Measurement Equivalence of the Center for Epidemiological Depression Scale in Racially/Ethnically Diverse Older Adults

by

Giyeon Kim

A dissertation submitted in partial fulfillment of the requirements for the degree of Doctor of Philosophy
School of Aging Studies
College of Arts and Sciences
University of South Florida

Co-Major Professor: David A. Chiriboga, Ph.D.
Co-Major Professor: Yuri Jang, Ph.D.
Victor Molinari, Ph.D.
Larry Polivka, Ph.D.
Brent J. Small, Ph.D.

Date of Approval:
September 10, 2007

Keywords: Depressive Symptoms, Measurement Equivalence, Health Disparities, Differential Item Functioning, CES-D

© Copyright 2007, Giyeon Kim
# TABLE OF CONTENTS

LIST OF TABLES .............................................................................................................. v

LIST OF FIGURES ........................................................................................................... vi

ABSTRACT ...................................................................................................................... vii

CHAPTER ONE: INTRODUCTION .................................................................................. 1
   Introduction and Statement of the Research Problem .............................................. 1
   Purpose of the Study ............................................................................................... 2
   Significance of the Study ......................................................................................... 5
   Background ............................................................................................................. 6
       Two Approaches for Cross-Cultural Depression Research ............................... 7
           Emic Approach to Cross-Cultural Depression Research ............................ 7
           Etic Approach to Cross-Cultural Depression Research ......................... 9
       Conceptual Model for Cross-Cultural Depression Research ....................... 13
       Measurement Equivalence in Cross-Cultural Depression Research ............. 17
           Establishing Conceptual Equivalence ......................................................... 18
           Establishing Metric Equivalence ............................................................... 19
           Establishing Structural Equivalence .......................................................... 20
       Cross-Cultural Research with the CES-D Scale ............................................. 22
           Description of the CES-D Scale ................................................................. 22
           DSM Criteria and the CES-D .................................................................. 24
           Cross-Cultural Applicability of the CES-D ............................................... 26
       Measurement Equivalence of the CES-D Scale ............................................. 31
       Analytic Strategies for Measurement Equivalence Research ....................... 35
CHAPTER TWO: STUDY 1

RACE/ETHNICITY-RELATED MEASUREMENT BIAS ............................................ 45
Introduction........................................................................................................... 45
Methods................................................................................................................ 51
Sample................................................................................................................... 51
Measures .............................................................................................................. 52
  CES-D Scale ...................................................................................................... 52
Statistical Analysis ............................................................................................... 52
  CFA DIF detection ............................................................................................ 54
  IRT DIF detection ............................................................................................ 56
Results.................................................................................................................. 57
  Descriptive Information of Sample................................................................. 57
  Descriptive Item Statistics ............................................................................... 59
  DIF Results ...................................................................................................... 59
    Whites versus Blacks ...................................................................................... 64
    Whites versus Mexican Americans ............................................................... 64
    Mexican Americans versus Blacks ............................................................... 65
    CFA versus IRT ............................................................................................. 65
Discussion............................................................................................................. 66

CHAPTER THREE: STUDY 2

SOCIODEMOGRAPHIC-RELATED MEASUREMENT BIAS ......................... 74
Introduction........................................................................................................... 74
Methods................................................................................................................ 78
Sample................................................................................................................... 78
Measures .............................................................................................................. 79
CHAPTER FOUR: STUDY 3

ACCULTURATION- AND LANGUAGE-RELATED MEASUREMENT BIAS ...... 100

Introduction....................................................................................... 100

Methods............................................................................................ 104

Sample............................................................................................... 104

Measures .......................................................................................... 104

CES-D scale .................................................................................... 104

Acculturation .................................................................................... 105

Instrument Language ........................................................................ 106

Acculturation × Language ............................................................... 106

Statistical Analysis ........................................................................ 106

Dimensionality test ......................................................................... 106

DIF analyses .................................................................................... 107

CFA DIF Detection .......................................................................... 108

IRT DIF Detection .......................................................................... 108

Results ............................................................................................. 109
Sample Characteristics .................................................................................. 109
DIF Results .................................................................................................. 111
  Acculturation-bias .................................................................................. 111
  Language-bias ......................................................................................... 114
  Acculturation × Language-bias .................................................................. 114
Discussion ..................................................................................................... 116

CHAPTER FIVE: DISCUSSION ....................................................................... 123
  Summary of the Study ............................................................................... 123
  Overview of Findings ............................................................................... 123
  Implications .............................................................................................. 125
    Methodological Implications .................................................................. 125
      Detecting DIF ...................................................................................... 125
      Explaining DIF .................................................................................. 125
      Dealing with DIF ................................................................................ 126
    Practice Implications ............................................................................ 128
  Limitations and Future Research ............................................................ 131
  Final Thoughts ......................................................................................... 134

REFERENCES .............................................................................................. 136

ABOUT THE AUTHOR .................................................................................. End Page
LIST OF TABLES

Table 1. Study 1 – Item Parameter Estimates for Whites, Blacks, and Mexican Americans ............................................................................................................... 55

Table 2. Study 1 – Descriptive Characteristics of the Sample.......................... 58

Table 3. Study 1 – Mean and Standard Deviation (SD) of the CES-D in Whites, Blacks, and Mexican Americans ................................................................. 60

Table 4. Study 1 – Item Responses of the CES-D Scale in Whites, Blacks, and Mexican Americans ................................................................. 61

Table 5. Study 1 – DIF Results from CFA and IRT Methods ................................ 62

Table 6. Study 2 – Descriptive Characteristics of the Sample.......................... 84

Table 7. Study 2 – Age DIF Results from CFA and IRT Methods ..................... 87

Table 8. Study 2 – Gender DIF Results from CFA and IRT Methods ............... 89

Table 9. Study 2 – Educational Attainment DIF Results from CFA and IRT Methods .................................................................................................................. 91

Table 10. Study 3 – Descriptive Information of the High and Low Acculturated Mexican Americans ................................................................. 110

Table 11. Study 3 – DIF Results from Confirmatory Factor Analysis (df = 341, Δ df = 2) .................................................................................................................. 112

Table 12. Study 3 – DIF Results from Item Response Theory (df = 157, Δ df = 4) .................................................................................................................. 113
LIST OF FIGURES

Figure 1. The Typology of the Relationship of Culture to Depression Adopted from the Sternberg (2004) Culture and Intelligence Model................................. 14

Figure 2. Outline for the Dissertation ............................................................................... 44

Figure 3. Item information function for Item 17 (‘I had crying spells’) showing Gender-DIF in the total sample and Mexican American race/ethnicity. ........... 90

Figure 4. Item information function for Item 17 (‘I had crying spells’) showing Education-DIF in the total sample and Black race/ethnicity. ............................ 92
This dissertation study was designed to examine measurement equivalence of the Center for Epidemiological Studies Depression (CES-D) Scale across White, African American, and Mexican American elders. Specific aims were to identify race/ethnicity-, sociodemographic-, and acculturation and instrument language-related measurement bias in the CES-D. Three studies were conducted in this dissertation to accomplish these aims. Two existing national datasets were used: the New Haven Established Populations for Epidemiologic Studies of the Elderly (EPESE) for the White and African American samples and the Hispanic Established Populations for Epidemiologic Studies of the Elderly (H–EPESE) for the Mexican-American sample. Differential item functioning (DIF) analyses were conducted using both confirmatory factor analysis (CFA) and item response theory (IRT) methods. Study 1 focused on the role of race/ethnicity on the measurement bias in the CES-D. Results from Study 1 showed a lack of measurement equivalence of the CES-D among Mexican Americans in the comparison with both
Whites and Blacks. Race/ethnicity-specific items were also identified in Study 1: two interpersonal relation items in Blacks and four positive affect items in Mexican Americans. Study 2 focused on identifying sociodemographic-related measurement bias in responses to the CES-D among diverse racial/ethnic groups. Results from Study 2 showed that gender and educational attainment affected item bias in the CES-D. The interaction between gender and educational level and race/ethnicity was also found in Study 2: Mexican American women and lower educated Blacks had a greater predisposition to endorse the ‘crying’ item. Focusing on Mexican American elders, Study 3 examined how level of acculturation and language influence responses to the CES-D. In Study 3, acculturation and instrument language-biased items were identified in Mexican American elders. Study 3 also suggested that acculturation-bias was entirely explained by whether the CES-D was administered in the English or the Spanish versions. Possible reasons for item bias on the CES-D are discussed in the context of sociocultural differences in each substudy. Findings from this dissertation provide a broader understanding of sociocultural group differences in depressive symptom measures among racially/ethnically diverse older adults and yield research and practice implications for the use of standard screening tools for depression.
CHAPTER ONE: INTRODUCTION

Introduction and Statement of the Research Problem

With the increased attention paid to racial and ethnic disparities in mental health, a growing body of literature suggests racial/ethnic disparities in depression among older adults (Perreira, Deeb-Sossa, Harris, & Bollen, 2005; Robison, Curry, Gruman, Covington, Gaztambide, & Blank, 2003). A number of studies have reported higher rates of depressive symptoms and disorders in some racial/ethnic minority groups than in White Americans (e.g., Dunlop, Song, Lyons, Manheim, & Chang, 2003; Minsky, Vega, Miskimen, Gara, & Escobar, 2003). However, reported prevalence rates of depression vary dramatically across diverse racial/ethnic groups, ranging from 1.5% to 25.4% (e.g., Mui, Burnette, & Chen, 2002; Saez-Santiago & Bernal, 2003; Weissman, Bland, Canino, Faravelli, Greenwald, Hwa et al., 1996). Given that culture—defined broadly, including racial/ethnic, age, gender, socioeconomic, geographical, and language variations—shapes values, attitudes, beliefs, and behaviors by a group of people (Sternberg, 2004), as well as influences disease manifestation and diagnostic labeling (Kleinman, 2004), cultural differences among racially/ethnically diverse elderly groups may affect the way in which people experience and express their depressive symptoms, which in turn may contribute to variations in the prevalence rates of depression.

A number of comparative studies on depressive symptoms with diverse racial/ethnic groups and even with subgroups of the major racial/ethnic groups have sought to identify similarities and differences, and explain differences across these groups
(Gregorich, 2006). While these efforts are praiseworthy, questions can be raised about the cross-group validity of the depression measures used. Group mean comparisons of depressive symptoms assume measurement invariance in which the same construct is being measured and the measurement used for comparison functions in the same way across all groups (Bravo, 2003; van de Vijver, 2001). Violations of this assumption can occur if individuals from particular cultures or subgroups are more or less likely to endorse specific items as a consequence of group membership, such as race/ethnicity, age, gender, educational experience, and language of administration (Jones, 2006). Researchers, however, have not paid enough attention to the cultural equivalence of depressive symptom measures, taking the underlying methodological assumptions for granted. Failure to establish measurement equivalence on depression may lead to inaccurate estimates of prevalence and misleading group comparisons (Hui & Triandis, 1985; Vandenberg & Lance, 2000). For this reason, comparative research on depressive symptoms is needed to distinguish differences arising from a lack of measurement equivalence, which is known as differential item functioning (DIF), from true differences in standing on the same latent trait.

