

The HBDI classifies individuals into one or more of four preference quadrants labeled A, B, C, and D. Validation data shows that a person classified in quadrant A favors factual, logical, rational, and mathematical thinking. A person classified in quadrant B favors administrative, controlled, and planned thinking with a rule and a place for everything. The B person focuses on perfection in details and works on one task at a time. A person classed in the C quadrant favors more musical, spiritual, talkative, emotional, and empathetic thinking. The C person is sensitive and receptive, trying to take the most out of an experience. Similarly, an individual classified in the D quadrant favors holistic, creative, and synthesizing thinking. The D person looks for new ideas, possibilities, oddities, and incongruities, thus tending to be a visionary. The four preference quadrants are not mutually exclusive. An individual may be classified as having a dominant preference in more than one quadrant. In fact, over 90% of the individuals who completed the HBDI have dominant preferences in multiple quadrants (Herrmann, 1994).

Figure 1 illustrates four profile examples resulting from the HBDI. The quadrant containing the greatest portion of the quadralateral area indicates a dominant preference quadrant. For example, the engineer is dominant in quadrant A and the musician is dominant in both quadrant C and D.

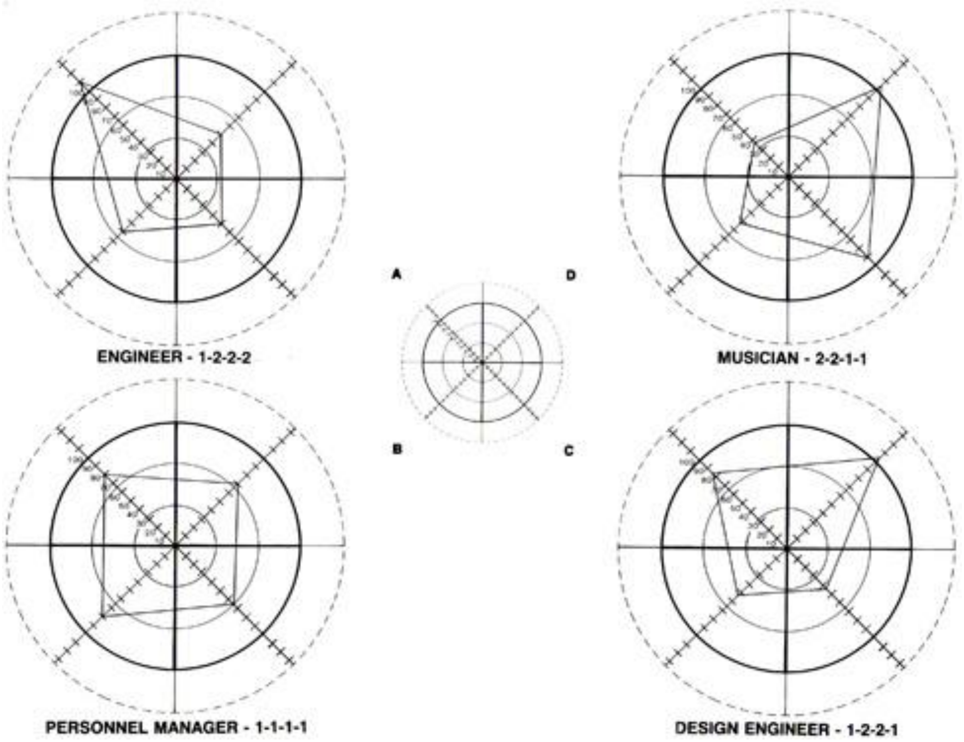


Figure 1. HBDI profile examples from *The Creative Brian* (Herrmann, 1994).

Rationale for Study

Herrmann International has translated and adapted the HBDI for administration into 18 different language groups and is interested in how well the instrument performs in each group. Although the HBDI is currently in widespread use, Herrmann International desired to take a closer look at the behavior of the instrument items to verify that they are measuring the intended preference trait, or construct, they were designed to measure before translation and adaptation. Differential item functioning (DIF) analysis was used in this study to analyze the HBDI at the item level. DIF is present when examinees from different groups have a different probability of selecting an item given that the examinees share the same level of the preference trait being measured by that item (Shepard, Camilli, & Averill, 1981).

It is important to note that Herrmann International refers to the HBDI as a *measurement instrument* rather than a test because they believe the term measurement instrument allows the participant to feel that each quadrant is equally important and that the HBDI is measuring the preferred quadrant or quadrants of the participant. Reference to the HBDI as a test may cause the participant to think he or she is in a right or wrong quadrant, not a preferred quadrant. In the HBDI and for this study, a correct response indicates the examinee with a particular preference trait selects the item that measures that particular preference trait. DIF literature uses the term *ability,* which does not apply to the HBDI because the HBDI measures preference traits, not ability. The HBDI traits are degrees of preference for the type of thinking found each of the four quadrants of the Whole Brain Thinking Model.

Statement of Purpose

The purpose of this project was to determine the direction and magnitude of DIF present in items in the HBDI with the U.S. English participants as the reference group and International English participants as the focal group. Differential functioning items were classified based on the amount of DIF present according to the guidelines proposed by Roussos and Stout (1996). In addition, a procedure to analyze DIF between other language groups of the HBDI was developed through documenting the process used to complete this study.

The study focused on the following three research objectives:

1. To identify the proportion of items that function differentially between the American English and International English examinees.
2. To classify individual items that function differentially according to the direction and magnitude of DIF detected and identify items with a large magnitude of DIF for suggested judgmental review of content.
3. To define a procedure for performing DIF analysis that can be replicated for additional language groups that take the HBDI.

Chapter 2: Literature Review

Testing and measurement instruments, including achievement, aptitude, and personality tests and surveys, are increasingly being translated and adapted into different languages and cultures (Allaoulf, Hambleton & Sireci, 1999; Gierl & Khaliq, 2001; Hambleton, Merenda, & Spielberger, 2005). For example, Spanish versions of the College Board's Scholastic Assessment Test (SAT) and the American Council on Education's General Educational Development (GED) test are currently in preparation for use in the United States (Hambleton et al., 2005). With businesses expanding to different continents, education spreading across languages, and technology increasing and facilitating the transfer of information, the need for measurement instruments worldwide has grown.

According to Hambleton et al. (2005), adaptation includes the following activities: (a) deciding whether or not a test could measure the same construct in a different language and culture, (b) selecting translators, (c) deciding on appropriate accommodations to be made in preparing the test for use in a second language, and (d) adapting the test and checking its equivalence in the adapted form. If measurement instruments are not adapted and administered properly, potential threats to the validity of the constructs being measured by the instrument will surface.

