CROSS-CULTURAL EQUIVALENCE OF BECK DEPRESSION INVENTORY-II (BDI-II) 
FOR PAKISTANI CANADIAN IMMIGRANTS AND OTHERS

by

Rida Waseem

A THESIS SUBMITTED IN PARTIAL FULLFILLMENT OF 
THE REQUIREMENTS FOR THE DEGREE OF 
MASTER OF ARTS 
in 
The Faculty of Graduate and Postdoctoral Studies 
(Measurement, Evaluation, and Research Methodology)

THE UNIVERSITY OF BRITISH COLUMBIA 
(Vancouver)

October 2017

© Rida Waseem, 2017
Abstract

The purpose of the study was to demonstrate the measurement equivalence of one of the most widely used depression measures, the Beck Depression Inventory-II (BDI-II) for the Canadian Pakistani immigrants and other Canadians. The data were collected from 400 university students in the Greater Toronto Area (GTA). To examine the measurement equivalence of the BDI-II, the data were analyzed using a multi-group confirmatory factor analysis (MG-CFA) and differential item functioning (DIF) analysis. The results of MG-CFA showed evidence of configural and metric invariance, however, no sufficient evidence for scalar invariance was found. The DIF analysis identified differential response patterns in three items for the two comparison groups. The expert reviews of items provided potential sources of inequivalence. The CFA, DIF analysis and expert reviews jointly provided evidence of inequivalence which suggests, caution should be taken when comparing depressive symptoms for Pakistani to non-Pakistani populations in Canada based on BDI-II. The results from the study are expected to help researchers in modifying and developing new measures for Pakistani or other South Asian ethnic groups.
Lay Summary

Researchers and clinicians often use psychological tests to assess individuals for mental health issues. Often, the test is used on all individuals regardless of ethnic background, assuming its validity and appropriateness. For example, the same test is often used for diverse ethnic groups to measure depression, assuming all groups have similar expressions of depression. It’s important to note that ethnic groups express depression in different ways and therefore, the same test may not be valid in assessing individuals from different cultural backgrounds. Therefore, this study examined whether a test (BDI-II) measuring depression could provide comparable information for an ethnic minority group as it does for the mainstream population in Canada. The results showed that the BDI-II does not provide comparable measurements of depression for the Pakistani Canadians. The findings from this study can help clinicians and researchers in using and developing appropriate tests for assessing different ethnic populations, including Pakistani Canadians.
Preface

The thesis is original, unpublished, and independent work of the author, R. Waseem. The thesis work was approved by University of British Columbia’s Behaviour Research Ethics Board [Certificate number: H16-01195] with title ‘Measurement Equivalence of Beck Depression Inventory-II (BDI-II)’. 
Table of Contents

Abstract.............................................................................................................................................. ii
Lay Summary ......................................................................................................................................... iii
Preface................................................................................................................................................ iv
Table of Contents .................................................................................................................................. v
List of Tables ......................................................................................................................................... vii
List of Figures ........................................................................................................................................ viii
List of Abbreviations .......................................................................................................................... ix
Acknowledgement ..................................................................................................................................... x
Dedication ................................................................................................................................................ xi

Chapter 1: Introduction ......................................................................................................................... 1
  1.1 Expression of Depression .................................................................................................................. 2
  1.2 Measurement Invariance ................................................................................................................... 4
  1.3 Purpose of Study ............................................................................................................................... 6
  1.4 Significance of Study ....................................................................................................................... 7

Chapter 2: Literature Review .................................................................................................................. 8
  2.1 Types of Equivalence ...................................................................................................................... 8
  2.2 Causes of Inequivalence .................................................................................................................. 9
  2.3 Expert Review ................................................................................................................................ 11
  2.4 Measurement Equivalence in Depression Measures ......................................................................... 12

Chapter 3: Method .................................................................................................................................. 18
  3.1 Measures ......................................................................................................................................... 18
  3.2 Participants and Procedure .............................................................................................................. 20
3.3 Internal Consistency........................................................................................................ 21
3.4 Confirmatory Factor Analysis.......................................................................................... 22
3.5 DIF Analysis..................................................................................................................... 23
3.6 Expert Review.................................................................................................................. 30

Chapter 4: Results.................................................................................................................. 32
4.1 The Sample ....................................................................................................................... 32
4.2 Reliability Analysis........................................................................................................... 33
4.3 Factor Analysis .................................................................................................................. 34
4.4 DIF Analyses .................................................................................................................... 35
4.5 Expert Reviews .................................................................................................................. 43
4.6 Summary ............................................................................................................................ 45

Chapter 5: Discussion................................................................................................................. 46
5.1 Limitations .......................................................................................................................... 49
5.2 Implications and Future Research ....................................................................................... 50

References..................................................................................................................................... 52

Appendices.................................................................................................................................. 65

Appendix A: Questionnaire ....................................................................................................... 65
Appendix B: Professional Profile and Feedback Form .............................................................. 67
Appendix C: Referral List for Resources in Greater Toronto Area ............................................. 70
Appendix D: Informed Consent ................................................................................................... 71
List of Tables

Table 4.1 Characteristics of the Study Sample.................................................................33
Table 4.2 Result for Measurement Invariance across groups........................................35
Table 4.3 Result for DIF detection by Ordinal Logistic Regression.................................36
Table 4.4 Result for DIF detection by Mantel-Haenszel...............................................38
Table 4.5 Total Variance for Focal group.....................................................................39
Table 4.6 Total Variance for Reference group...............................................................40
Table 4.7 Result for DIF detection by IRT analysis.......................................................42
Table 4.8 Items identified as DIF by three methods.....................................................43
List of Figures

Figure 4.1 Item Characteristic Curve of item 5 ................................................................. 36
Figure 4.2 Item Characteristic Curve of item 6 ................................................................. 37
Figure 4.3 Item Characteristic Curve of item 11 ................................................................. 37
List of Abbreviations

AERA  American Educational Research Association
APA   American Psychological Association
BDI-II Beck Depression Inventory-II
CFA   Confirmatory Factor Analysis
CFI   Comparative Fit Indices
DIF   Differential Item Functioning
EFA   Exploratory Factor Analysis
ICC   Item Characteristic Curve
IRT   Item Response Theory
ITC   International Test Commission
LR    Logistic Regression
MH    Mantel-Haenszel
NCME  National Council of Measurement in Education
OLR   Ordinal Logistic Regression
PCA   Principle Component Analysis
RMSEA Root Mean Square Error of Approximation
SRMR  Standardized Root Mean Square Residual
Acknowledgement

The thesis would not have been possible without continuous support and guidance of my supervisor, Dr. Kadriye Ercikan.

I am also thankful to my committee members, Dr. Robinder Bedi and Dr. Amery Wu, for their valuable feedback and advice.
This thesis is dedicated to the memory of Prof. Dr. Hamid Sheikh (1939-2015). His courses in Psychological Testing and Experimental Psychology inspired me to choose the MERM Program and eventually, complete this research study.
Chapter 1: Introduction

The number of immigrants choosing to make Canada their home is continuously growing (Statistics Canada, 2015). Immigrants who came to Canada before the 1970s were predominantly European. However, over time immigration patterns have changed significantly, with a greater influx of immigrants from Asia, Middle East, and Africa (Aycan & Berry, 1996).

While this influx has supported the labour needs of the Canadian economy, it also poses many challenges, some immigrants suffer from issues associated with psychological wellbeing, for example anxiety, depression and post-traumatic stress disorder (PTSD) (Aycan & Berry, 1996).

According to Kirmayer (2001) certain groups of immigrants are more vulnerable to mental health issues, especially those who face prosecution, violence, and exposure to war. Additionally, racism and discrimination in the host country can also increase the risk of mental health issues (Kirmayer, 2001).

Despite being at risk of mental health issues, immigrants are less likely to seek help than their Canadian born counterparts (Kirmayer, 2001). Though many studies (Ali, McDermott, & Gravel, 2004; Beiser, 2009; Hyman, 2004) have shown that upon arrival immigrants may have better mental health profiles than Canadian-born individuals, this argument cannot be applied to all immigrant populations; it depends upon circumstances such as experience to trauma, war, and prosecution (Stafford, Newbold, & Ross, 2011). Another reason that immigrants from certain ethnic groups may not make use of mental health services is due to a difference in the theoretical model of the Western health system, which sometimes fails to capture the beliefs and traditions of diverse immigrant groups (Lloyd, Pouwer, & Hermanns, 2012). These differences are reflected in many of the popular mental health assessment tools that inadequately reflect the
unique experiences of diverse ethnic groups (Islam, Khanlou, & Tamim, 2014; Lloyd et al., 2012).

1.1 Expression of Depression

Many studies have shown that depression is expressed differently in different ethnic groups (Islam et al., 2014; Kleinman, 1982). For example, depression has been shown to be expressed in affective ways in those who identify with Western culture, while in non-Western cultures depression is often expressed somatically (Islam et al., 2014). This somatization has also been observed in the South Asian population, which has even lead to the development of a separate measure, the Bradford Somatic Inventory (BSI). The BSI was developed using a Pakistani patient with a clinical diagnosis of depression. This assessment tool was developed out of a need to measure somatic symptoms of depression and anxiety (Mumford, Bavington, Bhatnagar, Hussain, & Naraghi, 1991).

In the study by Siddiqui and Shah (1997), a depression measure was developed and validated for the Pakistani population also observed the similar trends of somatic expression. In the phase of item generation, the authors asked the university student participants to recall depressive situations and the feelings associated with it. It was observed that a significant portion of statements recalled were about the depressive situations regarding bodily function such as loss of appetite, sleep disturbances and fatigue. To validate the reported somatic symptomatology of depression, the author also consulted psychiatrists working with the clinical population diagnosed with mental health issues; the psychiatrists also reported a significant proportion of symptoms that were somatic in nature (Siddiqui & Shah, 1997). Additionally, they also observed that few symptoms which are usually indicators of depression were not identified by the Pakistani population. For instance, suicidality was not reported as feelings associated with
depression by the participants in that study which might be due to religious orientation of the Pakistani population where most are Muslim and Islam, religion that discourage suicide or harming self by any means (Siddiqui & Shah, 1997). It was also observed that no item related to sex was obtained from the student’s statements. This might be due to conservative nature of the Pakistani culture in which a discussion about sexual topics is often considered a taboo. However, it can not be denied that depression don’t effect the sexual desires. It should also be noted that the participants were unmarried university students which affected how they respond to questions about sex (Siddiqui & Shah, 1997).

The literature also showed that in non-Western cultures, mental health issues are often attributed to the mystical or religious explanations. There is a tendency to associate depression and sadness with religious forces such as punishment from God. This has been showed by Haddad, Waqas, Qayyum, Shams and Malik (2016)’s study. In the study, a survey was conducted using a Revised-Depression Attitude Questionnaire (R-DAQ) with 700 non-psychiatric medical professionals in Lahore, Pakistan. It was found that one third of the participants reported that supernatural powers including a punishment from God as a cause of depression (Haddad et al., 2016). This showed that there is a tendency to associate ‘perceived’ bad deeds with punishment from God, which also makes them feel guilty. These guilty feelings are usually associated with feelings of punishment and hence they are related to religious orientation of the Pakistani people (Siddiqui & Shah, 1997).

The above-mentioned studies showed that some of the differential symptoms of depression reported by South Asian ethnic groups, including ability to endorse items related to guilt/punishment and tendency to describe depression with somatic symptoms. This signifies the importance of using an assessment tool that has cultural equivalence; otherwise it might threaten
the validity of the interpretations made using the test. The finding that it is important to ensure psychometric equivalence of the tests prior to their use on a different ethnic population has been argued consistently by measurement researchers. Despite its importance, researchers often ignore this requirement. In the case of mental health disorders like depression, if measures are used without establishing conceptual and measurement equivalence then the interpretations made using that measure will not be valid, and might affect the decisions made using those interpretations (Cuéllar & Paniagua, 2000).