Purpose of the Study

In light of the abovementioned lack of research on the cultural equivalence of depressive symptom measures, the main purpose of this dissertation study was to examine measurement equivalence of one of the most widely used depression assessment tools, the Center for Epidemiological Studies Depression Scale (CES-D), across diverse racial/ethnic groups and subgroups of the major racial/ethnic groups. Focusing on comparisons between three racial/ethnic elderly groups including Whites, Blacks, and
Mexican Americans, specific aims of this dissertation research were to identify culture-specific items in the CES-D scale functioning differently across diverse racial/ethnic groups and subgroups of age, gender, educational attainment, acculturation, and language within each racial/ethnic elderly group as well as to identify a core set of depressive symptom items in the CES-D scale across racially/ethnically diverse elderly groups and subgroups within the same racial/ethnic group.

The specific research questions for this dissertation study were the following:

*Research Question 1:* Do the measurement properties of the CES-D scale vary by race/ethnicity among older adults?; *Research Question 2:* Do the measurement properties of the CES-D scale vary by sociodemographic characteristics (i.e., age, gender, and educational attainment) within each racial/ethnic elderly group?; *Research Question 3:* Do the measurement properties of the CES-D scale vary by the level of acculturation and instrument language in older Mexican-Americans? Each research question was examined in a separate study (Study 1, 2, and 3).

In order to investigate three research questions, this dissertation study used two existing national datasets from the New Haven Established Populations for Epidemiologic Studies of the Elderly (EPESE) for older Whites and Black samples and the Hispanic Established Populations for Epidemiologic Studies of the Elderly (H–EPESE) for older Mexican American sample. The following data analyses were conducted: (1) ANOVA and t-test; (2) dimensionality test using exploratory factor analysis (EFA) and confirmatory factor analysis (CFA); and (3) differential item functioning (DIF) analyses using both confirmatory factor analysis (CFA) and item response theory (IRT) methods.
Research Question 1 was examined in Study 1 entitled “Measurement Equivalence of the Center for Epidemiological Studies Depression Scale Among White, Black, and Mexican American Respondents in the New Haven and the Hispanic EPESE.” The purpose of Study 1 was to identify race/ethnicity-related measurement bias in responses to the CES-D items in three racial/ethnic elderly groups. Study 1 hypothesized that the measurement properties of the CES-D scale would vary by race/ethnicity among older adults.

Research Question 2 was examined in Study 2 entitled “Sociodemographic-Related Measurement Bias in the Center for Epidemiological Studies Depression Scale: A Study of Racially/Ethnically Diverse Older Adults.” Study 2 focused on identifying sociodemographic-related measurement bias in response to the CES-D in three racial/ethnic elderly samples. Study 2 hypothesized that the measurement properties of the CES-D scale would vary by age, gender, and educational attainment within each racial/ethnic elderly group.

Focusing on older Mexican Americans, Research Question 3 was examined in Study 3 entitled “Measurement Bias in the Center for Epidemiological Studies Depression Scale in Older Mexican Americans: Effects of Acculturation and Language.” The purpose of Study 3 was to identify how level of acculturation and language influence responses to the CES-D in Mexican American elders. Study 3 hypothesized that the measurement properties of the CES-D would vary by the level of acculturation and instrument language in older Mexican Americans.
Significance of the Study

The goal of the overall dissertation is to evaluate the measurement equivalence of the CES-D across three racial/ethnic groups (Whites, Blacks, and Mexican Americans). This dissertation study represents the first multi-racial/ethnic and multi-cultural evaluation of the full version of the CES-D using older adults from two nationally representative datasets. Previous analyses have been limited to one or two racial/ethnic groups to evaluate the measurement equivalence of the CES-D. The effects of sociodemographic and cultural factors on symptoms and expressions of depressive symptoms can be explored and compared in this dissertation. More importantly, this dissertation study can identify sociodemographic- and culture-specific depressive symptom items that function differentially across cultures and sub-cultural groups within the same racial/ethnic group. This dissertation study can also identify a core set of depressive symptom items that function equivalently across cultures and subgroups of age, gender, educational attainment, and acculturation. This dissertation can eventually address limitations of current screening for depression in older racial/ethnic groups and improve the detection of depression among diverse groups of older adults, by providing the modified versions of the CES-D in each group. Identification of optimal depression screens may provide a basis for developing culturally appropriate and effective depression assessment tools in different racial/ethnic groups within the U.S. and cross-nationally.

Findings from this dissertation study can provide directions for future research and practice in the field. This study may be useful for screening, assessment, and treatment for depression in racially or ethnically diverse groups. When established
measures are used to screen and assess for depression in different racial/ethnic groups, we should recognize the risk that individuals with different cultural backgrounds may tend to be misclassified, leading to under- or over-diagnosis for depression. This may be related to under- or over-reporting, as well as under- or over-treatment for depression. This dissertation study may help inform the field with respect to culturally appropriate screens for depression in racially/ethnically diverse groups and sub-cultural groups within the same racial/ethnic group. This dissertation can be extended to other racial/ethnic groups, as well as other diverse cultural subgroups with important status characteristics. This may be particularly important when consequences of inaccurate measures could lead to misguided public policies. Finally, this line of research may be extended to other mental health measures to enable accurate comparisons among diverse racial/ethnic groups for future research.

Background

In this background section, this dissertation begins with three conceptual issues that underlie concerns about measurement equivalence: (1) two approaches for cross-cultural depression research including emic and etic approaches to cross-cultural depression research; (2) a conceptual framework for cross-cultural depression research, along with implications for different conceptualizations of depression and recommendations for future research; and (3) measurement equivalence issues in cross-cultural depression research including conceptual, metric, and structural equivalence. With special attention to one of the most widely used depression assessment tools, the Center for Epidemiologic Studies-Depression Scale (CES-D) scale, the cross-cultural applicability of the CES-D scale is discussed next. Focusing more on measurement
equivalence research, factors affecting measurement equivalence of the CES-D and analytic strategies for measurement equivalence are reviewed. Based on the extensive background for this dissertation, issues needed for future research are also presented in the following section.

Two Approaches for Cross-Cultural Depression Research

A central issue in cross-cultural comparative studies involves two different perspectives, which together have been called the ‘emic-etic paradigm’ (Brislin, Lonner, & Thorndike, 1973; Canino, Lewis-Fernandez, & Bravo, 1997; Pike, 1954; Triandis & Brislin, 1984). The emic approach reflects the inside perspective of the ethnographer, whereas the etic approach reflects the outside perspective of the comparativist researcher (Morris, Leung, Ames, & Lickel, 1999; Ng & Earley, 2006). The emic approach uses variables and observations that are culturally specific to a particular group, at a certain period in time, to develop an instrument (Rait & Burns, 1998). This does not allow for comparative research as it looks at variables in terms of language and culture and the instrument may not be relevant to other groups. The etic approach, on the other hand, is basically comparative, and is directed at eliciting standardized categories of phenomena out of local specificities (Canino et al., 1997).

Emic Approach to Cross-Cultural Depression Research

The emic approach attempts to describe the internal logic of a culture, its singularity, considering this a necessary step prior to any valid cross-cultural analysis (Chávez & Canino, 2005). Therefore, it does not allow cross-cultural comparisons using identical case definitions and standardized diagnostic interviews as case-finding instruments (Cheng, 2001). However, given that the emic approach focuses on examining a construct from
within a specific culture and understanding that construct as the people from within that
culture understand it (Schaffer & Riordan, 2003), the emic approach is particularly useful
in understanding the relatively unique features of the manifestation of depressive symptoms
and diagnosis in a given ethnic group.

For example, Patel and colleagues (2001) examined medical concepts of depression
in Zimbabwe, and found that one culture-specific terminology, kufungisisa (thinking too
much), has been shown to be closely linked to depression among Zimbabwean depressed
patients. Similarly, some researchers with an emic approach have highlighted that each
culture has its own somatic metaphors to describe depression: nervios (brain ache or
uncontrollable) in Mexican Americans (Jenkins, 1988); shenjing shuairuo (neurasthenia) in
Chinese (Parker, Gladstone, & Chee, 2001); Sadri dayeq alayya (My chest feels tight) and
Jesmi metkasser (broken body) in Dubai (Sulaiman, Bhugra, & De Silva, 2001). Along the
same lines, the emic approach has also influenced DSM-IV, which now includes an index
of culturally defined syndromes, as well as statements describing how culture influences
the prevalence, symptomatology, course, and clinical outcome of specific psychiatric
disorders (Canino et al., 1997).

Emic studies of depression have succeeded in developing some indigenous
depression instruments (Tanaka-Matsumi, 2001). For example, with an additional
inclusion of five Hopi illness categories (translated as worry, sickness, unhappiness,
drunken-like craziness, and disappointment), Manson, Shore, and Bloom (1985) developed
the American Indian Depression Scale (AIDS) among Hopi Indians. Other examples of
locally developed screening questionnaires include the Primary Care Psychiatric
Questionnaire (PPQ; Stinivasan & Suresh, 1990) in India, the Shona Symptom
Questionnaire (Patel, Simunyu, Gwanzura, Lewis, & Mann, 1997) in Zimbabwe, and the Chinese Health Questionnaire (CHQ; Cheng & Williams, 1986). Also, there are a few examples of indigenous structured interviews, including the Indian Psychiatric Survey Schedule (Shamasundar, Krishna Murthy, Prakash, Prabhakar, & Subbakrishna, 1986).

Although these indigenous depression measures share much with the questionnaires developed in Western culture and there is a high degree of agreement in care classification, the emic studies on depression have been criticized for neglecting the problem of observation bias (Canino et al., 1997). One of the critical issues in cross-cultural depression research with an emic approach is the lack of methodological homogeneity across studies focusing on different cultures. This can result in the inability to disentangle methodological from substantive factors when variability in cross-cultural comparisons is observed (Bravo, 2003). For example, as mentioned earlier, although Patel and colleagues (2001) suggested that culture-specific somatic symptoms are strongly associated with depression among Zimbabweans, criticisms may remain with respect to generalizability of the indigenous measurement and epidemiological testing of causal hypotheses on depression.

**Etic Approach to Cross-Cultural Depression Research**

The main assumption with the etic cross-cultural research on depression is that the etiology of depression is universal and key constructs of depression exist equally across all cultures. Epidemiologists and cross-cultural researchers often use the etic approach for the cross-cultural comparative study of depression (e.g., Breslau, Aguilar-Gaxiola, Kendler, Su, Williams, & Kessler, 2005; Cole, Kawachi, Maller, & Berkman, 2000; Gonzalez, Haan, & Hinton, 2001; Iwata, Turner, & Lloyd, 2002; Kessler & Ustun, 2004; Nguyen, Kitner-
Triolo, Evans, & Zonderman, 2004), emphasizing the search for equivalence across cultures and using similar methods, constructs, and measures across settings in order to increase the generalizability of the findings (Schaffer & Riordan, 2003; van de Vijver, 2001).