An example of an instrument that appeared to have threats to validity is found in Golden Rule, a lawsuit between the Golden Rule Insurance Company versus the Illinois Department of Insurance and the Educational Testing Service (ETS) filed in 1976. The plaintiffs alleged that the licensing test created by ETS and administered by the Illinois Department of Insurance discriminated against a minority group. An out-of-court settlement agreement was reached in November of 1984. As a result of the agreement, procedures and provisions intended to reduce discrimination in measurement practices were created. Legislation in other states proposed the use of these provisions and procedures in other testing situations. In January 1987, Gregory Anrig, the president of ETS, recanted the 1984 agreement because of “legislative proposals that go far beyond the limited terms of the original agreement.” The terms of the original agreement allowed ETS to adequately maintain the quality of the Illinois licensing exam by attempting to control for differences in ability between examinee groups. However, widespread use of the agreement in other situations ignores the possibility that such group differences on individual test items may validly reflect real differences in the trait being measured. ETS realized that comparisons on differences in the mean group performance of individual items cannot be made without matching examinees on the same level of the trait being measured. As a result, ETS began using differential item functioning to match examinees from different groups in order to determine if test items are measuring their intended constructs (Faggen, 1987; Haney & Reidy, 1987).

Along with the Golden Rule settlement and later practices developed by ETS, other organizations have organized standards and practices for the creation, adaptation, and implementation of measurement instruments (or tests) in order to ensure the instrument is measuring the intended constructs. Through a collaborative effort with other professional societies, the American Educational Research Association (AERA) (1999) has published standards for test developers, administrators, and users to promote the sound and ethical use of measurement instruments and to provide a basis for evaluating the quality of measurement practices.

The two standards listed below are of particular importance with respect to Herrmann International's desire to ensure the items of the HBDI measure the desired preference traits after translation and adaptation into other language groups:

> *Standard 7.3.* When credible research reports that differential item functioning exists across age, gender, racial/ethnic, cultural, disability, and/or linguistic groups in the population of test takers in the content domain measured by the test, test developers should conduct appropriate studies when feasible. Such research should seek to detect and eliminate aspects of test design, content, and format that might bias test scores for particular groups. (p. 81)
>
> *Standard 9.9.* When multiple language versions of a test are intended to be comparable, test developers should report evidence of test comparability. (p. 99)

These standards along with prior research suggest that translated and adapted instrument items should be checked for differential item functioning, which is a threat to the validity of the construct the items are intended to measure (Clauser & Mazor, 1998; Gierl, Rogers, & Klinger, 1999; Hambleton et al., 2005).

Differential Item Functioning

According to Shepard et al. (1981), DIF is present when examinees from different groups have a different probability of answering an item correctly after controlling for overall ability (p. 319). Control for overall ability indicates that examinees with similar levels of a trait should have the same probability of correctly answering a test item intended to measure that trait. A test item is then analyzed at the ability level. The process is repeated at each of the different ability levels and the results are aggregated across all ability levels (Clauser & Mazor, 1998).

Figure 2 provides a graphical example of an item that does not contain DIF. The example is from Item Response Theory (IRT), which uses mathematical functions that relate the probability of a correct response on an item to overall examinee ability. IRT is used here to help define DIF, but is not the method used to conduct the DIF analysis in the study that follows. The item characteristic curves (ICCs) illustrated in Figure 2 and Figure 3 describe how the probability of giving a correct response to this item varies as a function of examinees' trait level. The probability increases as examinees level of the trait increases, but the rate of increase is not constant. What is important in this example is that the item characteristic curves for the reference and focal groups are identical. In other words, at any given level of the trait, the members of the two groups have the same probability of answering the item correctly. This indicates that no DIF is present.

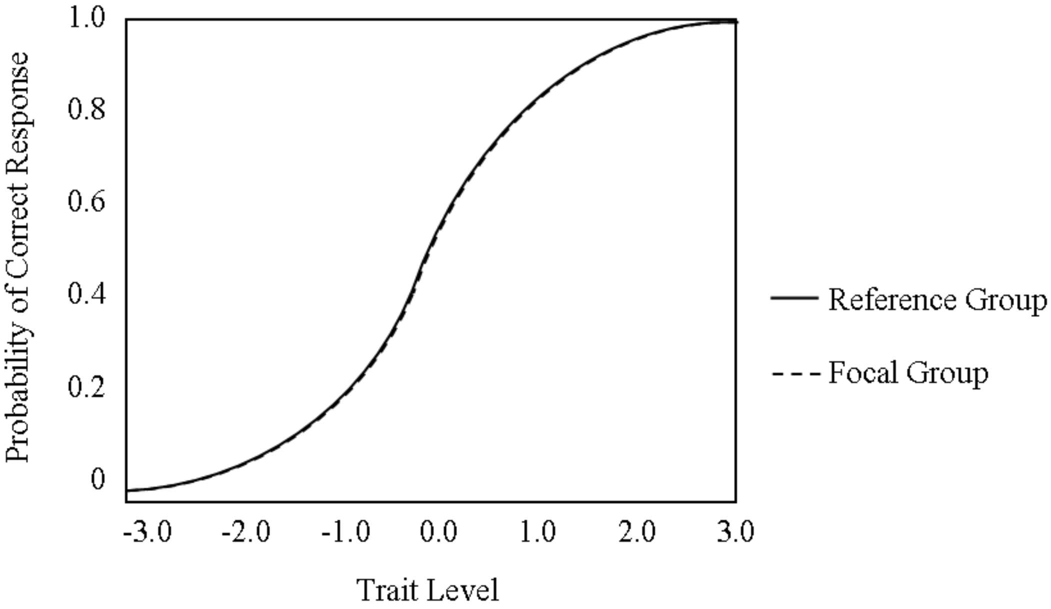


Figure 2. ICCs for a test item with no DIF between groups.

In Figure 3, the ICCs of the reference and focal groups are different, indicating that at each level of the trait, members of the reference group have a higher probability of answering correctly and the members of the focal group have a disadvantage. The greater the distance between the curves, the higher the magnitude of DIF (Anastasi & Urbina, 1997; Camilli & Shepard, 1994; Clauser & Mazor, 1998).