1.2 Measurement Invariance

There has been an abundance of research on South Asian immigrant’s mental health issues, yet there have been very few diagnostic tools that were developed or validated using essential testing standards.

Additionally, research has demonstrated that many different existing depression measures have been used on South Asian ethnic groups but these studies do not provide evidence for its cultural equivalence on the target population therefore, the lack of evidence in these studies make the findings invalid and limits it dissemination.

Researchers and clinicians often use qualitative methods (e.g. focus groups, interviews) to assess and diagnose mental health conditions because there has been limited quantitative measure available to use with South Asian population. These qualitative methods are often time consuming and require subjective interpretation of responses (Islam et al., 2014).

Efforts have been made to translate assessment tools to meet the needs of diverse ethnic or cultural groups. However, these translated measures have not been shown to be sufficient in understanding ‘cultural meaning’ and ‘cultural differences’ (Van de Vijver & Tanzer, 2004). The translated measures, for instance, do not give much information if the construct does not have the
same meaning for the comparison groups (Islam et al., 2014; Van de Vijver & Tanzer, 2004). Language involves unique concepts and thoughts which may not have similar meaning in the language to which it is being translated. Using a translated measure without addressing cultural equivalence can result in biased and invalid decisions (Lloyd et al., 2012).

The Standards for Educational and Psychological Testing (AERA, APA, & NCME, 2014) also stresses the importance of using assessment tools that are fair and reduce bias when applying it to a diverse range of individuals and subgroups in an intended population. According to Standards, it is also a responsibility of test administrators to ensure validity of score interpretations for the intended population. Similarly, the American Psychological Association’s (2002) ethical principles of Psychologists and the code of conduct also stresses the importance of using psychometrically sound assessment tools for assessing ethnic minorities (Okazaki & Sue, 1995). In psychological research, studies often compare different cultural groups on psychological constructs, yet disregard the psychometric properties of the measure and equivalence of construct being measured. According to Nuevo et al. (2009), such studies tend to assume that a psychological tool measures the same construct in all ethnic groups. However, this practice as stated by Nuevo et al. (2009), ‘ignores the impact that culture may have on the meaning of scale items and its internal structure’ (p. 157).

Vandenberg and Lance (2000) also suggested that the meaningful comparison can only be made when a measure shows measurement equivalence for groups being compared. The measure is said to have measurement equivalence when the same latent construct (such as depression) is measured in the same way across groups. Horn and McArdle (1992) also commented on the importance of using tools that have measurement equivalence:
‘The general question of invariance of measurement is one of whether or not, under different conditions of observing and studying phenomena, measurements yield measures of the same attributes. If there is no evidence indicating presence or absence of measurement invariance—the usual case—or there is evidence that such invariance does not obtain, then the basis for drawing scientific inference is severely lacking: findings of differences between individuals and groups cannot be unambiguously interpreted’ (p. 117).

If measures are used without establishing equivalence, then the interpretations made based on the results, will affect validity and may compromise the decisions based on those results (Cuéllar, & Paniagua, 2000). For example, if a depression measure does not show equivalence then it will impact the diagnosis as well as treatment for a person.

1.3 Purpose of Study

Considering the lack of research on the cross-cultural equivalence of depression tools, the present study examined the equivalence of one of the most widely known measures of depression, the Beck Depression Inventory-II (BDI-II). This study focused on measurement equivalence of BDI-II for the Pakistani-Canadian population in the Greater Toronto Area by comparing their responses with the non-Pakistani population as a reference group. The Pakistani population had been chosen for the research was due to an overwhelming lack of research regarding the measurement equivalence and cultural relevance of using depression measures. It is important to note that data was mostly collected from first generation or recent Pakistani immigrants. This is because first generation or recent immigrants are more likely to retain their cultural and family values compare to second generation or older immigrants. The second generation usually identifies with host country culture (Furnham & Shiekh, 1993).
This study answered the following research question:

Is there evidence of measurement equivalence of BDI-II for the Pakistani Canadian population?

To investigate this question, data was collected from students in the Greater Toronto Area (GTA) universities. Following data collection, CFA and DIF analyses were conducted which determined the equivalence of BDI-II for the Pakistani Canadian population. In addition, expert review was conducted to examine the cultural relevance of BDI-II items.

1.4 Significance of Study

The findings from this study provide crucial information about the measurement equivalence of BDI-II for the Pakistani population and whether it is a culturally appropriate measure for assessing severity of depression in Pakistani population in Canada. In addition, the review from experts help to shed light on some of the possible reasons for inequivalence. This study can also inform future research efforts concerned with developing and/or modifying existing mental health measures for the Pakistani population. As well, findings can inform researchers on other South Asian populations, since South Asian population share similar cultures, customs and sometimes language. However, careful consideration should be taken before applying to other South Asian communities since there are some cultural and linguistic variances within each South Asian community (Clarke, Peach, & Vertovec, 1990). This study can provide evidence for the need to further test the equivalence of psychological measures on other racially or ethnically diverse groups.
Chapter 2: Literature Review

In this chapter, the following three types of measurement equivalence are discussed: configural, metric and scalar equivalence. This is followed up with review of research on expert reviews and a review of literature on examining equivalence of depression measures, specifically the Beck Depression Inventory II, across different cultural groups. The emphasis is given to the statistical methods use to address equivalence in the cross-cultural comparison of the Beck Depression Inventory II or other depression measures.

2.1 Types of Equivalence

In this section three common types of equivalence are described: configural, metric and scalar equivalence (Milfont, & Fischer, 2015).

**Configural Equivalence.** Configural Equivalence is also known as configural invariance. It is considered a basic type of equivalence. This type of equivalence requires identical factors, i.e., whether there are same number of factors and same set of items that make up each factor for each group (Hambleton, Merenda, & Spielberger 2005). The configural equivalence is examined by constraining the number of factors and the loading patterns equal across groups. If configural invariance is met, it means that the same general concept is measured across the groups (Vandenberg & Lance, 2000). The metric invariance is tested as next level of measurement invariance.

**Metric Equivalence.** Metric equivalence, otherwise known as weak factorial invariance or metric invariance, is a type of equivalence used to determine if a measure has the same meaning and factors across groups being compared (Brown, 2006). The metric equivalence can be examined by constraining item-factor loadings to be equal and fitting the factor model to the data for each group. If metric invariance is met, it means that individuals from each group
responded to items in a similar way (Brown, 2006). The scalar or strong invariance is tested as next level of measurement.

**Scalar or Full Score Equivalence.** This type of equivalence requires the same factor loadings and equal intercepts across the group. This is considered the highest level of equivalence and assumes a bias free measurement (Vandenberg & Lance, 2000). The scalar equivalence can be examined by constraining and fitting the model to the data from both groups simultaneously. If there is a fit, it suggests that the model constraints are consistent with the data. The scalar equivalence can also be tested by comparing the fit of the metric and scalar models (Brown, 2006). If the scalar invariance is met, then this means that the same level of construct has the same score for items. It is important to ensure scalar invariance for cross cultural comparisons. If there is no evidence of scalar invariance then this indicates the presence of potential bias in responding to items (Vandenberg & Lance, 2000).

### 2.2 Causes of Inequivalence

Bias and equivalence are closely related concepts. It is important to discuss bias when examining equivalence because bias affects equivalence. What follows is an overview of the three main sources of bias: construct bias, method bias, and item bias (Van de Vijver & Tanzer, 2004).

**Construct Bias.** The construct bias can also be defined as the inequivalence of the construct between the cultures. For example, intelligence is defined in a different way in Western and non-Western cultures. In non-Western cultures, social aspects of intelligence are more prominent than in Western culture (Van de Vijver & Tanzer, 2004). This indicates that the construct of intelligence is defined in different ways for the two cultures and using the same measure may result in a score that will not be comparable across the comparison groups.
Method Bias. As the name indicates, it refers to the bias that comes from the different aspects of measurement methodology. The method bias is a result of three different sources, i.e., sample, administrator and instrument. Sample bias as described by Van de Vijver and Tanzer (2004) is caused by “incomparability of samples on aspects other than the target variable”. Examples of ‘other’ variables could include, demographics, testing format etc. For example, if we want to compare performance of two different groups on cognitive ability, and the groups differ in their educational background, then the differences in the construct (cognitive ability) being measured may not be accurate because it will be confounded by the comparison groups’ educational backgrounds (Hambleton et al., 2005; Van de Vijver & Tanzer, 2004).

The second type of method bias is administrator bias. This occurs due to administration problems like communication issues between test administrator and test taker, for example, lack fluency in language of testing and/or unfamiliarity of cultural background of test taker. Additionally, a subject’s lack of understanding of the experimenter’s instruction could also result in administration bias (Van de Vijver & Tanzer, 2004).

The third type of method bias is instrument bias, which occurs due to properties of the instrument/measure such as unfamiliarity of response format. This for example, can cause differing results in one cultural group compared to another (Van de Vijver & Tanzer, 2004).

Item Bias and Differential Item Functioning (DIF). Item bias occurs due to ‘distortions’ in item level characteristics, e.g., poor item translation, unfamiliarity/lack of cultural appropriateness of items, and culture specific factors associated with item wording (Van de Vijver & Tanzer, 2004). While DIF is identified when an item functions differently for two groups after having been matched on their latent ability or attribute (Holland & Wainer, 1993).
In cross cultural comparison studies, DIF typically occurs for reasons such as linguistic or cultural differences (Ercikan & Lyons-Thomas, 2011).

It is important to distinguish item bias from DIF because they have long been used interchangeably in the literature. DIF is a statistical property of an item, while item bias lies in the interpretation of the DIF that is observed (Sireci & Rios, 2013; Wiberg, 2007). Item bias is present when an item is statistically flagged to have DIF. The presence of DIF does not always indicate item bias (AERA et al., 2014; Sireci & Rios, 2013).

According to AERA et al. (2014), there must be an explanation for DIF to justify that the item is biased, which is typically done by different item bias analysis (e.g., content analysis, expert reviews etc.).

2.3 Expert Review

The previous sections have described commonly used methods (CFA & DIF analysis) for examining measurement equivalence. The expert review is also one of the methods used to evaluate items for measurement equivalence. It involves reviewing of items by individuals who have knowledge in the construct being compared and the comparison groups (e.g., gender, culture, language, ethnic, racial group) (Ercikan, Arim, Law, Domene, Gagnon, & Lacroix, 2010).

There have been many studies (Ercikan, 2002; Ercikan et al., 2010; Gierl & Khaliq, 2001; Gierl, Rogers, & Klinger, 1999) that have used expert review methods alongside statistical techniques (e.g. DIF analysis) for examining items which do not have measurement equivalence. One such study was conducted by Elosua and López-Jaúregui (2007), in which experts examined the items that were translated from Spanish to Basque language. The results showed a moderate degree of agreement between the items that had DIF and those identified by experts as having
linguistic differences in the two languages. The overall consistency in classifying items with DIF versus no-DIF between the two methods was almost 70%. The combined approach helps to explain the sources of DIF by linking different problematic elements of items identified by experts with DIF identification (Benítez, Padilla, Montesinos, & Sireci, 2016). For example, if a statistical analysis identified an item as DIF, the expert analysis on that item may provide insights about sources of differential functioning. This mixed method approach, combining statistical and qualitative expert reviews, provides a deeper understanding of DIF items (Benítez et al., 2016). In the current study, expert review method was also used. The details of the procedure are discussed in the method section.

The expert review also has a few limitations. It can identify the characteristics that are responsible for the DIF but it cannot identify the ‘how’ and ‘why’ these characteristics are responsible for DIF (Ercikan et al., 2010). The question of ‘how’ and ‘why’ can be answered by using a think aloud procedures in which participants verbalize their thoughts while answering the questions (Ercikan et al., 2010; Roth, Oliveri, Sandilands, Lyons-Thomas, & Ercikan, 2013). Think aloud procedure is an important method to gain deeper understanding of the sources of DIF, however, this study will only be limited to expert review due to time limits. Future research may want to explore using this procedure.