A prototypical etic study of depression is the international project of the World Health Organization (1983) on the diagnosis and classification of depression in Switzerland, Canada, Japan, and Iran. The goal of this study was to test feasibility of using standardized instruments of depressive disorders. Using the Schedule for Standardized Assessment of Depressive Disorders (WHO/SADD) by psychiatrists, 573 patients were diagnosed with depression in this study. On the basis of 39 symptoms of depression, WHO (1983) found that more than 76% of the depressed patients reported core depressive symptoms that included “sadness, joylessness, anxiety, tension, lack of energy, loss of interest, loss of ability to concentrate, and ideas of insufficiency” (p. 61). Suicidal ideation was also present in 59% of patients. The WHO project also discovered cross-cultural variation in the expression of depression, such as somatic complaints and obsessions which were not part of the original 39 symptoms of depression measured by the WHO/SADD. Variations existed both within and across cultures in this project.

More recently, WHO has also conducted several cross-national studies on depression. The cross-national depression study of Simon and his colleagues (2002) is an example. On the basis of the World Health Organization’s Psychological Problems in General Health Care (PPGHC) study, they examined prevalence rates of depression in 14 countries. The PPGHC study used the Composite International Diagnostic Interview (CIDI) for psychiatric symptoms and diagnoses. The authors found evidence that
prevalence rates of current major depression varied across cultures. They also reported that depression was universally associated with disability, but this association varied across cultures. They concluded that use of identical measures and diagnostic criteria might identify different levels of depression severity in different countries or cultures.

A number of studies with an etic approach have reported that even within the United States, prevalence estimates varied across racial/ethnic groups (e.g., Breslau et al., 2005; Dunlop et al., 2003; Gonzalez et al., 2001; Swenson, Baxter, Shetterly, Scarbro, & Hamman, 2000). For example, using a nationally representative sample from the National Comorbidity Survey Replication (NCS-R), Breslau and colleagues (2005) examined racial/ethnic variations in DSM-IV disorders. The Composite International Diagnostic Interview (CIDI) was used for the diagnostic assessment. Comparing non-Hispanic Whites, non-Hispanic Blacks, and Hispanics, the authors found significantly lower lifetime prevalence and risk of depression in both minority groups. Suggesting the presence of common protective factors across disorders for both minority groups, the authors concluded cultural differences might lead racial/ethnic minority groups to respond differently to the same survey questions regarding their psychiatric history despite similar levels of morbidity.

Reflecting on the importance of cultural influences on reporting depressive symptoms and the cultural appropriateness of measures, etic researchers have examined psychometric properties of depression measures (e.g., Crockett, Randall, Shen, Russel, & Driscoll, 2005; Foley, Reed, Mutran, & DeVellis, 2002; Nguyen et al., 2004). For example, using exploratory factor analysis, Foley and colleagues (2002) did not confirm Radloff’s original four-factor model in older African Americans. They found no distinction between
somatic complaints and depressed affect, and they identified one new factor, ‘social well-being’, that has not been reported in the general population. This suggests the presence of unique measurement properties of the CES-D in African Americans, as well as needs for additional research to assess the validity of CES-D across diverse cultural groups.

Etic studies of depression have developed a number of depression instruments. Assessment of depressive symptoms and disorder has been conducted with self-report (e.g., Center for Epidemiologic Studies Depression, Beck Depression Scale, Zung Depression Scale, Geriatric Depression Scale) and interviewer or clinician rating scales (e.g., Hamilton Rating Scale for Depression, Structured Clinical Interview for DSM-IV, WHO Composite International Diagnostic Interview). Although these self-report and interviewer rating instruments are based on symptom criteria that is geared to Western culture patients, these instruments have been used as standards for depressive symptoms across ethnic and cultural groups. As a result, the etic approach has been criticized for emphasizing reliability at the expense of validity (Canino et al., 1997). It may impose the appearance of cross-cultural homogeneity at the expense of a constricted conceptualization embedded in the instrumentation (Bravo, 2003). This limitation has been called the “cultural fallacy,” meaning the approach ignores cross-cultural differences in the nature or validity of depressive disorder (Bravo, 2003). For example, Medina-Mora and colleagues (2005) examined psychiatric disorders in Mexico using the CIDI instrument. However, given that culture-bound syndromes such as nervios exist in Mexico and may not be captured by the CIDI, the instrument based only on DSM-IV may have potential problems due to cultural fallacy.
Conceptual Model for Cross-Cultural Depression Research

A significant limitation of cross-cultural research conducted in the field of depression is that it lacks a clear theoretical framework. In order to advance cross-cultural comparative research on depressive symptoms, Sternberg’s (2004) culture and intelligence analytic model presents as one of the more appropriate analytic frameworks, although it was originally developed with a focus on cross-cultural intelligence research. Sternberg (2004) proposed a fourfold typology to capture the possible relationship of culture to intelligence. The four cells of the typology differ in two key respects: (1) whether or not there are cross-cultural differences in the nature of the mental processes and representations involved in adaptation that constitute intelligence and (2) whether there are differences in the instruments needed to measure intelligence (beyond simple translation or adaptation), as a result of cultural differences in the content required for adaptation (Sternberg, 2004, p.326).

The typology can be applied to cross-cultural comparative research on depression by reframing the previous discussion of etic and emic paradigm and of depression instruments used in the previous cross-cultural depression research. Taking all the above-mentioned conceptual considerations into account, we propose four types of the relationship of culture to depression, that are shown in Figure 1. The four types are summarized as follows: (1) Type I, the nature of depression is the same across cultures and the instruments used to measure it are the same; (2) Type II, the nature of depression is different across cultures, but the instruments used to measure it are the same across cultures; (3) Type III, the nature of depression is the same across cultures, but the instruments are not the same; (4) Type IV, the nature of depression is different across cultures and the instruments used to measure it are different.
### Dimensions of Depression

<table>
<thead>
<tr>
<th>Relation</th>
<th>Same</th>
<th>Different</th>
</tr>
</thead>
</table>
| Same     | Type I - Etic approach - Type I-based Depression Research  
  · Research on core depressive symptoms  
  · Research on depression prevalence rates  
  · Research on metric & structural equivalence of depression measures  
  · WHO projects based on DSM-based measures  
 | Type II - Etic approach - Type II-based Depression Research  
  · Research on different factor structure of instrument  
  · Research on different expression on depression  
  · Research on metric & structural non-equivalence of depression measures  
  · Research on somatization and depression |
| Different| Type III - Emic approach - Type III-based Depression Research  
  · Research on locally developed screening tools  
  · Research on symptom expression within one culture  
 | Type IV - Emic approach - Methods: focus groups, in-depth interview - Type IV-based Depression Research  
  · Research on culture-bound syndromes  
  · Research on somatization and depression  
  · Research on metric & structural equivalence of depression measures |

*Figure 1. The Typology of the Relationship of Culture to Depression Adopted from the Sternberg (2004) Culture and Intelligence Model*
In Type I, both the instruments and the ensuing dimensions of depression are the same across cultures. Studies that found core depressive symptoms across cultures using DSM-based measures are good examples, such as a prototypically etic study of depression (WHO, 1983). Indeed, a number of WHO studies of cross-national prevalence rates using DSM criteria (e.g., Breslau et al., 2005) are also based on this type. This line of research assumes that core depressive symptoms exist cross-culturally and cross-nationally and only levels of depressive symptoms are different across cultures. The argument based on Type I is that the nature of depression is precisely the same across cultures and that this nature can be measured identically without regard to culture using appropriate translations when necessary.

In Type II, the measures used to assess depression are the same across cultures, but the outcomes obtained from using those measures are structurally different as a function of culture. This type is close to a number of etic depression studies (e.g., Foley et al., 2002; Jang, Kim, & Chiriboga, 2005; Miller, Markides, & Blank, 1997) identifying that the same depression measures given in different cultures suggested that people from different cultures express their depressive symptoms in different ways. For example, as mentioned earlier, Miller and his colleagues (1997) argued strongly for a two-factor model of the CES-D among elderly Mexican Americans instead of the classic four-factor model that was derived from samples of the general American population. They believed qualitative differences in factorial structure exist between the two cultural groups.

Type III posits that there is no difference in the nature of depression but that a difference in the instruments used to measure it is necessary. As mentioned already in the emic approach, studies that developed indigenous depression measures (e.g., the
Chinese Health Questionnaire, the American Indian Depression Scale, the Shona Symptom Questionnaire in Zimbabwe) are examples for this type III (e.g., Manson et al., 1985; Patel & Mann, 1997). This approach basically assumes that instruments for depression within a given culture must be emic derived from within the context of the culture rather than from outside it, and argues that when the same instruments are used across cultures, meanings for depression to be assigned to the scores would be different from one culture to another.

In Type IV, both the nature of depression and the instruments are different as a function of the culture being investigated. A number of culture-bound syndromes included in DSM-IV are examples: some of these culture-specific symptoms have been reported to be associated with depressive symptoms, such as brain fag (West Africa), dhat (Indian), nervios (Latinos), and shenjing shairu (Chinese). This type argues that depression can be understood and measured only as an emic construct within a given culture and nothing about depression is necessarily common across cultures.

It is my opinion that the two major topics of the present dissertation, “culture and depression”, can be more fully understood, measured, and analyzed by applying the proposed types of the relationship of culture to depression. Because research using the previously-mentioned emic-etic paradigm has not paid sufficient attention to differences in the instruments needed to measure depression across cultures, the proposed typology may expand the traditional emic-etic approach by adding the key aspect of depression instruments used in cross-cultural research to the proposed typology. The proposed analytic framework has the potential for providing a broader understanding on racial/ethnic differences in depressive symptoms, as well as on the role of culture on
depressive symptoms. Additionally, the proposed analytic typology may be useful in the
development of culturally sensitive and informed screening and assessment of depression
within a given culture. Finally, the proposed analytic typology may serve as a basis for
future studies designed to extend the topic to other mental health research areas.

*Measurement Equivalence in Cross-Cultural Depression Research*

One major issue related to assessing the cross-cultural comparability of depressive
symptoms has been the equivalence of measures (Bravo, 2003; Crockett et al., 2005;
Liang, 2002; van de Vijver, 2001). Although it may be less of a concern when more
open-ended questions are administered by an interviewer, the equivalence issue is
particularly serious when self-report screening measures are involved (Liang, 2002).
Valid and reliable questionnaire items in one language often lose meaning and context
after translation. Even with accurate translation, the problem of different nuances unique
to different cultures may not be resolved (Bravo, 2003). As mentioned already, even
when using the DSM criteria for depression and a standardized depression instrument,
cultural factors still affect the way individuals express their depressive symptoms.
Failure to substantiate the equivalence in depression instruments is potentially serious
because it may lead to inaccurate prevalence rate and misleading group comparisons
(Vandenberg & Lance, 2000). Therefore, in order to make meaningful comparisons
across diverse population groups, researchers must establish equivalence of depression
measures in addition to the traditional reliability and validity requirements of instruments.