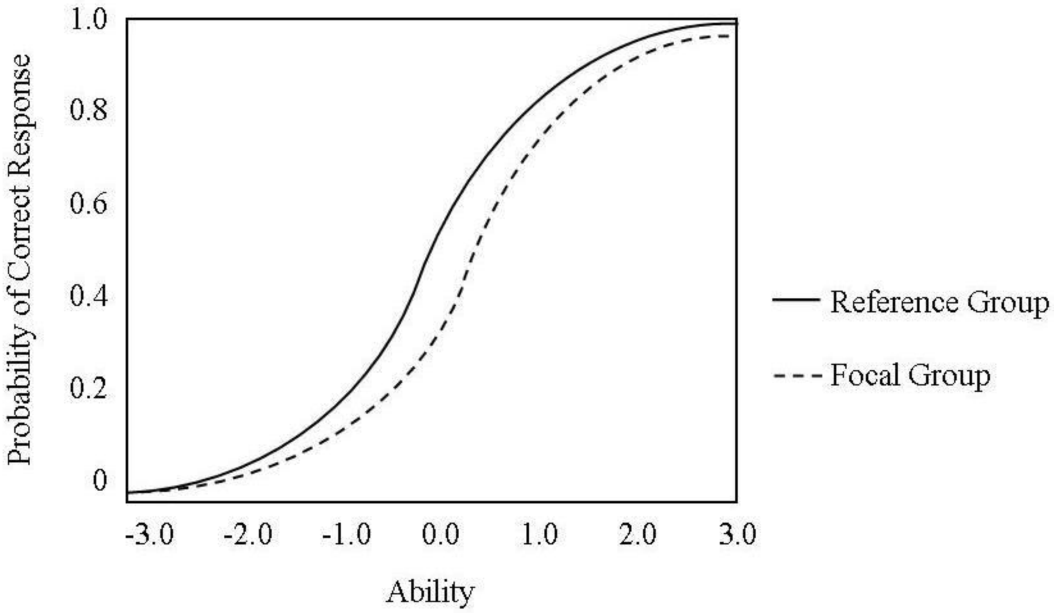


Figure 3. ICCs for a test item with DIF between groups.

Hambleton (1994) provides an example of an item containing DIF from a test using Swedish-English comparisons. In the test, English-speaking examinees were presented with this item:

Where is a bird with webbed feet most likely to live?

a. in the mountains

b. in the woods

c. in the sea

d. in the desert (p. 235).

In the Swedish translation the phrase “webbed feet” became “swimming feet” thereby providing an obvious clue to the Swedish-speaking examinees about the correct option for this item. The clue gives the Swedish-speaking examinees a different probability of answering the item correctly.

Detecting DIF

Psychometricians have developed multiple procedures for detecting DIF. The following two procedures were considered for use in this study due to their popularity and ease of use: Mantel-Haenszel and SIBTEST.

Mantel-Haenszel

The Mantel-Haenszel procedure (MH) is the most commonly used method for detecting DIF because it is conceptually simple, relatively easy to use, and provides a chi square test of significance (Clauser & Mazor, 1998; Millsap & Everson, 1993). The MH procedure is used to detect DIF in dichotomous items.

As defined by Millsap and Everson (1993), the MH procedure compares the performance of the reference and focal groups on all the items in a given instrument, one item at a time. The group designated as the focal group is the group that is believed to be disadvantaged by the presence of DIF in the instrument. The group designated as the reference group serves as a comparison group for the purpose of DIF detection. The performance of comparable members of both groups is contrasted. Typically the examinees’ observed total score on the test is the matching variable for establishing comparability between the groups.

The MH procedure has two major limitations. First, studies have shown that the MH statistic does not detect nonuniform DIF because the procedure sacrifices sensitivity in order to achieve greater power for detecting uniform DIF (Holland & Thayer, 1988; Millsap & Everson, 1993; Swaminathan & Rogers, 1990a, 1990b; Uttaro, 1992). Uniform DIF exists when the probability of correctly answering an item is greater for one group than the other group uniformly over all levels of ability. Nonuniform DIF exists when the probability of answering the item correctly is not greater across all levels of ability for any one group, also indicating that there is an interaction between ability level and group membership (Zumbo, 2001). The mathematical procedure used by MH to summarize DIF across the various ability levels tends to cancel out or minimize the observed DIF when it is nonuniform. Second, as indicated by Millsap and Everson (1993), theoretical studies performed by Meredith and Millsap (1992) and Zwick (1990) and simulation studies by Uttaro (1992) have shown that when the item responses are generated by complex IRT models, the MH procedure sometimes falsely detects DIF when no DIF is present. This result is due to the use of the total observed score as a matching variable for establishing comparability instead of a latent trait and is more serious with shorter tests (fewer than 20 items). The possibility of falsely detecting DIF is minimized with longer tests.

SIBTEST

An alternative method for detecting DIF is the Simultaneous Item Bias Test (SIBTEST) proposed by Shealy and Stout (1993). SIBTEST is a conceptually simple method, currently growing in popularity, that includes a test of significance based on the ratio of the weighted difference in proportion correct (for reference and focal group members) to its standard error. The matching criterion is a latent trait measure rather than the observed total score used by the MH procedure, thus eliminiating one of the limitaitons of the MH procedure. Estimation of the matching latent trait includes a regression-based correction that has been shown to be useful in controlling Type I error (Clauser & Mazor, 1998; Gierl, Jodoin, & Ackerman, 2000; Roussos & Stout, 1996; Shealy & Stout, 1993). SIBTEST was originally intended for use with dichotomous test items, but has since been extended to handle polytomous items (items with multiple correct responses such as a Likert Scale or a constructed-response item) (Clauser & Mazor, 1998). POLYSIBTEST is the extended SIBTEST procedure designed for use with polytomous items.

The statistical hypotheses tested by SIBTEST are as follows:

$$H_0: B(T) = P_{Ref} - P_{Foc} = 0$$

vs.

$$H_1: B(T) = P_{Ref} - P_{Foc} \neq 0$$

B(T) is the difference in probability of a correct response on the studied item between examinees in the reference and focal groups matched on the measured latent trait *T*. P_{Ref} is the probability of a correct response on the studied item for participants in the reference group and P_{Foc} is the probability of a correct response on the studied item for participants in the focal group. *B(T)* is zero when there is no DIF present and nonzero when DIF is present (Gierl & Khaliq, 2001; Gierl, et al., 2000). The latent trait *T* is estimated separately for the reference and focal groups by using the equation for the linear regression of true score on observed score from classical test theory. *T* is adjusted using a regression correction technique to ensure the estimated latent trait is comparable for examinees of both the reference and focal groups. This adjustment is made by averaging the observed latent trait for the reference and focal groups. Finally, *B(T)* is estimated using $\hat{B}$, the weighted sum of differences between the proportion of correct true scores on the studied item for examinees in the two groups across all score levels (Gierl et al., 2000; Roussos & Stout, 1996).

Reducing DIF

Judgmental reviews of content and back-translation are two methods that can be used to help reduce DIF. Each will now be discussed in further detail.