2.4 Measurement Equivalence in Depression Measures

As previously discussed, establishing equivalence is a prerequisite for cross cultural comparisons and helps to provide meaningful comparisons across cultural groups. There have been many studies which compared two or more cultural groups on depression; however, these studies rarely addressed issues concerning equivalence of measures across cultural groups. The few studies that have, will be reviewed here.
Canel-Çınarbaş, Cui, and Lauridsen (2011) examined the cross-cultural equivalence of BDI-II across Turkish and US undergraduate students. The Turkish students completed the translated and modified version of BDI-II while US students completed the original English version of BDI-II. The authors examined the equivalence using Multi-Group Confirmatory Factor Analysis (MG-CFA). They used the following fit indices in their model: Satorra-Bentler chi-square ($SB\chi^2$), Comparative Fit Indices (CFI), and Root Mean Square Error of Approximation (RMSEA). The indices showed the fit of the model for configural and metric equivalence. However, the item intercepts across the groups were not equal, indicating that the measurement equivalence did not reach the scalar equivalence level. The inequivalence led Canel-Çınarbaş and colleagues (2011) to further examine DIF by Poly-SIB test, which showed that out of 21 items, 12 (57%) items showed DIF while the other 9 did not show any DIF. However, when the items with DIF were examined by expert reviewers they only identified 1 of the items with equivalence and major translation problems and the remaining 11 items with minor translation problems. The expert reviews indicated different levels of equivalence than DIF analyses. The authors still concluded that DIF arises partly due to inequivalence of items of BDI-II and suggested that BDI-II is not a valid measurement tool for the Turkish population (Canel-Çınarbaş et al., 2011).

A study by Hooper, Qu, Crusto, and Huffman (2012) examined DIF of BDI-II in Black and White American students at university in Southeastern United States. They employed both CFA and IRT methodology. Using CFA, they found that 4 out of 21 items showed DIF, whereas when using IRT, only 3 out of 21 items showed DIF. Moreover, there were only two common items showed DIF using both CFA and IRT analysis. These items were measuring loss of interest in sex and self-criticalness. The analysis showed that there are differences in depressive
symptoms across Black and White American (Hooper et al., 2012). Another study by Hambrick et al. (2010) also compared the item responses of Black with White American students at a university in the Northeastern United States examining DIF using IRT-Likelihood Ratio method. Their results showed DIF for two of the items on BDI-II. The authors of this study concluded that cross-ethnic applicability of the BDI-II can be improved by removing two items that showed DIF (Hambrick et al., 2010). The two items that showed the significant DIF, punishment feelings and indecisiveness, were different than the findings of the Hooper et al. (2012)’s study. However, it is to be noted that the two studies were conducted in different regions of the US, which might be the cause of differences in results.

A study by Byrne, Stewart, Kennard, and Lee (2007) examined the measurement equivalence for the BDI-II across American and Hong Kong adolescents. The authors compared the responses of the Chinese version of the Beck Depression Inventory (C-BDI-II) with the original BDI-II. Using CFA, a strong evidence of measurement equivalence was found. However, the total mean scores on all factors were greater for Hong Kong adolescents. Among those factors, the somatic factor had one of the highest mean scores, this finding is consistent with the literature that shows that the somatization of depression is common in non-Western cultures (Islam et al., 2014; Kleinman, 1982). The authors indicated that the higher somatic scores among the Hong Kong adolescents reflect cultural differences in the expression of depression despite strong evidence of measurement equivalence (Byrne et al., 2007).

Cross-cultural equivalence has also been examined for other depression measures that focus on specific age groups, like elderly or children. Jang, Small, and Haley (2001) examined the factor structure of the Geriatric Depression Scale-Short Form (GDS-SF) across Korean and American elderly. The Korean elderly completed the translated non-validated version of GDS-SF
while American elderly completed the original version of BDI-II. The factor structure was examined with exploratory factor analysis using principal component analysis with varimax rotation. The comparison of factor structure showed differences across the two samples. The examination of a few items with low inter-item correlation also showed culturally irrelevant concepts that do not capture the depression construct in Korean culture. For example, they found that one of the items, “Do you prefer to stay at home rather than going out?”, seems irrelevant to the construct of depression in the Korean elderly. This is because in the West, staying home is often associated with isolating oneself, and elderly tend to live independently. Whereas, in Korean culture where it is typical for the elderly to remain living with their families, staying home may not have the same implications of isolation. These differences suggest that caution should be taken when using the measure for Korean elderly (Jang et al., 2001).

A study by Zhang et al. (2011) examined equivalence and factor structure across Dutch and Chinese elderly using the MG-CFA model. Both groups completed the translated version in their native language. The authors found support for a four-factor structure of CES-D, but only partial scalar and metric equivalence were supported. The results suggested that total scores can provide a meaningful comparison, but for few items caution should be taken when comparing depressive symptoms between Dutch and Chinese elderly populations. For example, with regards to the item on CES-D, “I thought my life had been a failure”, the Chinese elderly endorse the item more compared to the Dutch elderly (Zhang et al., 2011).

A study by Kim, Chiriboga, and Jang (2009) examined DIF across 3 racial groups in the United States. The study examined the measurement invariance across Black, Mexican, and White American elderly on CES-D with White American as the reference group. The CFA and IRT based DIF analysis was conducted to examine invariance. Both methods showed that Black
Americans tended to endorse two interpersonal items more highly compared to White Americans. For White and Mexican Americans, CFA identified 16 (80%) while IRT identified 19 (90%) items with DIF. The IRT identified all 16 items identified by CFA. The items with DIF showed that Mexican Americans were more likely to endorse negative items (e.g., “I had crying spells”). For the comparison of Mexican American and Black American, CFA identified 9, while IRT identified 17 DIF items. The 9 items identified by CFA were the same as those identified by IRT. The DIF showed that Mexican Americans are more likely to endorse positive items (e.g., “I enjoyed life”) (Kim et al., 2009). The findings from the study stress the need to interpret the results with caution when using with individuals from diverse cultural backgrounds.

A study by Wu et al. (2012) examined the measurement invariance of the Children’s Depression Inventory (CDI) by comparing it across Chinese and Italian primary school children. Both groups of children completed translated versions of CDI in their native language. Using MASC analysis, authors found a partial metric and scalar equivalence. The authors attributed the partial equivalence to language and cultural factors, since they had not evaluated if the two translated versions (Italian and Chinese) of CDI had the same meaning to the original version of CDI (Wu et al., 2012). It was suggested that Chinese and Italian children had interpreted the items differently than the original CDI items because of cultural differences. For example, one of the items on CDI was “I never do what I am told”, and it is believed that Chinese children interpret this item as referring to independence, or self-motivation, independent rather than the intended Western conceptualization of a deviant behavior. Their findings suggest that caution should be taken when comparing cross cultural differences (Wu et al., 2012).

Kalibatseva, Leong, and Ham (2014) also examined DIF of Composite International Diagnostic Interview (CIDI) depression module in Asian and European Americans. The authors
examined measurement invariance with Item Response Theory Differential Item Functioning (IRT DIF) analysis. They found that Asian Americans endorse items more often when they are related to somatic symptoms of depression, while European Americans were more likely to endorse items that were related to affective symptoms of depression (Kalibatseva et al., 2014). The DIF results did not demonstrated the equivalence of the few items for both European and Asian American. This shows cross-cultural differences in representations of depressive symptoms. Examining those differences will help to provide information about the development of culturally informed measurement tools for diverse cultures.

Examining the literature surrounding equivalence testing for depression measures, I unfortunately did not find a single study that examined the equivalence of any depression measure in the South Asian (Pakistani, Indian, Bangladeshi, Sri Lankan and Nepalese) population. However, literature showed many studies that used BDI on South Asian populations, including Pakistani and few studies (Alansari, 2005; Basker, Moses, Russell, & Russell, 2007; Butt, 2014; Datta, 2004; Husain et al., 2014; Khan, Marwat, Noor, & Fatima, 2015) reported reliability coefficient for their study sample. The reliability is not sufficient to provide evidence for cross-cultural equivalence and applicability of the BDI-II.

The literature search also showed that some studies reported a raw total score of measure or mean score of items and wrongfully claiming it an evidence of cross-group or cross-cultural equivalence. Borsboom (2006) cautioned to make such claims without investigating the measurement invariance of the measure.

No study was found in literature that examined cultural equivalence of any depression measure for equivalence. Given the huge gap in research on South Asian population, including the Pakistani population, there is a need to understand measurement issues in depression tools.
Chapter 3: Method

The purpose of the study was to examine the equivalence of using the BDI-II for a Pakistani population with non-Pakistani as a reference group. There are several methods used to examine the measurement equivalence including multi-dimensional scaling, exploratory factor analysis (EFA), confirmatory factor analysis (CFA) and differential item functioning (DIF) (Milfont & Fischer, 2015). CFA and DIF detection are the most common methods in cross-cultural comparison studies and therefore, was used to examine the equivalence of the BDI-II.

The method section describes the measures, including demographic questionnaire and BDI-II. In the second section, different procedures used to examine the equivalence of BDI-II are discussed, which includes CFA, DIF detection methods (Ordinal Logistic Regression, Mantel-Haenszel and IRT base Lord method), and expert review. The CFA provided the information about measurement equivalence, DIF methods identify items that function differently and expert review explained some of the reasons for inequivalence. The choice of these analyses was based on previous studies on equivalence of measures discussed in the literature review section of this paper.

3.1 Measures

Demographic Questionnaire. The participants were asked to complete a demographic questionnaire, which included questions about age, gender, ethnicity, year of study, country of birth, generation, and fluency of language (English and other languages). The sample demographic questions are shown in appendix A.

BDI-II. The BDI-II is the most commonly used screening instrument for assessing severity of depression (Beck, Steer, & Brown, 1996). The BDI-II is comprised of 21 self-administered items that are scored on a four-point Likert type response scale ranging from 0 (not
The responses are summed to create a total score that ranges from 0 to 63, with higher scores indicating greater severity of depression and lower scores indicating less severity of depression (Beck et al., 1996). The item content in the BDI-II included: (1) sadness, (2) pessimism, (3) past failure, (4) loss of pleasure, (5) guilty feelings, (6) punishment feelings, (7) self-dislike, (8) self-criticalness, (9) suicidal thoughts or wishes, (10) crying, (11) agitation, (12) loss of interest, (13) indecisiveness, (14) worthlessness, (15) loss of energy, (16) changes in sleeping pattern, (17) irritability, (18) changes in appetite, (19) concentration difficulty, (20) tiredness or fatigue, and (21) loss of interest in sex (Beck et al., 1996).

The items on the BDI-II are consistent with the Diagnostic Statistical Manual of Mental Disorders-IV Text Revised (DSM-IV-TR)’s depression criteria, which provides guidelines for diagnosis of psychological disorders to clinicians (Hooper et al., 2012).

Beck and colleagues (1996) also provided a criterion for the probable diagnosis of clinical depression, scores greater than 16 indicate the probability of diagnosing clinical depression. They suggested the criteria for interpretation of the scores as follows; 0 to 13 = minimal severity, 14 to 19 = mild severity, 20 to 28 moderate severity, and scores of 29 or greater indicate severe symptoms of depression (Hooper et al., 2012).

The BDI-II has shown to have excellent psychometric properties, as Beck and colleagues (1996) reported an internal consistency of 0.93 with college students and Farmer (2001) found an internal consistency of 0.92 with clinical outpatients. Additionally, Wang and Gorenstein (2013) examined the reliability of the BDI-II with diverse ethnic and geographic groups by conducting a comprehensive review of 118 studies. The authors found that the BDI-II reliability ranged from 0.73 to 0.96. The BDI-II has also shown evidence of construct validity, including high convergent validity with other depression measures, ranging from 0.66 to 0.86, which indicates
that the construct measured by BDI-II is theoretically related to other depression measures (Wang & Gorenstein, 2013).