Johnson (1998) reviewed articles on cross-cultural equivalence and reported that
more than 50 specific terms of equivalence (e.g., conceptual equivalence, criterion
equivalence, semantic equivalence, metric equivalence) have been discussed or
mentioned in the available literature on cross-cultural research. According to this review article, one of the most common equivalence addressed in the cross-cultural research was conceptual equivalence, followed by metric equivalence and structural equivalence (Johnson, 1998). Similarly, a number of researchers point out that cross-cultural measurement equivalence requires at least the abovementioned three interrelated conditions (Markides, Liang, & Jackson, 1990; van de Vijver, 2001). More importantly, these types of equivalence constitute a hierarchy in that both conceptual and metric equivalence are required for structural equivalence and metric equivalence assumes conceptual equivalence.

Establishing Conceptual Equivalence

Conceptual equivalence is the most basic type of equivalence and implies that research materials or observed behaviors have the same meaning in two or more cultures (Liang, 2002). Hui and Triandis (1985) identify conceptual equivalence as a necessary condition for making cross-cultural comparisons. Thus, unless two or more cultural groups share the same basic concept of depression, there is little purpose in determining whether measures of that construct are equally valid across groups (Crockett et al., 2005). An example of a possible research question for evaluating conceptual equivalence might be: “Do minorities and Whites think of depression in the same way?”

Conceptual equivalence can be evaluated by using several methods, such as back-translation, focus groups, random probes, and in-depth interviews. A major requisite for conceptual equivalence may be fidelity of translation. Ramirez and colleagues (2005) suggest that qualitative methods may be best suited to assess conceptual equivalence of existing depression measures as well as to uncover indigenous idioms of depression and
culture-bound syndromes. Qualitative methods, for example, can explore the relevance and appropriateness of depression concepts as well as the way people from different racial/ethnic backgrounds give meaning to a particular domain.

**Establishing Metric Equivalence**

Assuming conceptual equivalence, metric equivalence assures that a given measurement specification can be applied to different cultures (Liang, 2002). Metric equivalence occurs when the factor loadings of items in the depression instruments are invariant across the two or more cultural groups (Crockett et al., 2005). A number of studies evaluating the metric equivalence of depression instruments have used exploratory factor analysis (EFA; e.g., Callahan & Wolinsky, 1994; Foley et al., 2002), confirmatory factor analysis (CFA; e.g., Crockett et al., 2005; Nguyen et al., 2004), and item response theory (IRT: e.g., Cole et al., 2000; Iwata & Buka, 2002).

A number of studies have suggested that CFA may be the most versatile approach to evaluating metric and structural equivalence simultaneously (e.g., MacCallum & Austin, 2000; Raju, Laffitte, & Byrne, 2002). The assessment of measurement equivalence across cultural groups usually involves the use of multiple-group CFA (van de Vijver & Leung, 2000). The advantages of CFA with regard to cross-cultural research are that it 1) allows for detailed comparisons of factor models across cultural groups, 2) allows for a comparison of latent means, and 3) is not unduly influenced by relatively small sample sizes.

As an example of the confirmatory approach, on the basis of two surveys of African Americans and one survey of Caucasians, Nguyen and colleagues (2004) tested the metric equivalence of the CES-D scale. Using confirmatory factor analysis, they
demonstrated the equivalent number of factors and pattern of factor loadings across all three groups. Then, they tested the metric equivalence of the CES-D to see if the magnitudes of the factor loadings were equal across three racial/ethnic groups. Significant loading differences were found between African American and Caucasian groups, while significant loading similarities were found between the two surveys of African Americans. The authors pointed out that a higher number of loading similarities between two African American groups were expected given the same ethnicity and cultural backgrounds. Loading differences between African Americans and Caucasians were found in a number of both somatic complaints and depressive affect items, such as “I felt everything I did was an effort,” “My sleep was restless,” “I had crying spells,” and “I felt fearful.”

The study on elderly Mexican Americans by Miller and his colleagues (1997) is another example. They argued strongly for a two-factor model of the CES-D among elderly Mexican Americans instead of the original four-factor model (Radloff, 1977) which was derived from samples of the general American population. This study presents a case for qualitative differences in factorial structure between two cultural groups. Their findings raise a serious question as to whether the CES-D can be used for analyzing differences in depressive symptoms between elderly Mexican Americans and the general American population.

Establishing Structural Equivalence

Assuming both conceptual and metric equivalence, structural equivalence refers to similarities in the causal mechanism between a construct of depression and its consequences across different racial/ethnic groups (Liang, 2002). Considering that the
three types of equivalence constitute a hierarchy, depression instruments may be conceptually and metrically equivalent but not structurally equivalent across different racial/ethnic groups. Structural equation modeling (SEM; e.g., Crockett et al., 2005) and other techniques such as path analysis have been used to evaluate structural equivalence. A number of researchers point out that SEM is probably the most versatile approach to evaluating metric and structural equivalence simultaneously (Byrne & Watkins, 2003; Liang, 2002; MacCallum & Austin, 2000).

The study on Latino and Anglo adolescents of Crockett and colleagues (2005) is an example. The authors confirmed metric equivalence of a self-esteem measure across one Anglo and three Latino groups of adolescents, but not metric equivalence of the CES-D scale. They then tested structural equivalence of the CES-D between the four youth groups using multiple-group SEM. They identified similar relations between the CES-D and self-esteem across all four different groups (Anglo, Mexican, Cuban, and Puerto Rican) and concluded that results from the multiple-group SEM supported structural equivalence. However, it should be noted that this study is not a perfect example for structural equivalence because of the lack of metric equivalence of the CES-D scale across the Anglo and three Latino groups in this study. However, it is relatively rare to find studies on the structural equivalence of depression measures across different cultural groups.

Although a sizable number of studies have been conducted with regard to the cross-cultural comparison of depression, when two or more cultural groups are compared, descriptive and analytical techniques have been applied often without addressing issues concerning conceptual, metric, and structural equivalence (Liang, 2002). Given the
above review, assessing cross-cultural comparability of many widely used depression instruments seems to deserve the highest priority. Cross-cultural comparisons across racial/ethnic groups may not be justified without resolving these measurement concerns.

Cross-Cultural Research with the CES-D Scale

One of the major issues in cross-cultural depression research is how well depression instruments developed on samples of European Americans can assess depressive symptoms across racially or ethnically diverse groups. Because optimal depression screens and optimal cut-scores have not been identified for racially or ethnically diverse older adults, it may be important to review evidence for and against the utility of existing screening tools for depression across diverse cultural groups, as well as to identify optimal cut-scores for each existing depression screening tool. Given the fact that a substantial number of cross-cultural and cross-national studies on depressive symptoms have used the Center for Epidemiologic Studies-Depression Scale (CES-D) and the main purpose of this proposed dissertation study is to evaluate the measurement properties of the CES-D across diverse racial/ethnic groups, it may be meaningful to evaluate the cross-cultural applicability of the most widely used self-report depression instrument, the CES-D. The purpose of this section is (1) to provide a broad understanding on the CES-D scale and (2) to identify the cross-cultural applicability of the CES-D scale.

Description of the CES-D Scale

The CES-D was first developed in 1977, as a 20-item instrument with a dimensional response format (Radloff, 1977). The purpose of the CES-D is to determine the frequency and severity of current depressive symptoms in community samples.
Respondents are asked how often they have experienced each symptoms of depression during the past week on a four-point scale from “rarely or none of the time” to “most or all of the time.” Each item is scored 0-3, and total scores range from 0 to 60, with higher scores indicating more frequent depressive symptomatology. A suggested standard cutoff score for probable depression on the full-length questionnaire is equal to or higher than 16 (Andresen, Carter, Malmgren, & Patrick, 1994). CES-D scores of 16 to 26 are considered indicative of mild depression and scores of 27 or more are indicative of major depression (Zich, Attkisson, & Greenfield, 1990). Zich and colleagues (1990) found the stringent cutoff score of 27 was more useful for screening medical patients for depression than the standard cutoff score of 16.

Concern about respondent burden during surveys of older adults has led researchers to develop shortened versions of the original CES-D (Suthers, Gatz, & Fiske, 2004). To modify the original instrument, researchers have reduced the number of instrument items, as well as reduced the response format from four choices to either three or two choices. These shorter versions of the CES-D include the 11-item version administered in the Iowa Established Populations for Epidemiological Studies of the Elderly (EPESE), the 10-item version administered in the Boston EPESE, and an 8-item version administered in the Health and Retirement Study (HRS)/ Assets and Health Dynamics among the Oldest Old (AHEAD) survey of older adults (Suthers et al., 2004; Turvey, Wallace, & Herzog, 1999).

Factor analyses of the CES-D have been conducted since its initial development. In a recent meta-analysis of the factor structures of CES-D, Shafer (2006) identified that four specific factors have been generally supported by the majority of studies. These four
factors are generally described as Positive Affect, Depressed or Negative Affect, Somatic Symptoms, and Interpersonal Problems. However, there are no formal subscales for the CES-D. More importantly, a number of studies have not replicated four specific factors of the CES-D across diverse racial/ethnic groups (e.g., Crockett et al., 2005; Miller et al., 1997).

**DSM Criteria and the CES-D**

Numerous studies have examined the diagnostic accuracy of screening tests for depression. Several diagnostic instruments have been used to define the presence or absence of depression. The more frequently-used instruments include the following: Structured Clinical Interview for DSM-IV (SCID); Diagnostic Interview Schedule (DIS); CIDI; and Research Diagnostic Criteria (RDC). A sizable number of studies using self-report depression screening instruments have examined sensitivity and specificity for major depressive disorder, defined by a variety of criterion standards, many of which are based on DSM criteria.

According to the DSM-IV, the diagnosis of a major depressive episode requires the presence of five significant symptoms, which must include either a predominantly depressed mood and/or loss of interest or pleasure for at least a 2-week period of time (American Psychiatric Association, 1994; Kaplan & Sadock, 2003). Other symptoms may include: significant change in weight or appetite; insomnia or hypersomnia; psychomotor retardation or agitation; fatigue or loss of energy; feelings of worthlessness or excessive guilt; diminished concentration, loss of clarity of thought or indecisiveness; and recurrent thoughts of death (American Psychiatric Association, 1994; Kaplan & Sadock, 2003).
Previous research has compared CES-D scores to clinical depression based on DSM criteria (e.g., Haringsma, Engels, Beekman, & Spinhoven, 2004; Watson, Lewis, Kistler, Amick, & Boustan, 2004; Janssen, Beekman, Comijs, Deeg, & Heeren, 2006). For example, Beekman and colleagues (1997) compared the CES-D to major depression diagnoses determined by the Diagnostic Interview Schedule (DIS), which evaluates major depression based on the DSM-III. A cutoff score of 16 on the 20-item CES-D was found to have a sensitivity of 93.2%, specificity of 56.2%, and positive predictive value of 13.2% for major depression during the past year. Compared to false positives, true positives were more likely to be female and to have elevated anxiety, but did not differ on physical illness or cognitive performance.