Judgmental Reviews

After instrument items containing DIF have been identified, a judgmental review of the content of each item should follow the DIF analysis. In a judgmental review, reviewers are asked to study DIF items and propose possible reasons why these items are more difficult for one group of examinees compared to another (Allaoulf et al., 1999; Gierl & Khaliq, 2001). Reviews that yield interpretable results are essential for identifying items with translation differences and for controlling DIF in future adaptations of the test (Gierl et al., 1999).

Research has been conducted to help identify sources of DIF, thus narrowing the focus of a judgmental review of content. One such study carried out by Gierl and Khaliq (2001) involved a test administered to 6th and 9th grade students in the Canadian province of Alberta during 1997. The students were given the option to take the test in English or French. The test was originally constructed in English and translated into French.

After review of the results, an eleven member committee of testing specialists identified the following four sources (or explanations) of why the translated items may not behave the same as the original items:

1. Omissions or additions of words, phrases, or expressions that affect the meaning of an item
2. Words, expressions, or sentence structures that are inherent to one culture that do not have a direct parallel in another culture
3. Differences in words or expressions not inherent to language or culture
4. Format Differences such as, punctuation, capitalization, item structure, and typeface (pp. 32-33).

Allaoulf et al. (1999) performed a similar study using the *Psychometric Entrance Test* (PET), a high-stakes test used for admissions to universities in Israel (Belier, 1994). The PET is written in Hebrew and translated into five languages: (a) Arabic, (b) Russian, (c) French, (d) Spanish, and (e) English. Following an analysis of the Hebrew and Russian translation of the test, Allaouf et al. identified the following four causes of DIF:

1. *Changes in difficulty of words or sentences* - Even with accurate translation, some words or sentences became easier or more difficult. For example, an analogy item contained a very difficult word in the stem that was translated into a very trivial word. The translator was not aware of the difficulty of the original word, or of the importance of preserving that difficulty.

2. *Changes in content* - The meaning of the item changed in the translation, thus turning it into a different item. This could be due to an incorrect translation that changed

the meaning of the item or the translation of a word that has a single meaning into a word that has more than one meaning.

3. *Changes in format* - In some cases, changes in the format of the item were identified as the probable causes of DIF. For example, a shorter sentence became much longer. In another example concerning a translated sentence completion item, words that originally appeared only in the stem now appeared instead in all four alternative responses, thus making the item awkward. It should be noted that due to constraints of the Russian language, translating the item in this way could not be avoided.

4. *Differences in cultural relevance* - Differences in the relevance of item content to each culture was another source of DIF. In such cases, the item remained exactly the same but the two groups differed because of the cultural content of the specific item. This could be due, for example, to the content of a reading comprehension passage that was more relevant to one of the groups, or the content of a sentence completion item that was more familiar to one of the groups.

Back-translation

In addition to judgmental reviews of content, Gierl et al. (1999) suggest back-translation as an additional method of reducing DIF in adapted and translated tests. Back-translation is a popular and well-known judgmental method for evaluating the equivalence of two language forms (van de Vijver & Leung, 1997). In the basic design, the source language test is first translated into the target language, then back-translated into the source language by a different translator. The equivalence of the original source and target language forms is assessed by a reviewer or committee of reviewers who compare the original and back-translated source language forms for comparability in meaning (Brislin, 1970; Brislin, 1986; Gierl et al., 1999; Hambleton & Bollwark, 1991; Werner & Campbell, 1970).

Although back-translation and judgmental reviews may help reduce or explain the presence of DIF in translated and adapted instruments, time and resources may not always permit a thorough analysis to determine if DIF is controlled and if the validity of the construct the instrument items are intended to measure is preserved after translation and adaptation. It is important for instrument preparers and administrators to consider the possibility of DIF in instrument items and plan contingencies to ensure the items are measuring the constructs they were designed to measure after adaptation into different languages and cultures.

Summary

The use and adaptation of measurement instruments has been discussed as well as the importance of preserving validity of instrument items during adaptation, translation, and implementation. Examples of standards for instrument adaptation have been provided as well as a discussion on DIF analysis, including a description of the Mantel-Haenszel and SIBTEST procedures. Two methods of reducing DIF have also been discussed.

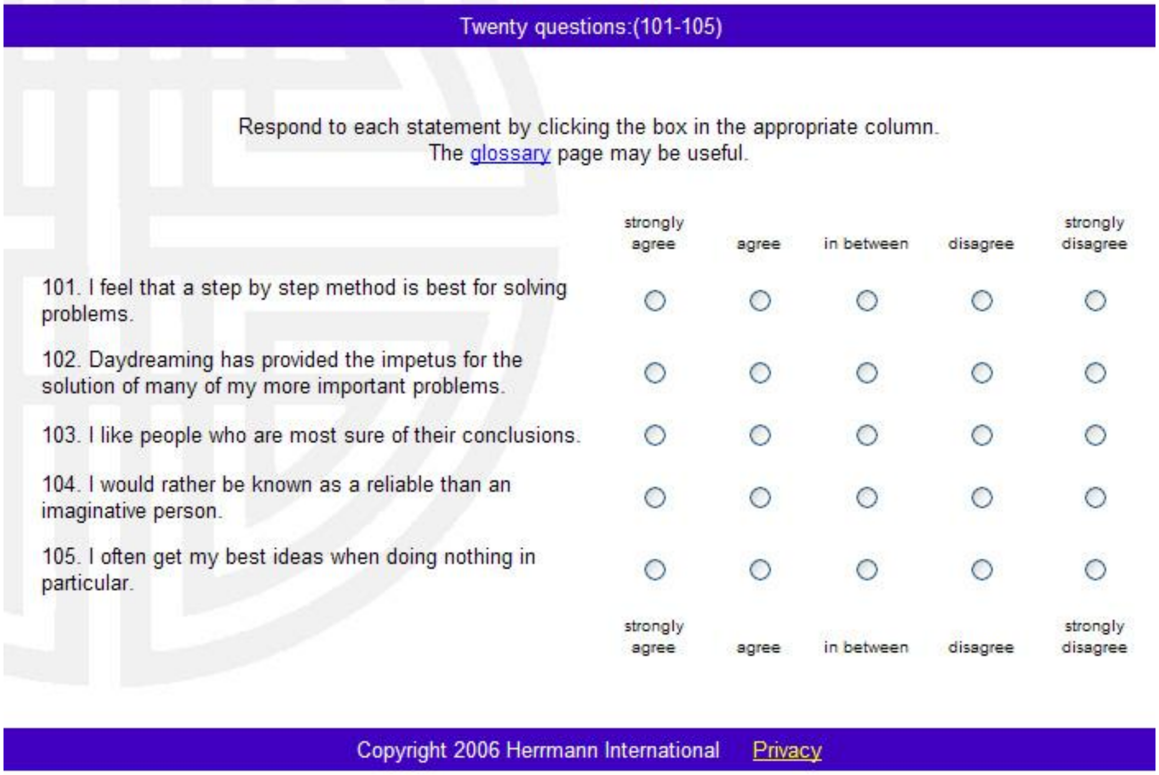
Chapter 3: Method

The purpose of this chapter is to provide information regarding the participants, details of the HBDI, and procedures used to conduct the DIF analysis. The guidelines used to classify the different levels of DIF detected are also presented.