The BDI-II shows adequate discriminant validity with Marlowe-Crowne Social Desirability Scale Form C (MCSD), with a correlation of .047 (Osman et al., 1997), indicating that the construct measured by the BDI-II is not related to another construct. However, the literature lacks research that examines the discriminant validity of BDI-II, and it has been suggested that more studies are needed to examine discriminant validity of BDI-II (Wang & Gorenstein, 2013).

In terms of factor structure, the BDI-II is considered multidimensional. Beck et al. (1996) reported two factors, somatic and cognitive dimension in both a clinical and non-clinical sample. A meta-analytic review of BDI-II conducted by Wang and Gorenstein (2013) also showed that most of the studies conducted across diverse ethnic and geographic groups had the same two-factor structure (two dimensions/component). However, a small number of studies also showed more than two factors (Wang & Gorenstein, 2013).

3.2 Participants and Procedure

Participants were recruited from universities in the Greater Toronto Area. There was a total of 400 participants, 200 Pakistani participants in the reference group and 200 participants in the focal group. A minimum of 400 participants was required for this study as it is considered a minimum criterion for DIF analysis (Reeve & Fayers, 2005, p. 71).

There was a total of 200 Pakistani individuals in the focal group. A total of 200 non-Pakistani individuals, 180 Caucasian and 20 other non-Caucasian, non-Pakistani ethnic group individuals were in the reference group. For comparing the focal group to the reference group, the demographics of the reference group were kept similar to the norming population of Beck’s
(1996) original study on non-clinical sample. Beck (1996)’s study used 120 students with 91% identified as Caucasians while rest identified as others.

To recruit participants for the study, advertisements were posted at university campuses in the Greater Toronto Area. The study required the participants from a specific ethnic group (i.e. Pakistani population) therefore; the study was also advertised through student run Pakistani clubs at universities in the Greater Toronto Area. The individuals interested in participating, contacted the researcher through email or phone number provided on the advertisement. The survey was conducted at a location and time convenient to the participant. Participants were also recruited by randomly approaching individuals at student centers at universities in the Greater Toronto Area.

The survey, which included demographic questions and the BDI-II, was the same for both the focal and reference group. To participate, individuals had to be over the age of 18. The participants were presented with an informed consent form explaining the purpose of the study. The sample informed consent form is shown in appendix D. After reviewing the consent form, the participants were asked to complete the self-administered survey, which took 10 minutes. At the end of the survey, participants were given an option to leave their email address if they were interested in the copy of the results and/or wanted to be included in a draw for a gift card, as part of the incentive for participation. Participants were asked to tear the paper and place it in the envelope provided by the researcher. Participants were offered resources for mental health support in their community. The sample resources document is shown in appendix C.

3.3 Internal Consistency

In the current study, a Cronbach alpha, the most common coefficient for Likert-type questions on the BDI-II, was used to examine internal consistency and comparability of measurement accuracy across the two groups. There are no absolute criteria for the Cronbach
alpha value however; when the measures are used to make inferences about an individual then
the value should be at least 0.85 as recommended by Anastasi (1988).

3.4 Confirmatory Factor Analysis

The MG-CFA (Jöreskog, 1971) was conducted using lavaan package to examine fit of
three models associated with measurement equivalence. The MG-CFA involves applying CFA
simultaneously to each model. The first model served as a baseline model, which tested
configural equivalence. This model demonstrated whether there was a similar factor structure
across the groups under comparison, the model was statistically tested by ensuring the equality
of factors. In the next model, the metric equivalence was examined to demonstrate whether the
measure had the same factor loading across groups. This was tested by constraining the factor
loadings to be equal with the varying intercepts. In the last model, the scalar equivalence was
examined to demonstrate whether items were interpreted in the same way by two groups. This
was tested by constraining the item intercepts to be equal. The decision to test next models
depends on the equivalence of previous models. If previous models have not demonstrated
equivalence then next model is not tested (Brown, 2006; Vandenberg & Lance, 2000).

The MG-CFA models were fitted to polychoric correlations using the robust weighted
least square estimation method (WLSMV). The WLSMV do not require the normality
assumption and works best with ordinal data (Brown, 2006). In the current study, the data
collected was from a non-clinical population therefore, non-normality in the data was expected.

There are different fit indices to test the fit of models. In the current study, in addition to
traditional $\Delta \chi^2$ (with critical value less than 0.01), few fit indices were also used. Hu and Bentler
(1998) provided the criteria for those indices as follows; Comparative Fit Indices (CFI) $\geq 0.95$,
Standardized Root Mean Square Residual (SRMR) \( \leq 0.08 \), and Root Means Square Error of Approximation (RMSEA) \( \leq 0.06 \) (Hu & Bentler, 1998).

Cheung and Rensvold (2002) showed that few fit indices are affected by different elements (such as, sample size, normality etc.), therefore it is recommended to used additional criteria for testing fit of models such as, \( \Delta \text{RMSEA} \leq 0.015 \) and \( \Delta \text{CFI} \leq 0.01 \) (Cheung & Rensvold, 2002).

### 3.5 DIF Analysis

DIF analyses were used to examine the item level equivalence across groups. An item is said to have DIF when it functions differently across two groups with same latent ability or attribute (Holland & Wainer, 1993). There are two types of DIF, uniform and non-uniform DIF. Uniform DIF occurs when one group is favored by other group at all the ability levels and can be visually observed by an Item Characteristic Curve (ICC) that it is parallel. The non-uniform DIF is observed when the probability of response is different at different levels and can be observed when the ICC is not parallel (Swaminathan & Rogers, 1990).

There are two different methods for examining DIF, parametric and non-parametric methods. The parametric methods require a specified statistical model and include methods such as: logistic regression, ordinal logistic regression, \( b \) parameter indices, likelihood ratio tests, general IRT LR, IRT LRT, Lord chi-squared test, log linear models, and mixed effect models (Hambleton 2006; Wiberg, 2007).

Non-parametric methods are not based on statistical models and do not require fit between data and model, some non-parametric methods include Mantel-Haenszel, Poly-SIBTEST, standardization, and chi-square methods (Wiberg, 2007).
There are many factors that affect the detection of DIF, including the type of test items (dichotomous, polytomous), test length, sample size, the type of DIF (uniform, non-uniform), statistical software, and the amount of DIF (Ercikan, Gierl, McCreith, Puhan & Koh, 2004; Matsumoto & Van de Vijver, 2011; Wiberg, 2007). Therefore, it has been suggested to use several DIF methods to deal with inconsistencies in results of DIF (Ercikan et al., 2004).

In the current study, three DIF detection methods were used: ordinal logistic regression, Mantel-Haenszel and IRT based Lord method. The ordinal logistic regression method can detect DIF in small samples, which is suitable for the current study because of its relatively small sample size, i.e. 400 participants (Wiberg, 2007). Additionally, the IRT Lord based methods and ordinal logistic regression methods can detect both uniform and non-uniform DIF, while the Mantel-Haenszel method can only detect uniform DIF. However, it is to be noted that IRT Lord based method also requires a large sample and strict assumptions (unidimensionality and local independence). When the assumptions cannot be met the Mantel-Haenszel method will serve as an alternative for the DIF analysis (Wiberg, 2007). The multiple methods can compensate the shortcomings of each method and decrease the false positive chance of detecting items with the DIF (Ercikan et al., 2004).

In the current study, three different DIF methods were used, it was expected to have some inconsistencies in the result as stated by Ercikan et al. (2004). Therefore, an item was considered to have DIF if it was flagged by at least two DIF methods.

**DIF Detection Methods.** Three DIF detection methods were used, the ordinal logistic regression, Mantel-Haenszel and IRT based Lord method.
**Ordinal Logistic Regression.** A ordinal logistic regression for the DIF analysis was initially introduced for dichotomous data by Swaminathan and Rogers (1990). It is a model based approach and allows the ability variable to interact with the grouping variable.

The logistic regression equation can be expressed as:

$$ p(u = 1/x) = \frac{e^{f(x)}}{1 + e^{f(x)}} $$

where,

$$ f(x) = \tau_0 + \tau_1\theta + \tau_2G + \tau_3\theta G $$

$p(u = 1/x)$ is the probability of person obtaining a correct answer given $x$, G is the group membership, $\theta$ is observed test takers ability, $\tau_0$ is the intercept, $\tau_1$ is the ability regression coefficient, $\tau_2$ is the grouping variable, and $\tau_3$ is interaction term between group and ability estimate (Swaminathan & Rogers, 1990).

This model was extended for polytomous data by few researchers (e.g., French & Miller, 1996; Zumbo, 1999). In the current study, Zumbo’s proposed model will be used.

Zumbo (1999) combined the Ordinal Logistic Regression (OLR) with $R^2$ measure of effect size to detect DIF in polytomous items. The OLR were carried out in three steps in hierarchical fashion. In the first step, only total score (conditioning variable) was entered an equation, in the second step a grouping variable was entered in equation and then in the last and third step interaction variable (grouping variable X total score) wase. The chi-square value was calculated for each step to test for significance. To classify items as having DIF, chi-square differences statistics were compared at different steps. To assess uniform DIF, differences of chi-square statistics between step 2 and 1 were examined, for non-uniform DIF the differences of chi-square statistics between step 2 and 3 were considered, and the chi-square test difference
between step 1 and step 3 give simultaneous statistical tests for uniform and non-uniform DIF (Zumbo, 1999). In each step, the coefficient of determination was calculated, which provided an estimation of variance accounted for by the variable added at each step, for example, the difference ($R^2\Delta$) of R square value from step 3 and step 1 gives estimates of effect size of non-uniform DIF. They can be represented through equation as

$$R^2\Delta = R^2_3 - R^2_1,$$

Zumbo and Thomas (1997) proposed the following criteria for effect size classification:

- Negligible Effect Size, $R^2\Delta < 0.13$
- Moderate Effect Size, $R^2\Delta = 0.13$ to 0.26
- Large Effect Size, $R^2\Delta = 0.13$ to 0.26

However, Jodoin and Gierl (2001) cautions researchers on using Zumbo and Thomas’s (1997) criteria because it identifies fewer items with DIF, also shown by Jodoin and Gierl’s (2001) simulation study. They demonstrated that, when using Zumbo and Thomas’s (1997) criteria, only 6.8 % of items were identified as having type B DIF, while 68.2 % of items were identified by Jodoin and Gierl’s (2001) as having type B DIF. Therefore, in the current study, Jodoin and Gierl’s (2001) criteria was used for identifying DIF, which is as follows:

- Negligible Effect Size, $R^2\Delta < 0.035$
- Moderate Effect Size, $R^2\Delta = 0.035$ to 0.7
- Large Effect Size, $R^2\Delta = 0.7$

The OLR method of DIF analysis has many advantages over other methods, as it allows estimation of effect size, has higher power and can be applied to small samples. It also gives information about both uniform and non-uniform DIF (Zumbo, 1990). However, if the sample
size of the reference and focal groups is unequal, it can falsely flag the items with DIF (Narayanan & Swaminathan, 1996).

**Mantel-Haenszel.** The Mantel-Haenszel procedure is the most commonly used non-parametric method. This method was originally developed to use for dichotomous data but has been adapted for used with polytomous data (Holland & Thayer, 1988; Matsumoto & Van de Vijver, 2011; Wiberg, 2007).

The MH method utilizes 2 X 2 contingency tables based on the number of correct responses for reference and focal groups (Dorans & Holland, 1992). The odd ratio of success for each item is compared for two groups under examination. The MH statistic is obtained by multiplying the odd ratio with -2.35 (Zwick, Thayer, & Lewis, 1999).