As another example, focusing on what gerontologists call “the old-old” adults (i.e., age 75 to 84), Watson and colleagues (2004) compared the Geriatric Depression Scale (GDS) and the CES-D to the SCID-IV depression diagnoses. They reported that the recommended cutoff points of 12 on the GDS and 16 on the CES-D performed poorly in detecting both major and minor depression. Using a traditional cutoff score of 16, for example, the CES-D showed a sensitivity of only 60% and a specificity of 89% for detecting major depression and a sensitivity of only 50% and a specificity of 86% for detecting minor depression. Given the findings, the authors suggested the need for more sensitive methods of screening in the healthy older adults.

Some studies have compared CES-D short forms to DSM-based instruments (e.g., Irwin, Artin, & Oxman, 1999; Suthers et al., 2004; Turvey et al., 1999). Irwin and colleagues (1999) compared their screening results from the CES-D against to the SCID as a criterion standard in a community-based sample of adults with known physical
illness. They found that scores of greater than or equal to 4 had 99% sensitivity and 84% specificity for major depression. As another example, on the basis of the AHEAD data Turvey and colleagues (1999) compared a yes/no format 8-item version of the CES-D to a DSM-based measure, the CIDI-SF. Results indicated that the clearest source of discordance between the two measures was found for those who were positive for depression on the 8-item CES-D but not currently depressed on the CIDI-SF. The authors reported that the discordance was explained by respondents who reported depressive symptoms on the 8-item CES-D, but did not meet the CIDI-SF criteria for symptom frequency or duration, and were therefore classified as noncases.

*Cross-Cultural Applicability of the CES-D*

The CES-D has been used in a sizable number of cross-cultural and cross-national studies on depressive symptoms. A number of previous studies have compared prevalence rates and means of the CES-D across racial/ethnic groups and found evidence of differences in both prevalence rates and means across those groups (e.g., Foley et al., 2002; Krause & Liang, 1992; Mackinnon, McCallum, Andrews, & Anderson, 1998). Comparing group means of the CES-D across four racial/ethnic groups (Japanese, Taiwanese, African Americans and Whites in the U. S.), Krause and Liang (1992) found that Japanese elders showed the lowest mean scores on overall depressive symptoms, followed by Taiwanese, Whites, and African Americans. More recently, Inoba and colleagues (2005) found that Japanese tend to have lower mean scores of the CES-D than Whites. In addition to the reported differences in the mean levels on the CES-D across cultural groups, the probable caseness rates of the CES-D also varied dramatically across diverse racial/ethnic groups, ranging from 3.5% to more than 30%. For example, Blazer
and colleagues (1998) showed that 9.5% for African Americans and 8.8% for Whites fall under the category of probable depression. Using the cut-off scores of the CES-D, reported prevalence rates across and within racial/ethnic groups are the following: 3.5% for Germans (Papassotiropoulos & Heun, 1999); 13.2% for Hispanics and 9.2% for Whites (Swenson et al., 2000); 14% of African Americans (Foley et al., 2002); 19.8% for African Americans (Baker, Velli, Freidman, & Wiley, 1995); 25.3% for Koreans (Cho, Nam, & Suh, 1998); 25.4% for Mexican Americans (Gonzalez et al., 2001); and more than 30% for Korean Americans (Jang et al., 2005).

Given the evidence for different means and rates of probable depression across racial/ethnic groups, a major issue with regard to cross-cultural applicability of the CES-D instrument is the extent to which such racial/ethnic group comparisons reflect true differences in the depressive symptoms or conversely, how much is due to measurement variance in the construct of interest. Some studies have led to results suggesting that the CES-D should undergo revision based on the target population. For example, Liang and colleagues (1989) observed lack of consensus on the factor structure of the CES-D and developed a 12-item version of the CES-D in Mexican Americans. In a study of Black and White Americans, Callahan and Wolinsky (1994) found significantly different factor structures for the four race/gender groups, and recommended dropping five items to maximize comparability across race by gender groups. Chapleski and colleagues (1997) also found that a 12-item version of the CES-D was a superior fit to American Indians than the original 20-item scales. Although the CES-D appeared robust in the face of these minor changes, McCallum and colleagues (1995) suggested these minor changes to
the original CES-D scale may create some potential risks, such as reduced reliability and lack of comparability in norms for screening.

Several studies have found that the CES-D has acceptable internal consistency as well as Radloff’s (1977) four-factor solution of depressive symptoms in different racial/ethnic groups (e.g., Blazer et al., 1998; Krause & Liang, 1993; Roberts, 1980). Two studies of Roberts and colleagues (Roberts, 1980; Roberts, Vernon, & Rhoades, 1989), for example, showed acceptable reliability of the CES-D in Mexican-Americans, African Americans, and Anglo Americans. Using confirmatory factor analysis, they (1989) found that a four-factor structure was supported in Anglo-American and Mexican American psychiatric patients. On the basis of the data from the Duke site of the EPESE, Blazer and colleagues (1998) also confirmed an original four-factor solution of the CES-D in both African Americans and Whites. Additionally, Mui and her colleagues (2002) reviewed cross-cultural studies on the CES-D and pointed out its usefulness for assessing depressive symptoms across cultures in older adults.

Although several studies have replicated the original four-factor solution, there is growing evidence for the existence of unique measurement properties of the CES-D across racial/ethnic groups. Factor analyses of the CES-D have yielded a wide range of factors across diverse populations, ranging from two to six factors (e.g., Crockett et al., 2005; Liang et al., 1989; Miller et al., 1997; Posner, Stewart, Marín, & Pérez-Stable, 2001). For example, on the basis of the Hispanic EPESE data, Miller and his colleagues (1997) argued strongly for a two-factor model of the CES-D among elderly Mexican Americans instead of a four-factor model which was derived from samples of the general American population. Chapleski and colleagues (1997) also found the factor solution of
the CES-D was inconsistent with the original four-factor solution when applied to American Indians. More recently, research on adolescents by Crockett and colleagues (2005) found evidence for differences in the CES-D factor structure even among Latino subgroups (Mexican Americans, Cuban Americans, and Puerto Rican Americans). This study showed that the original four-factor solution fit very well for Whites and Mexican Americans while the four-factor solution did not fit adequately for Cuban Americans and Puerto Rican Americans. Given the identified different factor structure for Cuban Americans and Puerto Rican Americans, the authors argued that Cuban and Puerto Rican Americans have somewhat different concepts of depression than Whites.

Several recent studies have addressed different response patterns to items of the CES-D across racial/ethnic groups (e.g., Cole et al., 2000; Iwata & Buka, 2002; Jang et al., 2005; Kim, Jang, Chiriboga, 2006). Cole and colleagues (2000) examined race/ethnicity–specific response tendencies on the CES-D utilizing differential item functioning (DIF) analysis. They found that African American elders were more likely to endorse items related to interpersonal problems, compared to White elders. Focusing on racially or ethnically diverse young adults, Iwata and colleagues (Iwata & Buka, 2002; Iwata et al., 2002) also used DIF analyses to assess variations in the manifestation. They showed that items related to somatic symptoms and positive affect function differently across racial/ethnic groups in both studies. Both African Americans and Native Americans tended to endorse somatic symptoms over depressive symptoms, and immigrant Hispanics and Japanese appeared to inhibit the expression of positive affect.

More recently, with special attention to the role of levels of acculturation on the CES-D, Jang and colleagues (2005) found that Korean American elders who are lower or
higher on acculturation showed significant differences in response patterns and factor structures of the CES-D. Additionally, DIF analysis suggested that those with lower levels of acculturation were more likely to inhibit responses to positive items of the CES-D. Expanding upon the findings from this study, Kim and colleagues (2006) examined variations in response to the CES-D by age and the levels of acculturation, with Koreans in Korea and Korean Americans. Utilizing multiple group confirmatory factor analyses (CFA), metric differences in the CES-D were identified across four cultural groups (low acculturated Korean Americans; high acculturated Korean Americans; Korean middle-aged adults; Korean older adults), and similar parameter estimates were found between low acculturated Korean Americans and Korean older adults. This line of research suggests the substantial cultural influences in expressing depressive symptoms.

Several conclusions can be drawn with regard to the applicability of each instrument and methodological concerns, although variations in samples and analytic techniques used in the previous research, as well as mixed results from the literature make it hard to compare findings. First, cross-cultural and cross-national studies on the CES-D confirm its general usefulness for assessing depression in diverse groups of older adults. Second, reported findings from the previous research suggest that cultural factors may impact the reporting of depressive symptoms. Third, changes in various items and factors may improve the metric and structural equivalence of the CES-D when the CES-D instrument is applied to culturally diverse groups. Fourth, more research on the comparative performance of various screening instruments with diverse racial/ethnic groups may be needed. Fifth, while various versions of the CES-D instrument demonstrate excellent reliability, it may need to be further calibrated against clinical
assessments of depression in diverse racial/ethnic groups. Finally, structural equivalence across cultural groups also needs to be tested and established, by relating the CES-D to other correlates of depressive symptoms.

**Measurement Equivalence of the CES-D Scale**

As mentioned above, although cross-cultural and cross-national studies on the CES-D scale confirm its general usefulness for assessing depression in diverse groups, previous studies have suggested group differences in the CES-D items across diverse racial/ethnic groups (e.g., Cole et al., 2000; Perreira et al., 2005; Roberts, Rhoades, & Vernon, 1990). Reported differences in the CES-D items were often found across subgroups of race/ethnicity (e.g., Cole et al., 2000; Iwata & Buka, 2002), age (e.g., Gatz & Hurwicz, 1990; Hays, Landerman, George, Flint, Koening, Land et al., 1998; Kessler, Foster, Webster, & House, 1992), gender (e.g., Posner et al., 2001; Stommel, Given, Given, Kalaian, Schulz, & McCorkle, 1993), acculturation (e.g., Chiriboga et al., 2007; Jang et al., 2005; Kim et al., 2006), and instrument language (e.g., Roberts et al., 1990), and other important variables. These group differences in the CES-D items can be observed because (1) the CES-D measure accurately assesses the underlying attribute for different groups when they actually differ on that attribute (true difference) or (2) the CES-D scale inaccurately assesses the underlying attribute for one or more groups when the groups do or do not truly differ on the attribute (artifactual difference). However, previous research has not fully identified whether such observed group differences in the CES-D items represent true differences, are due to differential item functioning (DIF), or are a combination of the two.
One major area found lacking in the literature is the assessment of differential item functioning (DIF) for the CES-D across subgroups. As mentioned above briefly, an item functions differentially if people from different subgroups with equal depression scores on the CES-D do not have the same probability of item endorsement. DIF occurs when the probability of responding to an item depends on both an individual’s depression and other construct-irrelevant factors, such as race/ethnicity, age, or gender (Teresi, 2002). DIF can be a threat to validity since measures containing DIF items may be invalid for between-group comparisons because their scores are indicative of attributes other than that which the test is intended to measure (Perkins, Stump, Monahan, & McHorney, 2006). Including DIF items in the CES-D scale can exaggerate or attenuate true differences between subgroups of race/ethnicity, age, gender, or other important demographic variables.