Instrumentation

The HBDI contains dichotomous items and some polytomous items that are operationally converted into dichotomous items through the score key developed by Herrmann International. The DIF analysis was conducted using the POLYSIBTEST procedure, an extension of the SIBTEST procedure that detects DIF in both dichotomous and polytomous data. POLYSIBTEST will ensure all items were analyzed accurately, in case an item does behave as a polytomous item even after being operationally converted. This procedure was selected because of its (a) ability to control Type 1 errors, (b) ability to detect uniform and nonuniform DIF, and (c) simplicity of use (Clauser & Mazor, 1998; Gierl et al., 2000). The score key also classifies individual items by corresponding quadrants. Previous reliability and validity analysis has been conducted to ensure the items are measuring their intended constructs in the American English Language (Herrmann, 1994). Recent analysis has been conducted to verify reliability and validity of items on other language versions of the HBDI.

The HBDI is available by paper and electronic versions through Herrmann International. Information regarding the HBDI is available at www.hbdi.com. The HBDI consists of 120 items distributed among ten sections to profile the participants' dominant preferences. Figure 4 is an excerpt from the online version of the HBDI. Upon completion of the HBDI, results are sent to the participant along with a packet of information describing the participant’s dominant preferences. Out of the 120 items, 105 items are directly used to profile the preferences. These 105 items are the items analyzed in this study.



Twenty questions:(101-105)

Respond to each statement by clicking the box in the appropriate column.
The glossary page may be useful.

	strongly agree	agree	in between	disagree	strongly disagree
101. I feel that a step by step method is best for solving problems.	○	○	○	○	○
102. Daydreaming has provided the impetus for the solution of many of my more important problems.	○	○	○	○	○
103. I like people who are most sure of their conclusions.	○	○	○	○	○
104. I would rather be known as a reliable than an imaginative person.	○	○	○	○	○
105. I often get my best ideas when doing nothing in particular.	○	○	○	○	○
	strongly agree	agree	in between	disagree	strongly disagree

Privacy

Figure 4. Items 101 to 105 of the HBDI.

Participants

Two of the 18 language groups that are currently administered by the HBDI were analyzed in this study: American English and International English. Participants of the HBDI select for themselves the language version of the instrument they wish to take. The American English and International English language versions were selected for analysis because both language versions take the same test. Since there is no language translation present, potential DIF due to item translation will not be a factor and the study will focus on DIF as a result of cultural adaptation. American English participants are classified as the reference group since American English is the original language of the instrument. International English participants are classified as the focal group. Since participants select their own language group, it is unclear if the participants select a language based on their language spoken, their language learned, their country of residence, or for some other reason (such as language spoken as a second language). Generally it is considered that the participants who select American English currently speak and understand English as it is spoken and understood in the United States, and participants who select International English speak and understand English as it is spoken outside of the United States. The instrument items for each language group are identical.

Data from the American English and International English participants who completed the HBDI between January 2, 2003 and December 11, 2005 were selected for the DIF analysis. This archived data set included 77,170 American English participants and 40,952 International English participants. Almost all participants were employed at the time of taking the HBDI. The analysis was based on a simple random sample of 7,000 cases from each group, the maximum sample size allowed by the software used. Of the 7,000 American English sample, 3,769 were males and 3,231 were females. The average age of the participants in this sample was 35. Of the 7,000 International English sample, 3,566 were males and 3,434 were females. The average age of the participants in this sample was 24.

Analysis

The software package DIF PACKAGE 1.7, developed by Louis Roussos and William Stout along with some of their Ph.D students, was used to run the

POLYSIBTEST procedure. SPSS and Excel were used to manage and organize the data sets.

Procedure

HBDI Items were classified into four categories (A, B, C, and D) to correspond with their preference quadrant. Some of the items are *ipsative*, meaning that a positive preference for one quadrant will result in a negative preference in another quadrant (Sax & Newton, 1997). For example, an ipsative item may correspond to the A and C quadrants. If the examinee answers the item correctly for Quadrant A, the opposite response of the item will be scored automatically in Quadrant C without giving the examinee the opportunity to answer the item for Quadrant C. This effect may provide erroneous results in the DIF analysis because the ipsative score will interfere with the calculation of the latent trait used as the matching criterion because the item scores will cancel each other. In order to avoid erroneous results due to ipsative scores canceling scores in other quadrants, the items associated with each quadrant were analyzed separately. The assumption that analyzing the items keyed for one quadrant at a time for DIF, where evidence exists that each quadrant score is unidimensional, is a fundamental assumption behind this research method. With ipsative or near-ipsative scores spanning different quadrants, this assumption deserves additional investigation in future research.

Upon completion of the analysis, items that manifested DIF were classified into three categories representing different magnitudes of DIF following guidelines proposed by Roussos and Stout (1996, p. 220) and adopted by Gierl et al. (2000, p. 8):

1. Negligible or A-level DIF: Null hypothesis was rejected and $|\hat{B}| < 0.059$

2. Moderate or B-level DIF: Null hypothesis was rejected and $0.059 \leq |\hat{B}| <$ 0.088
3. Large or C-level DIF: Null hypothesis was rejected and $|\hat{B}| \geq 0.088$.

The cutoff values used by Roussos and Stout (1996) are an adaptation of similar classification guidelines developed by ETS after conducting extensive research. Roussos and Stout (1996) converted the cutoff values proposed by ETS for use with the SIBTEST procedure. $\hat{B}$ is an estimate of the amount of DIF present with the null hypothesis stating $\hat{B}$ = 0 (no DIF present). Items classified as Large or C-level are recommended for further review. Steps of the process for conducting the DIF analysis are outlined in Appendix A so the analysis may be repeated for additional language groups of the HBDI.

Chapter 4: Results

Research Objective 1

Table 1 identifies the proportion of items that function differentially between the American English and International English examinees. The items were classified according to the guidelines proposed by Roussos and Stout (1996).