This can be represented on ‘delta scale’ or ‘delta metric’ as

\[ \text{MH D-DIF} = -2.35 \ln [\alpha_{\text{MH}}] \]

The DIF could be positive or negative and it can be classified into one of five categories; C-, B-, A, B+, and C+ (Zwick et al., 1999). A positive value indicates an item in a favour of a focal group while a negative value indicates an item in a favour of a reference group (Dorans & Holland, 1992). The details of this method can be found in following studies: Dorans and Holland (1992) and Holland and Thayer (1988).

In the MH method, DIF can be identified by both effect size and significance testing independently. However, it has been recommended to use a blended rule, using both effect size and significance testing for identifying an item with DIF. A study by Gómez-Benito, Hidalgo, and Zumbo (2013) showed that using both effect size and significance testing decreases the false positive rate of identifying DIF.
This blended rule is also used by Educational Testing Services (ETS) programs, as described by Zwick and Ercikan (1989) in the following way:

Type A items: Negligible DIF, items with MH D-DIF < 1 or statistically non-significant

Type B items: Moderate DIF, items with MH D-DIF > 1 & < 1.5 and statistically significant

Type C items: Large DIF, items with MH D-DIF > 1.5 and statistically significant

This method is considered powerful for detecting uniform DIF with small sample size. However, it is unable to detect non-uniform DIF, which is one of the limitations of this method (Matsumoto & Van de Vijver, 2011).

**IRT-Based Lord’s Method.** This method is based on the comparison of item parameters of the groups. According to Lord (1980), if the item parameters are invariant for groups or differ significantly for the groups, then the test is said to have items with DIF. The item parameters for two groups must be scaled on common metric or unit before comparing for statistical significance (Lord, 1980).

Lord (1980) proposed a chi-squared test to examine the difference between item parameters of two groups, with a null hypothesis for a given item, $i$, both $a_iR = a_iF$ and $b_iR = b_iF$ ($b_iR$ and $b_iF$ are difficulty parameters and $a_iR$ and $a_iF$ are discrimination parameters). The chi-squared test $\chi_i^2$ is as follows:

$$\chi_i^2 = v_i' \Sigma_i^{-1} v_i$$

$v_i$ is the vector difference in estimates of parameters between two groups. $\Sigma$ is the covariance of the matrix for the differences between parameters and the test statistics, $\chi_i^2$ has distribution with 2 degrees of freedom (Lord, 1980).
The Lord method was then improved by Langer (2008) where he suggested using an MML estimation, which supplemented EM algorithm allowing for a more accurate estimation of standard errors and item parameters (Woods, Cai, & Wang, 2013).

In the current study, IRTPRO was used to detect DIF. The discrimination and difficulty parameters of both groups were compared and tested for DIF using Wald $\chi^2$ statistics. The significant statistics of item were an indicator of DIF.

**IRT Assumptions.** There are some assumptions that need to be satisfied before using IRT based DIF methods. The following are the two most common assumptions: Unidimensionality and Local Independence (Hambleton, Swaminathan, & Rogers, 1991).

Unidimensionality. Only one factor or single dimension can be measured by the items in the test and is rarely met in the case of psychological measures. Therefore, the concept of ‘essential unidimensionality’ has been established. According to essential unidimensionality’ if a test data indicates presence of a dominant factor that influences test performance then the assumption can be adequately met (Hambleton et al., 1991).

The different criteria had been proposed to identify dominant factor in essential unidimensionality. The principal component analysis results in estimation of percentage of variance accounted for by the variables and eigenvalues that provide information related to essential unidimensionality. Reckase (1979) suggested that if percentage of variance for the first factor is at least 20 % then the measure can be considered as having essential unidimensionality. The ratio of first-to-second eigenvalues is another criterion, in which if the ratio of first-to-second eigenvalues is greater than 3 or 4 then this indicates essential unidimensionality (Gorsuch, 1983).
In the current research, exploratory factor analysis (EFA) was conducted using SPSS v.23. The EFA computed percentage of variance and eigenvalues, which was observed for unidimensionality. The percentage of variance of at least 20% and the ratio of first-to-second eigenvalues greater than 3 indicates the presence of essential unidimensionality.

The scree plots also determine the number of factors with each bend indicating a factor. It has been shown from studies that the interpretation of the scree plot is very subjective (Beavers et al., 2013). Therefore, in the current study a scree plot was not examined.

**Local Independence.** This assumption states that there should be no relationship between different responses of the individual in a test that cannot be accounted for by the targeted construct (Hambleton et al., 1991). If the individual responses are not independent of their responses on the other items then the violation of local independence occurs, referred to as local dependence (Embretson & Reise, 2000).

To test the local independence assumption, standardized chi-square statistics were computed for each pair of items using IRTPRO software. A value of an item pair less than 10 indicates presence of local independence (Cai, Du Toit, & Thissen 2011).

**Advantages and limitations.** The advantage of IRT Lord method is that it provides both uniform and non-uniform DIF as well it also tends to have relatively lower type 1 error. This method also has limitations, like other IRT methods, the method requires a moderate sample, with at least 200 participants per parameter, per group which limits its use on smaller sample sizes (Sireci & Rios, 2013).

**3.6 Expert Review**

To examine the sources of DIF, the BDI-II items were reviewed by expert reviewers all clinical psychologists with experience working with Pakistani populations in Canada and/or
Pakistan. The expert reviewers were selected by screening the local clinics websites for Pakistani Clinical Psychologists. They were contacted to inquire about their interest in participating in the study. Three of them agreed to participate in the study.

It is to be noted that there are no clear guidelines for choosing the number of expert reviewers. The previous studies (Benítez et al., 2016; Elsua & López-Jaúregui, 2007; Ercikan, 2002; Ercikan et al., 2004; Gierl et al., 1999; Gierl & Khaliq, 2001) have used 3 to 12 expert reviewers. In the current study, choice of three reviewers was subjective and depended on availability of Pakistani Clinical Psychologists. Those who agreed to participate in the study were sent a professional profile form and the feedback form. In the professional profile form, reviewers were asked about experience in counselling Pakistani clients. In the feedback form, the reviewers were asked to rate each item on the extent to which items are comparable and to identify the elements in the items that affect the comparability of items. The items ranked ranged from 0 (not comparable) to 3 (very comparable). The sample questions are shown in appendix B. This method of expert review rating was adapted from Benítez et al. (2016) and Ercikan (2002).

The reviewers were not informed about the items that showed or tended to show DIF since it could bias their reviews. Each reviewer rated the item independently at their convenience.
Chapter 4: Results

The purpose of the study was to examine the cultural equivalence of the BDI-II. A total of 400 participants from the Greater Toronto Area were included in this study. Participants provided verbal consent and completed a self-administered questionnaire consisting of demographic questions and the BDI-II.

In the first set of analyses, internal consistency was calculated to determine if the constructs measured were in the same level of accuracy for two groups. The second set of analyses, factor structures were examined to see if they were the same for each group. The third set of analyses included the DIF analysis using Ordinal Logistic Regression, Mantel-Haenzsel and IRT based Lord method, to determine if the items functioned differently for two groups. The fourth part of research involved qualitative reviews by experts for cultural relevance and equivalence of the items.

4.1 The Sample

The data were collected from universities in the Greater Toronto Area (GTA). Two hundred participants from each group completed the survey. The mean age of the participants was 20.51 (SD 2.6) with an age range of 18-30 years of age. All participants provided verbal consent before participating in the survey. There was a total of 142 (35.5 %) male and 258 (64.5 %) female participants. The focal group was comprised of 56 (28 %) male and 144 (72 %) female participants while reference group included 86 (43 %) male and 114 (57 %) female participants as shown in Table 4.1. The mean age of the participants in the reference group was 20.39 (SD 2.74) and in the focal group, it was 20.61 (SD 2.46) years of age. The participants were also asked to rate their English fluency from 1 (not fluent) to 5 (very fluent). The 162 (81 %) focal group and 185 (92.5%) reference group participants rated their English as ‘very fluent’.
A total of 157 (78.5 %) individuals in the focal group and 25 (13 %) individuals in the reference group identified themselves as first-generation Canadians (individuals who are not born in Canada).

Table 4.1

<table>
<thead>
<tr>
<th>Characteristics of the Study Sample</th>
<th>Focal (N=200)</th>
<th>Reference (N=200)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Demographics</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Male</td>
<td>56 (28 %)</td>
<td>86 (43 %)</td>
</tr>
<tr>
<td>Female</td>
<td>144 (72 %)</td>
<td>114 (57 %)</td>
</tr>
<tr>
<td>Age</td>
<td>Mean = 20.39 (SD 2.74)</td>
<td>Mean = 20.61(SD 2.46)</td>
</tr>
<tr>
<td>‘Very fluent’ in English</td>
<td>162 (81 %)</td>
<td>185 (92.5 %)</td>
</tr>
<tr>
<td>First Generation Canadian</td>
<td>157 (78.5 %)</td>
<td>25 (13 %)</td>
</tr>
</tbody>
</table>

The data were collected from a non-clinical population with only a few participants having extreme scores on depression. It was expected that the data would be skewed because fewer participants selected extreme response options (e.g., 2 and 3). Therefore, out of the four response options (0, 1, 2, 3), response categories 2 and 3 were combined into one response category as 2.

4.2 Reliability Analysis

The internal consistency coefficients were calculated separately for each group. The Cronbach alpha for the focal group (Pakistani) was 0.870 and the Cronbach alpha for the reference (non-Pakistani) group was 0.902. The reliability coefficient indicated strong internal consistency for both groups with only a slightly lower internal consistency for the focal group.
The ‘Cronbach alpha if item deleted’ for focal group was highest for Item 21 (Loss of interest in sex). By deleting this item, the Cronbach alpha could be improved and increased to 0.873.

For the reference group, ‘Cronbach alpha if item deleted’ for item 21 was 0.879, which was low compared if another item was deleted for the reference group.

The data also showed that the item 21 had more missing cases for the focal group (5.5 %) compared to the reference group (0.5 %). The missing data and high Cronbach alpha if item deleted’ might indicate that item 21 is irrelevant to construct of depression for the Pakistani immigrant population.

4.3 Factor Analysis

To examine the factor equivalence, a multi-group confirmatory factor analysis (MG-CFA) was conducted using lavaan package. A MG-CFA model was fitted to the data from two groups using WLSMV estimation. The model fit to data included configural, metric, and scalar invariance models.

In the first model, configural invariance was tested to determine if the overall model was equivalent across the groups. The model showed weak evidence for configural invariance with $\chi^2 = 594.649$, CFI= 0.850, RMSEA= 0.056, SRMR= 0.073. The configural invariance shows that same general concept was measured across each group.

Further analysis was conducted to test metric invariance to determine if the factor loadings were equal across the groups. The second model showed weak fit with $\Delta \chi^2 = 11.11, p < .9433$, CFI=0.872, $\Delta$CFI = 0.022, RMSEA= 0.051, $\Delta$RMSEA = -0.005, SRMR= 0.080 as shown in the Table 4.2. This suggests that individuals from each group responded to items in a similar way.
In the third and last model, scalar invariance was tested to determine if the factor loadings and item intercepts were equal across the groups. The model showed poor fit with $\Delta \chi^2 = 38.5$, $p = .0075$, CFI = 0.859, $\Delta$CFI = - 0.013, RMSEA = 0.052, $\Delta$RMSEA = 0.001, SRMR = 0.083. The Table 4.2 showed that two fit statistics (CFI and SRMR) of last model failed to meet the criteria for scalar invariance. Also, the difference of chi-square was significant and the $\Delta$CFI was greater than 0.01, which indicates that the scalar invariance was not met for the model. This means that the reference and focal group who were on the same level of construct have different score on a few items, which mean that a few items were biased for one group of participants. The results of the MG-CFA suggest that the BDI-II was not functionally equivalent for the Pakistani group.