The effect of race/ethnicity on the measurement properties of the CES-D has not been fully identified yet in the previous research and still remains questionable. Most previous studies of measurement bias in the CES-D scale have focused on subscale instead of individual item analysis to trace differential responses across racial/ethnic groups (Blazer et al., 1998; Callahan & Wolinsky, 1994; Nguyen et al., 2004; Perreira et al., 2005; Roberts et al., 1989). For example, Callahan and Wolinsky (1994) found differences in the CES-D factor structure by racial/ethnic group, although the study was limited due to missing data. More importantly, even some previous investigations of race/ethnicity bias in the CES-D scale did not include older populations for their analyses (e.g., Iwata & Buka, 2002; Iwata et al., 2002). Only a single study has used an appropriate method to assess item bias in the CES-D scale by race/ethnicity and tested the
effect of race/ethnicity on the CES-D items among older adults. Cole and colleagues (2000) compared two racial/ethnic groups and found evidence for racial/ethnic item bias in the CES-D scale. The authors suggested that Blacks were more likely to endorse higher levels of the two interpersonal problem items (“people are unfriendly” and “people disliked me”). To date, none of studies fully considered and tested DIF items in the CES-D scale across three or more racial/ethnic elderly groups. More investigation will be needed regarding race/ethnicity-specific DIF items in the CES-D scale, and future research on item bias in the CES-D needs to include Hispanic population. Additionally, for future research, it may be meaningful to explain why diverse racial/ethnic groups respond differentially to the items of the CES-D and what factors across diverse racial/ethnic groups can contribute to different response patterns to the CES-D items.

The effect of demographic characteristics (e.g., age, gender, educational attainment) on the measurement properties of the CES-D in diverse racial/ethnic groups has not been identified yet and still remains uncertain. Previous studies found different measurement properties of the CES-D scale across subgroups of age (e.g., Gatz & Hurwicz, 1990) and gender (e.g., Berkman, Berkman, Kasl, Freeman, Leo, Ostfeld et al., 1986; Callahan & Wolinsky, 1994). However, most studies on the measurement properties of the CES-D were limited to test the factor structure instead of testing item bias in the CES-D scale. Moreover, although some of previous studies have used appropriate methods to test item bias in the CES-D across demographic subgroups (e.g., Cole et al., 2000; Stommel et al., 1993), these studies did not consider the effect of demographic variables on the CEC-D items within each racial/ethnic group. For example, although Cole and colleagues (2000) conducted item-level biases in the CES-D across
subgroups of age, gender, and race/ethnicity, the authors did not analyze age- and gender-related measurement bias in the CES-D within each racial/ethnic group. Because each racial/ethnic group has its unique demographic profiles which may underlie racial/ethnic group differences (Alwin & Wray, 2005) and these cultural and social differences across diverse racial/ethnic populations may result in different responses to depressive symptom items (Nguyen et al., 2004), it may be meaningful to examine demographic-related item bias in the CES-D scale within each racial/ethnic group. Special lacking in the previous research was the effect of educational attainment: none of the previous research focused on the effect of educational attainment on the measurement properties of the CES-D among older adults.

Recent research suggests that acculturation and instrument language significantly affect different response patterns to items of the CES-D (Chiriboga, Jang, Banks, & Kim, 2007; Jang et al., 2005). Given that acculturation has been referred to simply as the degree to which people change when faced with the challenge of living in a cultural context differing from their own (Berry & Kim, 1988; Trimble, 2003), people who are more acculturated may be more likely to adopt the ways of thinking and expressing that characterize the host culture. In contrast, those who are less acculturated may be less accepting the new ways of thinking and expressing themselves, and instead hold onto the culture and behaviors that reflect their culture of origin. The differences in thinking, feeling, and expressing themselves may influence the ways in which depressive symptoms are organized and expressed, as well as have implications for the measurement nonequivalence of screening tools for depressive symptoms. In addition, literature suggests that language significantly affects the measurement properties of the CES-D in
racial/ethnic populations. Feelings reported in a native language may be expressed with
more emotion than those expressed in a second language (Cuellar & Roberts, 1984; Roberts et al., 1990). Given that previous studies showed higher correlations between
English proficiency and acculturation, it may be interesting to find out the effect of both
acculturation and instrument language on the measurement properties of the CES-D in
diverse racial/ethnic groups. Findings from this line of research may suggest the
substantial cultural differences even within the same racial/ethnic group.

*Analytic Strategies for Measurement Equivalence Research*

Researchers have used several analytic strategies and statistical methods to
evaluate racial/ethnic group differences in depressive symptoms itself. For example,
studies showing differences in prevalence rates of depression across racial/ethnic groups
have used chi-square test, providing sensitivity and specificity (e.g., Baker et al., 1995; Beals, Manson, Whitesell, Mitchell, DKovins, Simpson et al., 2005; Beekman et al., 1997; Madianos, Gournas, & Stefanis, 1992; Medina-Mora et al., 2005). Some
researchers have used receiver operating characteristic (ROC) analysis to provide
different cut-off points across racial/ethnic groups (e.g., Papassotiropoulos & Heun, 1999; Somervell, Beals, Kinzie, Boehnlein, Leung, & Manson, 1993). There have been
some studies using ANOVA or T-test to compare means of depressive symptoms across
cultural groups (e.g., Krause & Liang, 1992; Lee & Farran, 2004).

A number of studies evaluating factor structure of depression instruments have
used exploratory factor analysis (EFA; e.g., Callahan & Wolinsky, 1994; Foley et al., 2002; Van Tran, 1997). EFA is useful in a preliminary stage in identify the number of
factors and the underlying pattern of factor loadings within a given cultural group.
Although EFA has been used to assess metric equivalence, specifically, factorial invariance, EFA is insufficient for assessing comparability for several reasons. Liang (2002) suggested the following three reasons. First, in EFA, significance tests of the differences in factorial structure cannot be applied. Second, comparisons of factor configurations are accomplished by using Pearson correlation matrices. Such comparisons are confounded by variance differences between samples because correlation coefficients are standardized in terms of sample variances. Third, in EFA, little a priori specification is involved in deriving the factor structure. A number of other researchers have also noted the limitations of EFA (e.g., Teresi, 2002; van de Vijver & Leung, 2000).

Measurement equivalence is typically assessed through the use of statistical analyses comparing the properties of an instrument in two or more groups. In order to distinguish a lack of measurement equivalence (i.e., differential item functioning; DIF), a problem with the instrument, from impact, true differences in the trait distributions, researchers have used two popular methods capable of detecting DIF; one is confirmatory factor analysis (CFA) and the other is item response theory (IRT). A number of studies have suggested that CFA may be the most versatile approach to evaluating metric and structural equivalence simultaneously (e.g., MacCallum & Austin, 2000; Raju et al., 2002). The assessment of measurement equivalence across cultural groups usually involves the use of multiple-group CFA (van de Vijver & Leung, 2000). The advantages of CFA with regard to cross-cultural research are that it 1) allows for detailed comparisons of factor models across cultural groups, 2) allows for a comparison of latent means, and 3) is not unduly influenced by relatively small sample sizes.
There are advantages ascribed to IRT, which has also been reported as a powerful method for examining measurement equivalence at the item or scale level (Bingenheimer, Raudenbush, Leventhal, & Brooks-Gunn, 2005). Assuming that all items reflect the same underlying construct, IRT allows for the comparison of scale scores even when individuals or groups of individuals answer different items. In addition, IRT parameters are subpopulation invariant. They do not depend on the distribution of trait scores in examinee groups, in contrast to more conventional analyses where item means depend on sample particulars (e.g., the same item can have a low mean in one group and a high mean in another) (Hulin, Drasgow, & Parsons, 1983; Meade & Lautenschlager, 2004; van de Vijver & Leung, 2000).

There have been a number of studies recently that have compared CFA and IRT DIF detection procedures (e.g., Meade & Lautenschlager, 2004; Raju et al., 2002). These previous studies compared traditional CFA-based procedures with IRT-based procedures that differed with regard to hypothesis testing strategies and targets of analysis. Recently, however, Stark, Chernyshenko, and Drasgow (2006) proposed and tested a common strategy for detecting DIF with CFA and IRT based on the likelihood ratio (LR) test. Their method, which involved comparing statistically correct free-baseline models with a series of constrained models that examined item loadings/discrimination parameters and intercepts/location parameters simultaneously, showed higher power and low Type I error rates across simulation conditions for both CFA and IRT.

Issues Needed for Future Research

Given the preceding review of conceptual issues in cross-cultural depression research and cross-cultural research with the CES-D scale, some conceptual and
methodological concerns and problems can be identified and these issues may be in need
for future research in the field of depression. First, with regard to dataset issues, the
majority of studies with the CES-D scale have used data collected from one racial/ethnic
or cultural group to test measurement equivalence of the CES-D. To assess cultural
group differences in the CES-D scale, as well as to make a meaningful comparison of
depression with the CES-D scale across diverse racial/ethnic groups, collecting
comparable data may deserve the highest priority. In addition, future research should pay
attention to find nationally representative datasets with inclusions of the original version
of the CES-D scale and racially/ethnically diverse elderly groups.

Second, when two or more racial/ethnic groups are compared, descriptive and
analytical techniques have been applied often without addressing issues concerning
conceptual, metric, and structural equivalence. Given that cross-cultural measurement
equivalence requires at least three interrelated conditions which are conceptual, metric,
and structural equivalence (Markides et al., 1990; van de Vijver, 2001), equivalence of
the instruments must be established to make a meaningful comparison across diverse
racial/ethnic groups. Currently, much more remains to be learned about the cross-cultural
comparability of the CES-D scale.

Third, although previous research raised the possibility of the nonequivalence of
the CES-D across different racial/ethnic groups, cross-cultural depression studies have
not paid enough attention to how the CES-D measure can be improved to screen for
depression. The consequences of using nonequivalent instruments may be potentially
serious for several reasons. For example, if the CES-D commonly used to screen for
cases of depression is valid and accurate for one cultural group but less for another
cultural group, applying the standard cut-scores of the CES-D may lead to misclassification in the second group, resulting in false positives, false negatives, or both. Thus, researchers should pay more attention to making the CES-D more reliable and culturally valid, as well as to establishing measurement equivalence of the CES-D.