Table 1

Proportion of HBDI Items That Show Evidence of DIF

	Quadrant			
Magnitude	A	B	C	D
Negligible	86.4%	70.0%	81.0%	81.2%
Moderate	4.5%	16.7%	14.3%	6.3%
Large	9.1%	13.3%	4.8%	12.5%
Total Number of Items	22	30	21	32

Research Objective 2

Tables 2 and 3 display the magnitude and direction of DIF detected and classification level for each item. For example, Item 1 in Quadrant A is classified as Negligible with a $\hat{B}$ = 0.029. This indicates the expected probability of correctly answering this item is 0.029 in favor of the reference group. A negative $\hat{B}$ indicates the item favors the focal group. $\hat{B}$ is an estimate of *B(T),* which was previously defined as the difference in probability of a correct response on the studied item between examinees in the reference and focal groups matched on the measured latent trait *T*. The items in Table 2 favor the reference group and the items in Table 3 favor the focal group.

Table 2

Classification and Magnitude of DIF Items that Favor the Reference Group

	Quadrant A		Quadrant B		Quadrant C		Quadrant D	
Magnitude	Item	$\hat{B}$	Item	$\hat{B}$	Item	$\hat{B}$	Item	$\hat{B}$
Negligible	1	0.029	2	0.000	2	0.005	2	0.005
	7	0.027	11	0.045	5	0.001	4	0.018
	9	0.042	12	0.050	13	0.002	5	0.035
	10	0.026	16	0.037	15	0.035	8	0.037
	12	0.028	20	0.043	17	0.018	10	0.008
	13	0.020	23	0.038	21	0.008	11	0.040
	14	0.009	25	0.010			14	0.000
	15	0.025	30	0.053			15	0.003
	18	0.008					16	0.009
	19	0.032					17	0.009
	21	0.001					18	0.041
							20	0.048
							29	0.045
							30	0.055
							31	0.022
Moderate			5	0.072	1	0.061	21	0.059
			8	0.074	8	0.083		
			21	0.075				
Large	3	0.097	3	0.096	7	0.135	1	0.110
			6	0.119			24	0.094
			24	0.163				

Table 3

Classification and Magnitude of DIF Items that Favor the Focal Group

	Quadrant A		Quadrant B		Quadrant C		Quadrant D	
Magnitude	Item	$\hat{B}$	Item	$\hat{B}$	Item	$\hat{B}$	Item	$\hat{B}$
Negligible	2	-0.026	1	-0.054	3	-0.004	3	-0.024
	5	-0.023	4	-0.046	4	-0.050	6	-0.028
	6	-0.012	9	-0.040	6	-0.049	9	-0.006
	11	-0.026	10	-0.021	10	-0.041	12	-0.023
	16	-0.042	13	-0.009	11	-0.021	13	-0.057
	17	-0.005	14	-0.014	12	-0.010	19	-0.022
	20	-0.001	15	-0.012	14	-0.014	22	-0.048
	22	-0.032	17	-0.019	16	-0.009	23	-0.040
			19	-0.053	18	-0.010	25	-0.026
			22	-0.042	19	-0.033	27	-0.031
			26	-0.032	20	-0.013	28	-0.005
			27	-0.009				
			28	-0.050				
Moderate	8	-0.075	7	-0.075	9	-0.068	32	-0.083
			29	-0.069				
Large	4	-0.107	18	-0.150			7	-0.151
							26	-0.090

Research Objective 3

The DIF analysis was completed in three stages, identified by the use of different software packages. In the first stage, the data were retrieved from Herrmann International and imported into SPSS in order to prepare the data for analysis. Data preparation included coding each member of the reference and focal groups, keying the item responses according to quadrant, and separating the data into datasets for Quadrants A, B, C, and D.

The second stage involved the use of Microsoft Excel. The purpose of this stage was to organize the item responses into a text file free from any character delimitation so that DIFPACK will properly read the data.

The data were analyzed using DIFPACK in the third stage. This stage also involved obtaining the results of the DIF analysis. A more detailed list of steps used to complete the three stages is found in Appendix A.

Chapter 5: Discussion

Conclusion

Of the 105 items in the HBDI, only 11 items were classified as *Large* in the amount of DIF present and are recommended for judgmental review of content. Items recommended for review in Quadrant A include 3 and 4. Items in Quadrant B to be reviewed include 3, 6, 18, and 24. Item 7 in Quadrant C and Items 1, 7, 24, and 26 in Quadrant D are also recommended for review. Overall, the items in the HBDI are functioning well with a low amount of DIF detected. Although only those items classified as *Large* are being recommended for review, items classified as *Moderate* may be reviewed for additional improvement to the HBDI. However, due to the low amount of DIF present in these items, the cost of a review may outweigh the benefit of further reducing DIF.

Item Review

Since the items are presented to examinees in the same language, a difference in cultural relevance, as described by Allaoulf et al. (1999), is the main plausible source of DIF. Gierl and Khaliq (2001) further describe differences in cultural relevance as words, expressions (idioms), or sentence structures that are inherent to one culture that do not have a direct parallel in another culture. Other sources as indicated by studies from Gierl and Khaliq (2001) and Allaoulf et al. (1999) are not relevant because they deal more with translation of items. Selected items classified as having *Large* amounts of DIF will now be presented. Additional items will be made available upon request.

Quadrant A, item 4. This item comes from the Work Elements section of the HBDI. The instructions for this item are as follows:

> Rate each of the work elements below according to your strength in that activity, using the following scale: 5 = work I do best; 4 = work I do well; 3 = neutral; 2 = work I do less well; 1 = work I do least well. Enter the appropriate number next to each element. Do not use any number more than four times.
>
> 4. ___ financial aspects

Item 4 favors the reference group with a $\hat{B}$ = -0.107. The source of DIF is unclear. A cultural difference in interpretation of this item is one possible source of DIF. The two groups may be interpreting the meanings differently. Another possible source of DIF is the participants in one group may lack confidence in their ability to work with financial aspects. This difference in self-confidence in doing financial work could affect the selection of this item despite the trait level the participants possess for this item.

Quadrant B, item 24 and Quadrant D, item 26. This is the same item. One response is scored as Quadrant B and the other response is scored as Quadrant D. In Quadrant B, the item favors the reference group with a $\hat{B}$ = 0.163. In Quadrant D, the item favors the focal group with a $\hat{B}$ = -0.09. The instructions for this item are as follows:

> Respond to each statement by checking the box in the appropriate column.
>
> 24. I sometimes get a kick out of breaking the rules and doing things I am not supposed to do.

Strongly Agree	Agree	In Between	Disagree	Strongly Disagree
□	□	□	□	□

One possible source of DIF in this item is the idiom “get a kick out of breaking the rules”. In the American English culture, this idiom is used as a slang term to indicate that some sort of gratification is achieved by breaking the rules. In other cultures, the idiom may not exist and the word “kick” could indicate a punishment is received by breaking the rules. Thus the idiom could change the intended meaning of this item depending on how the examinee interprets it.