Table 4.2

<table>
<thead>
<tr>
<th>Model</th>
<th>$\chi^2$</th>
<th>$\Delta$df</th>
<th>$\Delta \chi^2$</th>
<th>CFI</th>
<th>$\Delta$CFI</th>
<th>RMSEA</th>
<th>$\Delta$RMSEA</th>
<th>SRMR</th>
<th>Invariance</th>
</tr>
</thead>
<tbody>
<tr>
<td>Configural</td>
<td>594.649</td>
<td>-</td>
<td>-</td>
<td>0.850</td>
<td>-</td>
<td>0.056</td>
<td>-</td>
<td>0.073</td>
<td>weak</td>
</tr>
<tr>
<td>Metric</td>
<td>583.530</td>
<td>20</td>
<td>11.11</td>
<td>0.872</td>
<td>0.022</td>
<td>0.051</td>
<td>- 0.005</td>
<td>0.080</td>
<td>weak</td>
</tr>
<tr>
<td>Scalar</td>
<td>622.076</td>
<td>20</td>
<td>38.5*</td>
<td>0.859</td>
<td>-0.013</td>
<td>0.052</td>
<td>0.001</td>
<td>0.083</td>
<td>No</td>
</tr>
</tbody>
</table>

*Note.* *p* < .01.

*Note.* Bold figures indicate criteria is not met for invariance.

### 4.4 DIF Analyses

The DIF analysis allowed for the comparison of item equivalence for both groups. As suggested by previous research (Ercikan et al., 2004), in this study three methods were used to verify the DIF status of items. The following DIF detection methods were used in the study: Ordinal Logistic Regression (Swaminathan & Rogers, 1990), Mantel-Haenszel (Holland & Thayer, 1988) and Item Response Theory Lord-Based (Lord, 1980).

**DIF Detection by Ordinal Logistic Regression.** The ordinal logistic regression (OLR) was conducted using SPSS v.23 utilizing a methodology suggested by Zumbo (1999). The OLR
method identified three items with DIF with all of them having uniform DIF (shown in Table 4.3). The effect size for all three items were ‘negligible’ ($R^2 \Delta < .035$), as classified by Jodoin and Gierl’s (2001) criteria.

Table 4.3

Result for DIF detection by Ordinal Logistic Regression

<table>
<thead>
<tr>
<th>Item</th>
<th>DIF</th>
<th>Uniform-DIF</th>
<th>Non-Uniform DIF</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Chi square</td>
<td>$p$</td>
<td>Effect Size</td>
</tr>
<tr>
<td>Item 5</td>
<td>11.338</td>
<td><strong>0.0034</strong></td>
<td>0.031</td>
</tr>
<tr>
<td>Item 6</td>
<td>9.067</td>
<td><strong>0.0107</strong></td>
<td>0.031</td>
</tr>
<tr>
<td>Item 11</td>
<td>8.909</td>
<td><strong>0.0116</strong></td>
<td>0.025</td>
</tr>
</tbody>
</table>

*Note.* Bold figures indicate presence of DIF.

![Figure 4.1. Item Characteristic Curve of item 5](image-url)
Figure 4.2. Item Characteristic Curve of item 6

Figure 4.3. Item Characteristic Curve of item 11
**DIF Detection by Mantel-Haenszel.** The Mantel-Haenszel procedure was also used to identify items with DIF. The MH procedure was conducted using jMetrik software. The MH method identified the three items as having DIF shown in Table 4.4. The ICCs were drawn using jMetrik which also showed three items with DIF as shown in Figures 4.1, 4.2 and 4.3. The ICCs are parallel to each other, indicating the presence of uniform-DIF. The items identified by MH method was same as indicated by ordinal logistic regression. However, the effect size was stronger than what was classified in the ordinal logistic regression. The effect size was largest for item 5 as indicated by capital letters CC while for item 6 and item 11 the effect size was moderate as indicated by capital letters BB. The positive sign indicates that these items were endorsed by the focal group (Pakistani participants) whereas, the negative sign indicates that item was favoured by non-Pakistani participants (reference group). The order of magnitude of the chi-square values were similar to that identified using OLR DIF method e.g. Item 5 with highest magnitude.

<table>
<thead>
<tr>
<th>Item</th>
<th>Chi square</th>
<th>P value</th>
<th>Class</th>
<th>Effect Size</th>
</tr>
</thead>
<tbody>
<tr>
<td>Item 5</td>
<td>13.72</td>
<td><strong>0.001</strong></td>
<td>CC+</td>
<td>Large</td>
</tr>
<tr>
<td>Item 6</td>
<td>5.27</td>
<td><strong>0.001</strong></td>
<td>BB+</td>
<td>Moderate</td>
</tr>
<tr>
<td>Item 11</td>
<td>6.78</td>
<td><strong>0.001</strong></td>
<td>BB-</td>
<td>Moderate</td>
</tr>
</tbody>
</table>

*Note. + (positive) sign indicates item in favour of Focal group and – (negative) sign indicates item in favour of Reference group.*

*Note. Bold figures indicate presence of DIF.*
IRT Based DIF Analysis. The IRT method of DIF detection was conducted by IRTPRO software. The following assumptions of IRT were examined before conducting DIF analysis: unidimensionality, local independence and model fit.

The dimensionality of the BDI-II was evaluated using Principle Component Analysis (PCA) by SPSS v.23. PCA provided eigenvalues and percentage of variances for each group as shown in the Table 4.5 and Table 4.6.

Table 4.5

<table>
<thead>
<tr>
<th>Component</th>
<th>Initial Eigenvalues</th>
<th>Extraction Sums of Squared Loadings</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Total</td>
<td>% of Variance</td>
</tr>
<tr>
<td>1</td>
<td>5.748</td>
<td>27.370</td>
</tr>
<tr>
<td>2</td>
<td>1.824</td>
<td>8.686</td>
</tr>
<tr>
<td>3</td>
<td>1.557</td>
<td>7.412</td>
</tr>
<tr>
<td>4</td>
<td>1.422</td>
<td>6.772</td>
</tr>
<tr>
<td>5</td>
<td>1.162</td>
<td>5.536</td>
</tr>
<tr>
<td>6</td>
<td>1.024</td>
<td>4.876</td>
</tr>
<tr>
<td>7</td>
<td>.934</td>
<td>4.450</td>
</tr>
<tr>
<td>8</td>
<td>.898</td>
<td>4.276</td>
</tr>
<tr>
<td>9</td>
<td>.864</td>
<td>4.112</td>
</tr>
<tr>
<td>10</td>
<td>.788</td>
<td>3.752</td>
</tr>
<tr>
<td>11</td>
<td>.657</td>
<td>3.130</td>
</tr>
<tr>
<td>12</td>
<td>.609</td>
<td>2.898</td>
</tr>
<tr>
<td>13</td>
<td>.555</td>
<td>2.643</td>
</tr>
<tr>
<td>14</td>
<td>.523</td>
<td>2.490</td>
</tr>
<tr>
<td>15</td>
<td>.469</td>
<td>2.232</td>
</tr>
<tr>
<td>16</td>
<td>.442</td>
<td>2.104</td>
</tr>
<tr>
<td>17</td>
<td>.371</td>
<td>1.765</td>
</tr>
<tr>
<td>18</td>
<td>.327</td>
<td>1.559</td>
</tr>
<tr>
<td>19</td>
<td>.304</td>
<td>1.448</td>
</tr>
<tr>
<td>20</td>
<td>.289</td>
<td>1.378</td>
</tr>
<tr>
<td>21</td>
<td>.233</td>
<td>1.111</td>
</tr>
</tbody>
</table>
Table 4.6

Total Variance for Reference group

<table>
<thead>
<tr>
<th>Component</th>
<th>Initial Eigenvalues</th>
<th>Extraction Sums of Squared Loadings</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Total</td>
<td>% of Variance</td>
</tr>
<tr>
<td>2</td>
<td>1.569</td>
<td>7.470</td>
</tr>
<tr>
<td>3</td>
<td>1.184</td>
<td>5.638</td>
</tr>
<tr>
<td>4</td>
<td>1.124</td>
<td>5.354</td>
</tr>
<tr>
<td>5</td>
<td>1.101</td>
<td>5.243</td>
</tr>
<tr>
<td>6</td>
<td>1.051</td>
<td>5.006</td>
</tr>
<tr>
<td>7</td>
<td>.956</td>
<td>4.552</td>
</tr>
<tr>
<td>8</td>
<td>.928</td>
<td>4.421</td>
</tr>
<tr>
<td>9</td>
<td>.797</td>
<td>3.794</td>
</tr>
<tr>
<td>10</td>
<td>.723</td>
<td>3.445</td>
</tr>
<tr>
<td>11</td>
<td>.672</td>
<td>3.200</td>
</tr>
<tr>
<td>12</td>
<td>.637</td>
<td>3.033</td>
</tr>
<tr>
<td>13</td>
<td>.598</td>
<td>2.849</td>
</tr>
<tr>
<td>14</td>
<td>.582</td>
<td>2.773</td>
</tr>
<tr>
<td>15</td>
<td>.552</td>
<td>2.629</td>
</tr>
<tr>
<td>16</td>
<td>.470</td>
<td>2.236</td>
</tr>
<tr>
<td>17</td>
<td>.401</td>
<td>1.908</td>
</tr>
<tr>
<td>18</td>
<td>.360</td>
<td>1.715</td>
</tr>
<tr>
<td>19</td>
<td>.306</td>
<td>1.457</td>
</tr>
<tr>
<td>20</td>
<td>.294</td>
<td>1.400</td>
</tr>
<tr>
<td>21</td>
<td>.251</td>
<td>1.196</td>
</tr>
</tbody>
</table>

The Table 4.5 and 4.6 showed that the percentage of variance for the first factor of each group was 27.37 for the focal group and 30.68 for the reference group. The ratio of first-to-second eigenvalues was 3.151 for focal group and 4.106 for reference group as calculated below.
For Focal Group,
\[
\frac{\lambda_1}{\lambda_2} = \frac{5.748}{1.824} = 3.151
\]

For Reference Group,
\[
\frac{\lambda_1}{\lambda_2} = \frac{6.443}{1.569} = 4.106
\]

This analysis demonstrated that the percentage of variances for first factor of focal and reference group was greater than 20% as shown in Table 4.5 and 4.6 and the eigenvalue ratios were greater than 3 i.e. 3.151 and 4.106 for focal and reference group, indicating the presence of one dominant factor, hence essential unidimensionality assumption was met.

The local independence assumption was tested to demonstrate each item was independent of responses to other items. To test the local independence, chi-square (LD \(\chi^2\)) was calculated using IRTPRO software. The LD \(\chi^2\) values that exceed 10 are suggestive of local dependence. The analysis showed that no item pairs exceeded 10 indicating that the local independence assumption was met.

The fit of the items to the graded response model tested was adequate and resulted in following statistics, \(M_2=380.98, p=0.045, \text{RMSEA }=0.02\). The S-X\(^2\) item level fit indices also indicated an adequate fit with only few items that showed statistically significant values.

After testing the IRT model assumptions, the DIF analysis was conducted using IRT PRO software. Out of 21, the same three items identified by OLR and MH were identified as DIF by IRT method. The two items (item 5 and 6) were in favour of the focal group (Pakistani) and one
item (item 11) was in favour of the reference group (non-Pakistani). Item 5 was identified with both uniform and non-uniform DIF, item 6 was identified to have uniform DIF and item 11 had non-uniform DIF as shown in Table 4.7.

Table 4.7

<table>
<thead>
<tr>
<th>Items</th>
<th>DIF</th>
<th>Non-Uniform DIF</th>
<th>Uniform DIF</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>$\chi^2$</td>
<td>$\chi^2$</td>
<td>$\chi^2$</td>
</tr>
<tr>
<td>Item 5</td>
<td>14.1</td>
<td>3.2</td>
<td>10.9</td>
</tr>
<tr>
<td></td>
<td>0.0028</td>
<td>0.0455</td>
<td>0.0043</td>
</tr>
<tr>
<td>Item 6</td>
<td>7.4</td>
<td>0.1</td>
<td>7.4</td>
</tr>
<tr>
<td></td>
<td>0.0410</td>
<td>0.9770</td>
<td>0.0251</td>
</tr>
<tr>
<td>Item 11</td>
<td>8.5</td>
<td>4.7</td>
<td>3.8</td>
</tr>
<tr>
<td></td>
<td>0.0366</td>
<td>0.0301</td>
<td>0.1508</td>
</tr>
</tbody>
</table>

Note. Bold figures indicate presence of DIF.