Fourth, and perhaps most important, the effects of race/ethnicity and demographic characteristics on the measurement properties of the CES-D have not been identified and still remain questionable. Given that some of racial/ethnic differences in mental health may be explained by differences in the age, gender and socio-economic composition of each racial/ethnic group (U.S. Department of Health and Human Services, 2001), differences in the demographic characteristics (age, gender, and educational attainment) of each racial/ethnic group may also explain different responses to the CES-D scale, which may in turn contribute to detecting DIF items that function differently across subgroups of each racial/ethnic group. By identifying demographic-related item bias in the CES-D within each racial/ethnic group, future research can explain why diverse racial/ethnic groups respond differentially to the certain items of the CES-D scale. Currently, much more remains to be learned about demographic-specific depressive symptom items that function differentially across cultures and sub-cultural groups within each racial/ethnic group, as well as a core set of depressive symptom items that function equivalently across cultures and subgroups of age, gender, and educational attainment.

Fifth, acculturation-related item bias in the measurement properties of the CES-D has not been extensively identified among diverse racial/ethnic groups. Because literature suggests that level of acculturation is an important variable to explain differences in thinking and expressing depressive symptoms within racial/ethnic minority
group (Chiriboga, 2004; Chiriboga et al., 2007; Myers & Rodriguez, 2003) and even explained racial/ethnic differences in depression by differences in age, gender, and SES of each group may not be meaningful without controlling for acculturation levels within racial/ethnic groups (Myers & Rodriguez, 2003), it may be expected to find the unique role of acculturation on item biases in the CES-D, as well as other depressive symptom inventories. In addition, given that acculturation levels and language proficiency are highly interrelated and instrument language itself significantly affects measurement bias in racial/ethnic populations, it will be meaningful to examine the combination role of acculturation and instrument language. Findings from this line of research may suggest the substantial cultural differences even within the same racial/ethnic group.

Lastly, although at present, CFA and IRT are two popular methods to detect a lack of measurement equivalence, none of published studies involving item level comparisons in depressive symptom instruments compared CFA and IRT results. Considering that only some simulation studies are now beginning to compare CFA and IRT DIF detection procedures and appear in the research literature (e.g., Gonzalez-Roma, Hernandez, & Gomez-Benito, 2006; Stark et al., 2006), testing DIF analyses with both CFA and IRT may be clearly needed for future study in cross-cultural depression research. Moreover, it may be very meaningful to use a common strategy for DIF detection across both CFA and IRT and compare the results. The combined use of the two statistical approaches may increase the accuracy for testing metric equivalence of depressive symptom inventories as well as other mental health instruments and may be the first step towards applying the integrative method of detecting DIF across diverse groups.
Study Design

This dissertation was designed to examine cultural differences in the measurement properties of the CES-D among racially/ethnically diverse older adults within the framework of integrating Sternberg’s Type I (the nature of depression is the same across cultures and the instruments used to measure it are the same) and Type II (the nature of depression is different across cultures, but the instruments used to measure it are the same across cultures) regarding the relationship of culture to depression. As mentioned earlier, three major hypotheses for the dissertation were (1) the measurement properties of the CES-D would vary by race/ethnicity among older adults, (2) the measurement properties of the CES-D scale would vary by age, gender, and educational attainment group within each racial/ethnic elderly group, and (3) the measurement properties of the CES-D would vary by the level of acculturation and instrument language in older Mexican Americans. Three steps for the dissertation were designed to test the abovementioned three hypotheses in a separate study and the outline for the three studies was shown in Figure 2.

Study 1 was designed to examine Research Question 1. The purpose of this study was to test whether the CES-D items function differentially across three racial/ethnic group, older Whites, Blacks, and Mexican Americans. Study 1 was expected to identify race/ethnic-specific items in the CES-D that function differentially across three racial/ethnic groups, as well as a core set of depressive symptom items in the CES-D that function similarly across three racial/ethnic groups. Findings from Study 1 were expected to address limitations of the CES-D as a tool for making cross-racial/ethnic comparisons.
Study 2 was designed to examine *Research Question 2*. This study aimed to assess the role of sociodemographic characteristics (i.e., age, gender, and educational attainment) on the measurement properties of the CES-D within each racial/ethnic elderly group (Whites, Blacks, and Hispanics). Study 2 hypothesizes that the measurement properties of the CES-D scale will vary by age, gender, and educational attainment group within each racial/ethnic elderly group. For the purpose of this Step 2, differentiation was made with respect to age (younger than 75 vs. 75 or older), gender (male vs. female), and educational attainment (Less than 8\textsuperscript{th} grade vs. 8\textsuperscript{th} grade or more) variables for each racial/ethnic group. Results from Study 2 was expected to identify age-, gender-, and educational attainment-specific items in the CES-D within each racial/ethnic elderly group, as well as a core set of items in the CES-D functioning equivalently across subgroups of age, gender, and educational attainment within each racial/ethnic elderly group. Findings from Study 2 may provide a broader understanding of cultural group differences in the CES-D among racially/ethnically diverse older adults and be a basis for developing culturally equivalent depression measures across diverse racial/ethnic groups and subgroups within each racial/ethnic group.

Study 3 was designed to examine *Research Question 3*. The purpose of Study 3 was to identify the measurement properties of the CES-D across acculturation levels and instrument language among older Mexican Americans. In this step, this dissertation was expected to identify acculturation and instrument language-specific items in the CES-D that function differentially in older Mexican Americans, as well as a core set of CES-D items that function equivalently across subgroups within the Mexican American group. With special attention to the role of acculturation and instrument language on the
measurement properties of the CES-D, differentiation was made with respect to the level of acculturation (low vs. high acculturated groups) and instrument language (Spanish vs. English) for older Mexican Americans. Findings from Study 3 may provide evidence of measurement nonequivalence of the CES-D scale even within a specific subgroup of the overall Hispanic population.
Study 1:
Research Question 1:
Do the measurement properties of the CES-D vary by race/ethnicity?

New Haven EPESE (N = 2,340)
Hispanic EPESE (N = 2,623)

Whites (N = 1,876)
Blacks (N = 464)
Mexican Americans (N = 2,623)

1. Age: 1) <75
   2) 75 or older
2. Gender: 1) Male
   2) Female
3. Education: 1) <8th grade
   2) 8th or more

Study 2:
Research Question 2:
Do the measurement properties of the CES-D vary by sociodemographic characteristics within each racial/ethnic group?

1. Age: 1) <75
   2) 75 or older
2. Gender: 1) Male
   2) Female
3. Education: 1) <8th grade
   2) 8th or more

Research Question 3:
Do the measurement properties of the CES-D vary by acculturation and instrument language in older Mexican Americans?

1. Acculturation: 1) Low Acculturation
   2) High Acculturation
2. Instrument Language: 1) Spanish
   2) English
3. Acculturation x Instrument Language: 1) Low Acculturation - Spanish
   - English
   2) High Acculturation - Spanish
   - English

Figure 2. Outline for the Dissertation
CHAPTER TWO: STUDY 1

RACE/ETHNICITY-RELATED MEASUREMENT BIAS

Introduction

A growing body of literature documents racial/ethnic disparities in depressive symptoms (e.g., Coyne & Marcus, 2006; Dunlop, Song, Lyons, Manheim, & Chang, 2003). A number of cross-cultural and cross-national studies on these depressive symptoms have found evidence that prevalence rates of probable depression vary dramatically across diverse racial/ethnic groups, with the rates ranging from 1.5% to 32.0% (e.g., Blazer, Landerman, Hays, Simonsick, & Saunders, 1998; Gonzalez, Haan, & Hinton, 2001; Dunlop et al., 2003; Foley, Reed, Mutran, & DeVellis, 2002; Mui, Burnette, & Chen, 2002). For example, Gonzalez and colleagues (2001) reported that 25.4% of Mexican American elders showed evidence of probable depression, while Foley and colleagues (2002) showed 14.0% of African American elders fall under the category for probable depression. Because of cultural differences across diverse racial/ethnic groups, in fact, it has become a virtual truism in cross-cultural research that racially/ethnically diverse groups manifest different prevalence rates of probable depression and different group means on standard inventories. However, the question of whether differences across diverse groups have any practical implications for how the currently existing depression instruments should be used remains in controversy.

The relationship of culture to depression has been explained by two different perspectives, the emic and the etic (Canino, Lewis-Fernandez, & Bravo, 1997; Pike,
The emic approach to cross-cultural depression research assumes that every culture has its own unique ways of manifesting and expressing symptoms and suggests using variables and observations that are tailored to the particular cultural group when developing a depression instrument (Rait & Burns, 1998). This approach argues that when the same instruments are used across cultures, meanings of the depression scores may differ from one culture to another. Following this approach has led to the development of several indigenous depression instruments, such as American Indian Depression Scale (AIDS; Manson et al., 1985) and the Shona Symptom Questionnaire (Patel, Simunyu, Gwanzura, Lewis, & Mann, 1997) in Zimbabwe.

In contrast, etic researchers assume that the nature of depression is essentially the same across cultures and that, using appropriate translations when necessary, this nature can be measured identically without regard to culture (e.g., World Health Organization, 1983). Etic studies of depression have led to the development of a number of depression instruments including self-report scales (e.g., Beck Depression Scale, Center for Epidemiologic Studies Depression Scale, Geriatric Depression Scale) and interviewer or clinician rating scales (e.g., Structured Clinical Interview for DSM-IV, WHO Composite International Diagnostic Interview). This line of research assumes that core depressive symptoms exist cross-culturally and cross-nationally and that the key issue is variation in levels of depressive symptoms across cultural groups.

Although emic and etic approaches have their own views on the relationship of culture to depressive symptoms, where these perspectives basically agree is that culture shapes values, attitudes, beliefs, and behaviors (Sternberg, 2004) and may affect ways in which people experience and express their symptoms of depression (Kleinman, 2004).
However, it is unfortunate that neither approach has been able to provide clear guidelines as to how to analyze cultural influences on existing depressive symptom inventories.