In addition to a possible misinterpretation of wording, cultural appearance is another possible explanation of DIF in this item. Breaking rules may be a more acceptable behavior in one culture than another. This would result in a response that reflects how the examinee wants to be perceived by society instead of a response that is related to the construct.

Quadrant C, item 7. Item 7 in Quadrant C favors the focal group with a $\hat{B} =$ 0.135. Below is the item description:

> Select eight adjectives which best describe the way you see yourself. Enter a 2 next to each of your eight selections. Then change one 2 to a 3 for the adjective which best describes you.
>
> 7. __ spiritual

The HBDI glossary defines *spiritual* as having to do with spirit or soul as apart from the body or material things (Herrmann, 2007). Although the word spiritual is defined, the examinee is not required to use the glossary, and thus may rely upon his or her own interpretation of the meaning of this word. The fact that the word is presented outside of any specific context probably increases the likelihood that different examinees will interpret it differently. Also, in this section of the HBDI, there are 25 items to select from.

Item 7 may be incorrectly omitted by the examinee simply because of the appeal of eight other items (the directions indicate to select only eight items).

Quadrant D, item 7. Item 7 in Quadrant D is found in the same section as Item 7 in Quadrant C. The item is as follows:

> Select eight adjectives which best describe the way you see yourself. Enter a 2 next to each of your eight selections. Then change one 2 to a 3 for the adjective which best describes you.
>
> 7. __ holistic

Item 7 in the D Quadrant favors the reference group with a $\hat{B}$ = -0.151. Possible explanations for DIF found in this item parallel the explanation given for Item 7 in Quadrant C.

Overall, the adaptation of the HBDI into the International English language group is successful. Only 11 of the 105 items contain sufficient DIF to warrant further review.

Recommendations

As a result of the review of *Large* magnitude DIF items, Herrmann International should consider the following recommendations:

1. Encourage the examinees to use the glossary.
2. Avoid using idioms.
3. Identity items through internal validity analysis followed by judgmental content reviews that are not measuring intended constructs and create and evaluate additional items that will serve as replacements.

Closing Remarks

With the increased use of measurement instruments across different languages and cultures, instrument developers and administrators have a responsibility to promote the sound and ethical use of all types of measurement instruments and to provide a basis for evaluating the quality of testing practices (AERA, 1999; Allaoulf et al., 1999; Gierl & Khaliq, 2001; Hambleton et al., 2005). Herrmann International is adhering to this principle through evaluating the performance and content of the HBDI. Since DIF is a plausible threat to the validity of the constructs the HBDI items are intended to measure, a DIF analysis adds value to Herrmann International's desire to improve the HBDI (Clauser & Mazor, 1998; Gierl, Rogers, & Klinger, 1999; Hambleton et al., 2005). The procedures used to carry out the DIF analysis in this study can easily be replicated so that Herrmann International may conduct DIF analysis for other language versions of the HBDI.

The results of this study indicate it is possible for Herrmann International to perform a DIF analysis on the HBDI, even with ipsative items. The procedures used provide the necessary steps to perform a DIF analysis on subsequent versions of the HBDI. It is important to remember that ordinarily, the purpose of a DIF analysis is to identify items that appear to give an advantage to members of one group over another group. With preference traits, advantage or disadvantage is not an important issue. The real issue is whether substantial DIF is a signal that construct-irrelevant variance exists in the item's meaning as perceived by the two groups. By design, the item is intended to measure the same trait the same way in both groups. DIF statistics alone do not determine if the item is measuring its intended construct, but they can flag items for further review.

Schedule

Table 4 defines the schedule of the study. The duration of the study was extended to compensate for vacation and work schedules of myself and the committee members.

Budget

Materials, including data for analysis and SPSS, were made available at no cost by Dr. C. Victor Bunderson of the EduMetrics Institute as a result of the validation work EduMetrics has been conducting for Herrmann International. In this work, EduMetircis is helping Herrmann International analyze other reliability and validity aspects of the HBDI. The Harold B. Lee Library (HBLL) at Brigham Young University (BYU) provided reference materials at no charge. DIFPACK and miscellaneous printing supplies were the only monetary expenses incurred for the study. Table 5 lists the budgeted and actual expenses. A grant from Applied Psychological Measurement Inc. (APM) was used to purchase DIFPACK. APM included an additional $15 in the grant to cover for unanticipated shipping costs. Labor incurred an additional five hours of cost because of incompatibility issues between DIFPACK and SPSS. Microsoft Excel was used to format the data to conform to DIPACK’s data format.

Table 4

Schedule

Event	Participants	Delivery
Discuss possible ideas for study	Jared Lees Dr. C. Victor Bunderson Dr. Van Newby	October 2006
Begin preliminary research	Jared Lees	October 2006
Initial meeting to begin proposal	Jared Lees Dr. Richard Sudweeks	January 2, 2007
End preliminary research	Jared Lees	January 2, 2007
Begin proposal and literature review	Jared Lees	January 2, 2007
Progress report	Jared Lees Dr. Richard Sudweeks	January 22, 2007
Progress report	Jared Lees Dr. Richard Sudweeks	March 20, 2007
Progress report	Jared Lees Dr. Richard Sudweeks	March 26, 2007
Progress report	Jared Lees Dr. Richard Sudweeks	May 2, 2007
End proposal and literature review	Jared Lees	May 8, 2007
Progress report	Jared Lees Dr. Richard Sudweeks	May 8, 2007
Schedule proposal defense	Jared Lees	May 8, 2007
Proposal defense	Jared Lees Dr. Richard Sudweeks Dr. C. Victor Bunderson Dr. Charles Graham	May 23, 2007
Submit proposal revisions	Jared Lees	June 21, 2007
Begin study / data analysis	Jared Lees	June 22, 2007
End study / data analysis	Jared Lees	July 6, 2007
Schedule final defense of study	Jared Lees	July 27, 2007
Defense of study	Jared Lees Dr. Richard Sudweeks Dr. C. Victor Bunderson Dr. Charles Graham Dr. Diane Strong-Krause	August 15, 2007
Submit revisions	Jared Lees	August 24, 2007

Table 5

Budget

Expense	Estimated Cost	Actual Cost
DIFPACK Student Version	$200	$215
SPSS	$0	$0
Paper and print supplies	$15	$15
Labor (20 hours at $15/hour)	$300	$375*
Total	$515	$605

*Actual labor cost was based on 25 hours

References

Allaoulf, A., Hambleton, R. K., & Sireci, S. G. (1999). Identifying the causes of DIF in translated verbal test items. *Journal of Educational Measurement, 36,* 185-198.