Consistency among DIF Detection Methods. The results showed that all three methods identified the same three items as DIF, however the degree of effect size varied for each method. For all items, Mantel-Haenszel showed a stronger effect size compared to the ordinal logistic regression. These differences could be attributed to the differences in estimating the DIF for three methods as discussed earlier in literature.

The OLR and MH method identified all items with uniform DIF while IRT methods showed some differences. The IRT method identified item 11 with non-uniform DIF while other methods verified the item with uniform DIF.
Table 4.8

*Items identified as DIF by three methods*

<table>
<thead>
<tr>
<th>Items</th>
<th>DIF Detection Methods</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>LR</td>
</tr>
<tr>
<td>Item 5 (guilty)</td>
<td>Uniform (N)</td>
</tr>
<tr>
<td>Item 6 (punishment)</td>
<td>Uniform (N)</td>
</tr>
<tr>
<td>Item 11 (agitation)</td>
<td>Uniform (N)</td>
</tr>
</tbody>
</table>

*Note.* N indicates negligible DIF, M indicates moderate DIF.

### 4.5 Expert Reviews

The equivalence of the BDI-II items was analyzed by two female and one male expert reviewers. The expert reviewers were Pakistani-Canadian clinical psychologists with 15-17 years of experience in counseling Pakistani descent clients in Pakistan (mostly in province of Punjab) and Canada (mostly in Ontario or British Columbia). The reviewers were fluent in three languages including English, Urdu and Punjabi. All the reviewers received their master degree and post-masters diploma in Clinical Psychology from a same institute in Lahore, Pakistan. All reviewers had doctorate degrees in Clinical Psychology. Two reviewers completed their doctorate degree in US while one completed in Pakistan. The reviewers evaluated the equivalence of BDI-II for Pakistani population. They were not aware of the items that showed DIF in the study or tendency to show DIF. The reviewers were asked to rate the items from 0 to 3 if they were comparable across Pakistani culture with 0 being least and 3 being most comparable. If the reviewers rated any items less than 2 they were asked to identify any ‘term’ or ‘expression’ in the item that can indicate a different meaning.
The ratings of reviewers identified several items with interpretation issues including item 3, 5, 6, 7, and 21 and rated the degree of comparability from 0 to 3. The identified items also included the items (item 5 and item 6) that were identified as DIF in this study earlier. The ratings of these two items were lowest compared to the rest of the items, which indicated that item 5 and item 6 were identified as having comparability issues. The reviewers suggested that these items, especially item 5 and 6, can be interpreted in religious context by Pakistani participants. For example, the Pakistani participants tend to associate punishment and guilty feelings with the perceived wrong doing of religious obligations, which indicates that these feelings are often not associated with depression (Siddiqui & Shah, 1997). The reviewers recommended some rewording of the item, so that participants would be less likely to answer the question in a religious context. This suggests that clinicians and researchers should be cautious when using the BDI-II on Pakistani participants since there is a tendency to have high score on certain items despite the participants having low depression scores.

Furthermore, the reviewers also suggested that additional somatic items be included to better capture the depressive symptomology of Pakistani participants since somatic expression of depression is prominent in non-Western cultures compared to Western cultures (Bhui, Bhugra, Goldberg, Sauer & Tylee, 2004).

The reviewers also commented on item 21 which is about ‘loss of interest in sex’, the reviewers expressed a concern that the questions related to sex will less likely be answered by Pakistani participants as it might make them uncomfortable. A similar finding was presented by Siddiqui and Shah (1997) scale development study. In the item development stage, author asked the participants to list statements about feeling and behavior during depressive situations. The authors were not able to find a single statement that showed effect on sexual activity. The lack
representation of item reflected a conservative culture of the Pakistani culture which restricts expression of sexual desire discussion of related topics are considered a taboo (Siddiqui & Shah, 1997). Additionally, a study by Mumford et al. (2005) on the development of a Pakistan Anxiety and Depression Questionnaire (PADQ) also observed similar results. In item testing phase, the authors found a ‘less favourable psychometric results’ on the item measuring libido, which led to the removal of the item from the questionnaire.

4.6 Summary

The results of the analysis to determine the equivalence of the BDI-II for Pakistani immigrant and non-Pakistani groups were described. The reliability analysis showed a high degree of internal consistency for the two groups, with slightly lower internal consistency for the focal group compared to the reference group. The lower internal consistency for the focal group may be due to an item 21 measuring ‘loss of interest in sex’, which had high degree of missing data for the Pakistani-Canadian group. The findings were also consistent with Siddiqui and Shah’s (1997) study discussed earlier.

The multi-group confirmatory factor analysis demonstrated evidence for configural invariance and metric invariance but there was no evidence for scalar invariance, suggesting that a few items were not equivalent across both groups. The DIF analysis methods identified the items that functioned differently for two groups. All three DIF methods identified the same set of three items (item 5, 6, and 11) that showed DIF but with differences in type and strength of DIF. The results of the DIF were consistent with literature.

In the last stage, experts reviewed the BDI-II items for cultural appropriateness. The reviewers recommended rewording of some items. The experts also advised adding items that focus on somatic symptomology of depression.
Chapter 5: Discussion

This study used a convenience sample of students from universities in the Greater Toronto Area to examine the equivalence of the Beck Depression Inventory-II (BDI-II). Specifically, the CFA and DIF methods were used to analyze the equivalence of the BDI-II in Pakistani-Canadians, with non-Pakistanis utilized as a reference group. In addition, expert feedback was also gathered on the cultural equivalence of the BDI-II.

In the first set of analysis, a MG-CFA was conducted to investigate factorial invariance. The results indicated the presence of weak configural and metric invariance, however there was no evidence of a scalar invariance across two groups. This result suggests the presence of partial measurement invariance (i.e. similar factor structure and equal factor loadings but unequal item intercepts). In other words, configural evidence indicates that the common factor was associated with identical items, which implies that same construct of depression was measured across each group. The metric evidence indicates that the strength of the relationship between factors and items of the BDI-II were similar, determining that both groups responded to the measure in a similar way. The absence of scalar invariance indicates that differences in the BDI-II scores cannot be fully explained by the mean differences in the trait, surmising that some items were interpreted differently by two groups. The results of MG-CFA indicated the presence of partial measurement invariance, hence the comparison of scores could not be meaningful. This result stresses the importance of being cautious when comparing cross-cultural differences across groups. Using the measures without evidence of measurement equivalence will increase the chance of false positive and false negative diagnoses in assessing depression. Using the BDI-II, the inferences from the cross-group comparison can be questionable.
The DIF analysis also supported comparability issues with the BDI-II. The DIF was conducted to investigate measurement invariance of the BDI-II at the item level. Three DIF detection procedures were used to increase the accuracy of identifying DIF. Results showed three items with DIF, where two of the items functioned differently for Pakistani groups: item 5 (guilty feelings) and item 6 (punishment feelings). The result from the DIF analysis denotes that Pakistanis are more likely to endorse punishment and guilt related items.

No previous research has examined the BDI-II at the items level for a Pakistani population or any other South Asian ethnic groups, however, previous studies have described the association of depression with guilt or punishment feelings. In one such study by Haddad and Colleagues (2016), non-psychiatric, Pakistani medical professionals were asked about the causes of depression. Almost one quarter of these study participants described the cause of depression a result of supernatural forces, including punishment from God. The Haddad et al. (2016) study also concluded that this kind of association is due to a more negative attitude towards mental health issues. By putting the responsibility on supernatural forces, the cause of pathological behaviour is not examined; and in turn, provides justification for behaving in a certain way (Haddad et al., 2016).

Furthermore, Siddiqui and Shah (1997) stated that individuals in Western cultures experience guilt and punishment feelings, however, the feelings of guilt experienced by Pakistanis are often associated with perceived wrongdoing of religious obligations when compared to the Western concept of guilt. These strong feelings of guilt are related to the religious orientation of Pakistanis, suggesting that despite lower depression scores, there is a tendency for Pakistani participants to have high scores on items related to guilt or feelings of punishment (Siddiqui & Shah, 1997). Previous studies also reported feelings of guilt as
component of depression for non-Western populations (Bhugra, Gupta, & Wright, 1997; El-Islam, 1969).

It should be noted that in the current study, the participants were not asked about their religious orientation or degree of religiosity. To further understand the association of punishment feelings with religious compulsions, future research can use a think aloud protocol. In a think aloud protocol, participants verbalize their thoughts while answering questions. These verbalizations would help determine if there is any influence of religiosity when responding to questions in the BDI-II regarding punishment and guilt feelings.

The DIF analysis also showed that non-Pakistani participants, mostly comprised of Caucasian-Canadians, endorsed an item related to agitation. This finding is consistent with Carmody’s (2005) study in which the BDI-II scores for Asian Americans and Caucasian Americans were compared. In this study, Caucasian Americans endorsed an item with higher scores on agitation compared to Asian Americans. Moreover, literature has also found that in some ethnic or cultural groups, depression accompanied with agitation is more likely to be associated with constructs other than depression, such as stress (Kim et al., 2016; Lawrence, 2014; Oei, Sawang, Goh, & Mukhtar, 2013).

The results from expert reviewers also identified that 5 of 21 items (3, 5, 6, 7 and 21) that did not have the same meaning for both cultural groups. The reason for these differences was attributed to the terms and expressions used in the items. The reviewers recommended using the specific term for certain items for Pakistani populations. For example, item 6 asked the participants about ‘punishment feelings’. However, adjusting the language of the question to be specific to Pakistani’s culture or religious orientation may have led to different results since there is a tendency to associate punishment feelings with religious beliefs (e.g. punishment from God).
In general, clinicians and researchers should be cautious when using the BDI-II on Pakistani participants since there is a tendency to have high scores on items 4 and 5 despite having lower depression scores. Furthermore, more somatic items are encouraged because depression is more likely to be presented by somatic items in non-Western culture compared to Western cultures (Bhui et al., 2004). For item 21 (loss of interest in sex) there is a higher chance that the Pakistani group of subjects would not answer this question due to cultural restrictions regarding sexual desire. This is supported by data in the current study as there was more missing data for item 21 compared to any other items of the BDI-II for Pakistanis.

It should also be noted that the data in the study were collected from undergraduate university students, who are less likely to be married. Moreover, Pakistani culture restricts expression of sexual desires in unmarried individuals, which could be the reason for missing data. Hence, item 21 was determined to be irrelevant as an indicator of depression.

### 5.1 Limitations

There were of course limitations with this study. First, there was a small sample size of only 200 participants in each group, which is often considered as a minimum sample size for conducting a CFA (Boomsma, 1983; Comrey & Lee, 1992; Hoelter, 1983) and DIF analysis (Reeve & Fayers, 2005).

For the DIF analysis, a larger sample size can increase the power in detecting items with DIF (Scott et al., 2009; Sébille et al., 2010). However, it is to be noted that the current study utilized three DIF methods, including the ordinal logistic regression (OLR) method, which is suitable for testing DIF in small samples. The results of the OLR method were also very similar to the other methods in this study. Hence, there is no visible impact of small sample size on the results.
For CFA, the larger sample sizes often improve the fit indices, however there is no rule of thumb for sample size (Meade & Bauer, 2007). There is a chance that the small sample size in the study affected the fit indices or difference tests, since research has shown that the chi-square difference tests are sensitive to a very large or small sample size (Brannick, 1995). Therefore, a more acceptable sample size could improve the fit indices.

Secondly, this study used data from university students, which were similar in age and education, limiting the generalizability of results to the general population. However, since there is a very limited knowledge on the equivalence of the BDI-II, the findings in this study still provide valuable insight into the equivalence of the BDI-II or other depression measures for Pakistani individuals.

5.2 Implications and Future Research

The instruments that are developed using a norming population from a specific country cannot be valid for the other populations. The BDI-II was originally normed on US population, with mostly Caucasian American participants. Researchers and clinicians should be aware of the psychometric evidence and equivalence of the measure before using it on certain populations.

To date, there has been no research providing evidence of equivalence of the BDI-II for Pakistani group and other South Asian groups. The results of the study deemed that BDI-II is not equivalent for the Pakistani group. The non-equivalent results of BDI-II in the study inform researchers and clinicians to exercise caution in using the BDI-II scores for Pakistani participants. Furthermore, the results of the study clearly demonstrated that there were certain items that were more salient for one group. For instance, the Pakistani participants endorsed items related to punishment and guilty feelings. This strong endorsement indicates the tendency to have higher scores on these items despite having an overall low score on depression.
Therefore, it should be noted that the depression score in different ethnic groups might not be representative of actual depression because of differences in its expression. If clinicians are aware of such differences, they could potentially avoid underdiagnoses or misdiagnoses of depression.

The findings of this study will also help researchers validate existing measures, and also to develop new measures. However, since this study was conducted in Canada, future research should be conducted in Pakistan to ensure the generalizability of results.

This study will also help understand the expression of depression in other South Asian ethnic groups (i.e. Sri Lankan, Indian, Bangladeshi, Nepalese), since they share a similar culture to Pakistanis. This similarity in cultures suggests that the findings from this study could be applied to other South Asian groups. However, a careful consideration should be taken in this regard since there are some differences within South Asian ethnic groups (Clarke et al., 1990).

In the current study, for the purpose of comparison, the reference group was kept very similar to the normative data used in Beck et al.’s (1996) study, of which the majority were Caucasian university students. Using a sample similar to normative data for the reference group assured that the differences observed in interpretation of item are not due to differences in sample demographics. The current study involved non-psychiatric university students. A further study can explore the equivalence in the Pakistani clinical sample by using a reference group similar to Beck et al.’s (1996) normative sample of clinical outpatients.
References


Lawrence, S. (2014). 31-year-old female shows marked improvement in depression, agitation, and panic attacks after genetic testing was used to inform treatment. Case Reports in Psychiatry, 2014, 1-4. doi:10.1155/2014/842349


Inventory: A five-country analysis from the ODIN study. *Journal of Affective Disorders, 114*(1), 156-162. doi:10.1016/j.jad.2008.06.021


Appendices

Appendix A: Questionnaire

Participant Demographic Form

Measurement Equivalence of Beck Depression Inventory (BDI-II)

1. What is your age? ________
2. Identify your gender?
   a) Male  b) Female  c) Other, please specify ________
3. Identify your race and specify ethnicity in the blank?
   a) South Asian........... (Pakistani, Indian, Sri Lankan, Bangladeshi, Nepalese etc.) ________
   b) Southeast Asian...... (Filipino, Indonesian, Malaysian, Vietnamese, etc.) _____________
   c) Caucasian............ (German, Canadian, Scottish, Russian, Polish, etc.) _______________
   d) Mediterranean........ (Italian, Greek, Maltese, Spanish, etc.) _______________
   e) Black..................... (Jamaican, African-American, Ethiopian, Kenyan, etc.) __________
   f) Hispanic................ (Mexican, Latin American, Honduran, etc.) ______________
   g) Middle Eastern........ (Persian, Turkish, Lebanese, Egyptian, Kurdish, etc.) __________
   h) Native American....... (Cherokee, Iroquois, Metis, etc.) __________
   i) Other/mixed.......... Please specify________
4. Year of study _____

5. (a) What is your country of birth? ____________?
   (b) If you are not born in Canada, how long have you been in Canada? ____________

6. What describes you best?
   - First Generation Canadian (born outside Canada)
   - Second Generation Canadian (born in Canada with at least one parent born outside Canada)
   - Third Generation Canadian (born in Canada with both parents born in Canada)
   - Other ______________

7. (a) How would you rate your fluency in English?
   1  2  3  4  5
   (not fluent) (very fluent)

   (b) What is your first language of choice ____________, please rate your fluency
   1  2  3  4  5
   (not fluent) (very fluent)

   (c) What is your second language of choice ____________, please rate your fluency
   1  2  3  4  5
   (not fluent) (very fluent)

~Thank you for your participation~

.unlink
.unlink
.unlink
.unlink

Please tear along the dotted line above and leave it in an envelope provided by the researcher.

- Are you interested in the final results of the study?  □ Yes  □ No
- Would you like to be considered in a draw of a $50 gift card?  □ Yes  □ No

If you checked ‘yes’ to any of the above question, leave your email address: ____________
Appendix B: Professional Profile and Feedback Form

Professional Profile

Measurement Equivalence of Beck Depression Inventory (BDI-II)

8. Please provide details about your experience working with the Pakistani clinical population in Canada or anywhere else (for more space, please attach a paper).

____________________________________________________________________________

____________________________________________________________________________

____________________________________________________________________________

____________________________________________________________________________

9. Please provide details about your education and training (for more space, please attach a paper).

____________________________________________________________________________

____________________________________________________________________________

10. In what province do you currently practice?
   a) British Columbia
   b) Ontario
   c) Other (please specify) _______________

11. Would you like to be identified/acknowledged in publications/graduate thesis? Please explain?

____________________________________________________________________________

____________________________________________________________________________
Feedback Form

Measurement Equivalence of Beck Depression Inventory (BDI-II)

Purpose of the study
Researchers and psychologists often employ tests/assessment tools that are not suitable for use with diverse populations. There are often certain linguistic or cultural elements in a test that effect the score of the individuals, threatening the validity of assessment/diagnostic decisions made based on test results.

The purpose of the study is to examine if a common psychological assessment tool i.e. BDI-II is culturally appropriate to use with Pakistani population.

Instructions:
In the first column, you are asked to rate each item from 0 to 3 on an extent to which the items of BDI-II are comparable for Pakistani culture. If you rate any item less than 3, please explain in the next columns if there is any item’s term or expression that seem to have different meaning. Also, explain if you have any suggestions to improve the item.

Table adapted from Benítez et al. (2016) and Ercikan (2002)

<table>
<thead>
<tr>
<th>Questions</th>
<th>Items are comparable across Pakistani culture (Rate from 0 to 3)</th>
<th>If items are not comparable across culture, is there any term with different meaning</th>
<th>If items are not comparable across culture, is there any expression with different meaning</th>
<th>Other concerns</th>
<th>How the item can be rewritten?</th>
<th>Suggestions for improvement</th>
</tr>
</thead>
<tbody>
<tr>
<td>Item 1</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Item 2</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Item 3</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Item 4</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>-------</td>
<td>---</td>
<td>---</td>
<td>---</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Item 5</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Item 6</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Item 7</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Item 8</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Item 9</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Item 10</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Item 11</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Item 12</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Item 13</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Item 14</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Item 15</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Item 16</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Item 17</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Item 18</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Item 19</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Item 20</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Item 21</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

~Thank you for your participation~
Appendix C: Referral List for Resources in Greater Toronto Area

Referral List for Community Resources

Measurement Equivalence of Beck Depression Inventory (BDI-II)

If you think that you are experiencing depression or sadness and you would like to deal with it, the following resources might help.

1. Family Service Toronto - Walk-In Counselling
   355 Church Street
   Toronto ON M5B 1Z8

2. South Asian Community Health Services
   25 Lady Stewart Boulevard
   Brampton ON L6S 3Y2

3. Brampton Civic Hospital- Outpatient Psychiatry
   2100 Bovaird Drive E
   Brampton, ON L6R 3J7

4. WoodGreen Community Services – Toronto
   815 Danforth Avenue
   Toronto, ON M4J 1L2

For immediate assistance:

1. Toronto Distress Centre
2. Contact Centre Telecare Peel
Appendix D: Informed Consent

Sample Consent Form

Measurement Equivalence of Beck Depression Inventory (BDI-II)

Principal Investigator: Dr. Kadriye Ercikan, professor in the Department of Educational and Counselling Psychology, and Special Education (ECPS) at the University of British Columbia (UBC) is the faculty supervisor overseeing this research.

Co-Investigator: Rida Waseem, a Master’s student in Measurement, Evaluation, and Research Methodology (MERM) program in the Department of Educational and Counselling Psychology, and Special Education (ECPS) at The University of British Columbia (UBC), is conducting this research for her Master’s thesis.

Study Introduction:
You are invited to take part in a study that will examine if a depression questionnaire is a helpful measure for diverse cultures.

Voluntary Participation:
Your participation is voluntary. You have the right to refuse to participate in this study. Before you decide, it is important for you to understand what might be involved. This consent form will confirm that you understand the purpose of this study, your role in this study and your rights as a participant. If you decide to participate, your decision is not binding and you may choose to withdraw from the study at any time without any consequences.

Why are we doing this study?
Research has shown that individuals from different cultures describe depression differently. The purpose of this study is to determine if the questionnaire measuring depression, called Beck Depression Inventory (BDI-II), is a helpful measure of depression in diverse cultural groups.

What is involved in this study?
If you decide to take part in the study you will be asked to complete a brief questionnaire. The questionnaire will have two parts. The first part includes questions about demographic information. The second part includes a set of 21 questions asking about symptoms of depression. you will be asked to rate your severity of symptoms.
How long will it take?
The questionnaire will take 5-10 minutes to complete. However, it will not be timed and you may take as long as you want on any question.

Results of the study:
The results of the study will be reported in a graduate thesis and may also be published in a journal articles and/or books.

If you are interested in obtaining the results of the study, you will be asked to check a box and provide an email address at the end of the survey questionnaire. You will be asked to tear the email address and leave it in the envelope provided by the researcher. The email address will be kept separate from your responses on the questionnaire.

What are the benefits of participating?
There are no direct benefits anticipated from participating in the study. However, your responses might help researchers to learn more about depression in diverse cultural and ethnic groups.

Is there any way being in this study could be bad for you?
We do not think there is anything in this study that could harm you or be bad for you. Some of the questions may seem distressing. If at any time you feel uncomfortable answering the study questions, you do not have to answer the question and you have a right to withdraw from the study. Please let the researcher know if you have any concerns. At the end of the survey, a document outlining options for community referral for mental health support will be offered.

How will your privacy be maintained?
Your confidentiality will be respected. Information that discloses your identity will not be released. The survey questionnaire will identify you only by code number. The completed survey will be kept in a locked filing cabinet. If you identify yourself by name or use any other identifying information intentionally or unintentionally, it will be erased from the survey as soon as it is noticed.

The data extracted from the paper based survey questionnaires, will be stored in a password protected computer and files. The participants will not be identified in any reports of the completed study by name or any other identifying information.

Will you be paid for your time/taking part in this research study?
At the end of survey, you will be asked to check a box and leave your email if you would like to be considered in the draw of a $ 50 gift card. You will be asked to tear the email address and leave it in an envelope provided by the researcher. The email address will be kept separate from your responses on the questionnaire.

Who can you contact if you have questions about the study?
If you have any questions about the study. Please contact Rida Waseem or Dr. Kadriye Ercikan.

If you have any concerns or complaints about your rights as a research participant and/or your experiences while participating in this study, contact the Research Participant Complaint Line in
the UBC Office of Research Ethics at 604-822-8598 or if long distance e-mail RSIL@ors.ubc.ca or call toll free: 1-877-822-8598.

Consent
Taking part in the study is entirely up to you. You have the right to refuse to participate in this study. If you decide to take part, you may choose to pull out of the study at any time without giving a reason and without any negative impact on you.

If the questionnaire is completed, it will be assumed that consent had been given.