American Educational Research Association, (1999). *Standards for educational and psychological testing.* Washington, D.C. American Educational Research Association.

Anastasi, A., & Urbina, S. (1997). *Psychological testing* (7th ed.). Upper Saddle River, NJ: Prentice-Hall.

Beller, M. (1994). Psychometric and social issues in admissions to Israeli universities. *Educational Measurement: Issues and Practice, 13*(2), 12-21.

Brislin, R. W. (1970). Back-translation for cross-cultural research. *Journal of Cross-Cultural Psychology, 1,* 185-216.

Brislin, R. W. (1986). The wording and translation of research instruments. In W. J. Lonner & J. W. Berry (Eds.), *Field methods in cross-cultural research* (pp. 137-164). Newbury Park, CA: Sage.

Camilli, G., & Shepard, L. A. (1994). *MMSS: methods for identifying biased test items.* Thousand Oaks, CA: Sage.

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

Faggen, J. (1987). Editorial. *Educational Measurement: Issues and Practices, 9*(2), 5-8.

Gierl, M. J. & Khaliq, S. N. (2001). Identifying sources of differential item and bundle functioning of translated achievement tests: A confirmatory analysis. *Journal of Educational Measurement, 38,* 164-187.

Gierl, M. J., Jodoin, M. G., & Ackerman, T. A. (2000, April). *Performance of Mantel-Haenszel, Simultaneous Item Bias Test, and Logistic Regression when the proportion of DIF items is large.* Paper presented at the annual meeting of the American Educational Research Association, New Orleans, LA.

Gierl, M. J., Rogers, W. T., & Klinger, D. A. (1999). Using statistical and judgmental reviews to identify and interpret translation DIF. *Alberta Journal of Educational Research, 45*, 353-376.

Hambleton, R. (1994). Guidelines for adapting educational and psychological tests: A progress report. *European Journal of Psychological Assessment, 10,* 229-244.

Hambleton, R. K. & Bollwark, J. (1991). Adapting tests for use in different cultures: Technical issues and methods. *Bulletin of the International Testing Commission, 18,* 3-32.

Hambleton, R. K., Merenda, P. F., & Spielberger, C. D. (Eds.). (2005). *Adapting educational and psychological tests for cross-cultural assessment.* Mahwah, NJ: Lawrence Erlbaum Associates, Inc.

Haney, W. M., & Reidy, E. F. (1987). Editorial. Educational Measurement: Issues and Practices, 9(2), 4.

Herrmann International products and services (2007). Retrieved May 12, 2007, from http://www.hbdi.com/WholeBrainProductsAndServices/index.cfm

Herrmann, N. (1994). *The creative brain.* Kingsport, TN: Quebecor Printing Group.

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.

Meredith, W., & Millsap, R. E. (1992). On the misuse of manifest variables in the detection of measurement bias. *Psychometrika, 57*, 289-311.

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

Millsap, R. E., & Meredith, W. (1992). Inferential conditions in the statistical detection of measurement bias. *Applied Psychological Measurement, 16,* 389-402.

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

Sax, G., & Newton, J. W. (1997). *Principles of educational and psychological measurement and evaluation* (4th ed.). Belmont, CA: Wadsworth Publishing Co.

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

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

Swaminathan, H., & Rogers, H. J. (1990b, April). *A comparison of the logistic regression and Mantel- Haenszel procedures for detecting differential item functioning.* Paper presented at the meeting of the American Educational Research Association, Boston MA.

Uttaro, T. (1992). *Factors influencing the Mantel-Haenszel procedure in the detection of differential item functioning.* Unpublished doctoral dissertation, Graduate Center, City University of New York.

van de Vijver, F, & Leung, K. (1997). *Methods and data-analysis for cross-cultural research.* Thousand Oaks, CA: Sage.

Werner, O, & Campbell, D. T. (1970). Translating, working through interpreters, and the problem of decentering. In R. Naroll & R. Cohen (Eds.), *A handbook of method in cultural anthropology*. New York, NY: Columbia University Press.

Zumbo, B. D. (2001, April). *Investigating DIF by statistical modeling of the probability of endorsing an item: logistic regression and extensions thereof.* Paper presented at the National Council for Measurement in Education.

Zwick, R., Donoghue, J. R., & Grima, A. (1993). Assessing differential item functioning in performance tasks. *Journal of Educational Measurement, 30,* 233-251.

Appendix A: DIF Analysis Procedure

The following steps outline the procedure used to complete the DIF analysis for this study. Herrmann International may follow these steps to conduct DIF analyses on other language groups of the HBDI.

Steps

SPSS

1. Acquire original SQL data file from Herrmann International
2. Sort cases by Flag 136 = 1 (for AE)
3. Randomly select 7000 cases
4. Run coding syntax developed by EduMetrics to code questions into Key 2
5. Separate questions into A, B, C and D quadrants
6. Create a new dataset for each quadrant
7. Repeat steps 1 through 6 for Flag 136 = 2 (Intl English)
8. Import each of the 8 datasets into Excel

Excel

These steps are necessary to format the data into a file that can be read by DIFPACK

1. For each dataset, strip out all of the variable names and additional data so only item responses remain
2. Remove all decimals – DIFPACK cannot read in data with decimals
    a. Quadrants B and D contain an item set with some values that equal 0.5.
    b. Separate these items and create a subset for quadrant B and D
    c. Multiply all values in the subset by 2 to remove values that equal 0.5.

3. Use the CANCATENATE function in Excel to combine all of the item responses into one cell. The result should be one column with 7000 rows (1 row for each case/participant)
4. Search for NULL values using the FIND command in Excel. Replace the NULL values with 0. This takes care of any missing data
5. Copy the row into a new worksheet
6. Save the new worksheet as a Text (MS-DOS) file – this file can be read into DIFPACK

DIFPACK

1. Load the Text (MS-DOS) files into DIFPACK for the Reference group and Focal group of one quadrant
    a. For example, quadrant A for American English and quadrant A for International English
2. Verify the data using the Verify button in DIFPACK. Each dataset should have 7000 Examinees. Quadrant A has 22 items. Quadrant B has 17 items. Quadrant B1 (subset of B) has 13 Items. Quadrant C has 21 items. Quadrant D has 19 Items. Quadrant D1 (Subset of D) has 13 Items
3. Identify the Reference and Focal Groups (For this study, the reference group is American English and the Focal Group is International English)
4. Specify the output file
5. Select each item as a Suspect Item
6. Select “Test each SI separately”
7. Run the POLYSIBTEST program leaving all other options as default

8. View output