One unresolved issue in assessing depressive symptoms of different cultural groups has been the equivalence of measures (e.g., Crockett, Randall, Shen, Russel, & Driscoll, 2005; Liang, 2002). Measurement equivalence is of particular importance in cross-cultural depression research because if depressive symptom measures have differential meanings or validity across diverse cultural groups, group comparisons would be misleading and the prevalence rates of depression would be inaccurate (Crockett et al., 2005). When self-report depression measures are used, researchers should pay particular attention to the issue of measurement equivalence since it may be unclear whether different depressive symptom scores across cultural groups are caused by actual differences in depression or problems with the depression instrument used (i.e., differential item functioning; DIF). If people from different cultures with equal depression scores are more or less likely to endorse specific depressive symptom items as a consequence of cultural grouping, those specific items endorsed more or less by a group of people will function differently across groups, or show DIF (Teresi, 2002). Measures containing DIF items may be invalid for between-group comparisons because their scores are indicative of attributes other than that which the test is intended to measure (Perkins, Stump, Monahan, & McHorney, 2006). Therefore, a high priority in assessing cross-racial/ethnic comparability of many widely used depression instruments should be to distinguish a lack of measurement equivalence, which is known as DIF, from impact, true group differences in standing on the same latent trait, as well as to find appropriate methodology that can detect the measurement bias accurately.
Given the abovementioned importance of detecting DIF in cross-cultural studies, researchers have generally used one of two methods capable of detecting DIF; confirmatory factor analysis (CFA) or item response theory (IRT). Most studies examining DIF in depression instruments have only used one of these methods. The methods, however, may not yield the same results, since they differ in terms of their mathematical models (e.g., linear model vs. nonlinear models), target of analysis (e.g., testing invariance of loading parameter first and intercepts later vs. testing invariance of discrimination and location parameters simultaneously), and strategies for hypothesis testing (testing free baseline model vs. testing constrained baseline model) (Raju, Laffitte, & Byrne, 2002; Stark, Chernyshenko, & Drasgow, 2006). The potential discrepancy in results raises serious questions as to how to interpret results. At present studies involving item level comparisons in depressive symptom instruments have not included both CFA and IRT results, and only now are some simulation studies beginning to compare these two DIF detection procedures (e.g., Meade & Lautenschlager, 2004; Raju et al, 2002; Stark et al., 2006). It is therefore meaningful to employ both CFA and IRT methods. One of the objectives of the present study was to use a common strategy across both CFA and IRT as a means of detecting DIF in the widely-used Center for Epidemiological Studies Depression Scale (CES-D).

Since its development on samples of European Americans, the CES-D has been used in a substantial number of cross-cultural studies on depressive symptoms. Several studies on the CES-D have found acceptable internal consistency (Mui et al., 2002) and have confirmed its general efficacy in detecting depression across diverse racial/ethnic groups of older adults, including African Americans (e.g., Blazer et al., 1998) and
Mexican Americans (e.g., Gonzalez et al., 2001; Swenson, Baxter, Shetterly, Scarbro, & Hamman, 2000). Despite its general usefulness in cross-cultural application, there has been growing evidence that the CES-D has unique measurement properties across racial/ethnic groups (e.g., Callahan & Wolinsky, 1994; Crockett et al., 2005; Foley et al., 2002; Miller, Markides, & Blacks, 1997; Nguyen, Kitner-Triolo, Evans, & Zonderman, 2004; Perreira, Deeb-Sossa, Harris, & Bollen, 2005; Roberts, Vernon, & Rhoades, 1989). For example, Miller and colleagues (1997) argued strongly for a two-factor model of the CES-D among elderly Mexican Americans instead of the classic four-factor model that was derived from samples of the general American population. Foley and colleagues (2002) also found evidence for differences in the CES-D factor structures among older African Americans: they found no distinction between depressive affect and somatic symptom factors. As mentioned above, however, most previous studies investigating unique measurement properties of the CES-D scale in diverse cultural groups have focused on subscale analyses of the CES-D. Given that cultural differences across diverse racial/ethnic groups may be related to unique patterns of response (e.g., McHorney & Fleishman, 2006), these subscale analyses may not be enough to understand and determine unique psychometric properties of the CES-D (Teresi, 2002). The potential of the CES-D items to function differentially across multi racial/ethnic elderly groups thus becomes a priority in cross-cultural research.

A few studies have used DIF methods to investigate the CES-D item bias by race/ethnicity and found evidence of differential function (e.g., Cole, Kawachi, Maller, & Berkman, 2000; Iwata & Buka, 2002; Iwata, Turner, & Lloyd, 2002). In one DIF analysis conducted among White, Japanese, Native American and Argentinean
undergraduates, Iwata and Buka (2002) found evidence that Whites are predisposed to endorse positive CES-D items and that Japanese and Argentineans are more likely to inhibit endorsement of the same positive items. Generalizations from this study were limited due to the use of a small and non-representative sample. Testing DIF among representative samples of younger adults, Iwata and colleagues (2002) suggested that two of four positive items (‘I felt hopeful about future’ and ‘I enjoyed life’) showed DIF and were over-endorsed by African Americans, and Hispanics showed tendencies to inhibit the expression of positive affect. In the field of gerontology, only two studies (Cole et al., 2000; Yang & Jones, in press) investigated race/ethnic item differences on the CES-D. Using the same New Haven EPESE data set employed in the present investigation, both studies found evidence that Blacks were more likely than Whites to endorse two interpersonal relation items (‘people are unfriendly’ and ‘people disliked me’). To date, no study has fully considered and tested DIF items in the CES-D across three or more racial/ethnic elderly groups. Moreover, none have included older Mexican Americans, the largest subgroup of Hispanics in the United States, for identifying DIF items on the CES-D. For this reason, more investigation is clearly needed regarding race/ethnicity-specific DIF items in the CES-D scale among diverse groups of older adults including Mexican Americans.

Taking the abovementioned issues together into account, the purpose of this present study was to examine the cultural equivalence of the CES-D items across three racial/ethnic elderly groups, including Whites, Blacks, and Mexican Americans. Specifically, the present study focused on identifying race/ethnicity-related DIF items in the CES-D scale that function differentially, as well as a core set of CES-D items that
function equivalently across racially/ethnically diverse elderly groups. In order to take a conservative approach to identifying DIF, only items identified by both the CFA and IRT methods as IDF were treated as such.

Methods

Sample

The present study used two national datasets. The New Haven Established Populations for Epidemiologic Studies of the Elderly (EPESE) provided the White and Black samples and the Hispanic Established Populations for Epidemiologic Studies of the Elderly (H-EPESE) provided the Mexican American sample. These two datasets were selected because of (1) their inclusion of older adults aged 65 or older; (2) their use of the original 20-item CES-D with four-point rating scale; and (3) their inclusion of diverse racial/ethnic groups, especially Blacks in New Haven EPESE and Mexican Americans in the H-EPESE. The New Haven EPESE is a longitudinal study of community-dwelling participants aged 65 or older collected in one of four geographic locations (East Boston, New Haven, Iowa, and North Carolina) and includes Whites (N = 2,283) and Blacks (N = 529) at baseline (1982). The H-EPESE is also a longitudinal study of Mexican American participants (N = 3,050) aged 65 or older that collected from Texas, New Mexico, Colorado, Arizona, and California and was modeled after the design of the EPESE studies in order to compare with other populations in 1993-4.

Using the first waves of the New Haven EPESE and H-EPESE, subjects (2,283 Whites, 464 Blacks, and 2,623 Mexican Americans) were included in the analyses if they responded to all 20 CES-D items. Using listwise deletion method, 407 subjects of Whites (17.8 % of the total Whites), 65 subjects of Blacks (12.3 % of the total Blacks), and 427...
subjects of Mexican Americans (14.0 % of the total Mexican Americans) were excluded in the analyses due to missing data. Removed participants in each group had similar characteristics in terms of gender distribution, but were more likely to be older for all three racial/ethnic groups.

*Measures*

**CES-D Scale**

Both the New Haven EPESE and H-EPESE used the original 20-item version of the CES-D (Radloff, 1977). Respondents in both datasets were asked to report how often each symptom was experienced during the past week, and their symptoms were rated on a 4-point likert scale, with categories presented in the following order: “rarely or none of the time (coded as 0),” “some or a little of the time (coded as 1),” “much of the time (coded as 2),” and “most or all of the time (coded as 3).” The four positive items were reverse-coded and scale scores were computed by summing across twenty items to produce total scores ranging from 0 (no depressive symptoms) to 60 (severe depressive symptoms). Scores of 16 or higher are typically viewed as evidence of probable depression (Andresen, Carter, Malmgren, & Patrick, 1994). Internal consistency of the CES-D was satisfactory in the present sample: \( \alpha = .86 \) for Whites, \( \alpha = .84 \) for Blacks, and \( \alpha = .88 \) for Mexican Americans.

*Statistical Analysis*

Sample characteristics were first compared for three racial/ethnic groups. The ANOVA and the chi-square test were used to test for group differences, with \( p < .05 \) indicating a statistically significant difference. For descriptive purposes, the prevalence of probable depression based on the CES-D (total score ≥ 16) was compared using a chi-
square test. Individual item mean and standard deviation (SD) for the CES-D scale were also compared, with p < .05 indicating a statistically significant difference. Information on item responses of the 20 CES-D items was provided for Whites, Blacks, and Mexican Americans.

Because the DIF detection methods used in this investigation assumed that a single dominant factor underlies item responses, the unidimensionality of the CES-D was investigated using a principal component analysis (PCA) via SPSS and a confirmatory analysis via LISREL 8.8 (Jöreskog & Sörbom, 2006). The PCA results indicated that while the analysis produced 4 factors for Whites (1st factor = 5.86, % variance = 29.31; 2nd factor = 1.32, % variance = 6.61; 3rd factor = 1.20, % variance = 5.99; 4th factor = 1.05, % variance = 5.26), 5 factors for Blacks (1st factor = 5.49, % variance = 27.45; 2nd factor = 1.39, % variance = 6.98; 3rd factor = 1.17, % variance = 5.88; 4th factor = 1.08, % variance = 5.43; 5th factor = 1.02, % variance = 5.07), and 3 factors for Mexican Americans (1st factor = 6.84, % variance = 34.20; 2nd factor = 2.27, % variance = 11.35; 3rd factor = 1.14, % variance = 5.70), the ratio of first to second eigenvalue was 4.41, 3.95, and 3.01, respectively, which suggests that the data are essentially unidimensional (Lord, 1980; Stout, 1990). This result was supported by one-factor confirmatory factor analyses, where goodness-of-fit indices for three racial/ethnic groups all exceeded .90, indicating generally good fits of the one-factor model to the data. In these analyses Whites showed slightly better goodness-of-fit indices than Blacks and Mexican Americans (for Whites, CFI = .95, NFI = .94, NNFI = .95; for Blacks, CFI = .93, NFI = .90, NNFI = .93; and for Mexican Americans, CFI = .91, NFI = .90, NNFI = .90). These results confirmed the overall unidimensionality of the CES-D.
After verifying that the CES-D data overall were sufficiently unidimensional, the application of IRT and CFA DIF detection using the likelihood ratio tests proceeded. In both methods, this study followed the general approach to hypothesis testing described by Stark et al. (2006), since the latter showed high power and low Type I error rates across a wide variety of simulation conditions. In essence, the authors suggested testing for DIF by a common strategy that can be implemented in both CFA and IRT, which is called *free baseline with Bonferroni correction*. From CFA, this strategy involves use of a fully free baseline model (with the exception of a single referent item), which is statistically appropriate as the basis for subsequent nested model comparisons where one item at a time is constrained to be equal across groups. From IRT, this strategy incorporates the ideas of simultaneous comparisons of item parameters (discrimination-loadings and locations-intercepts) and strict $p$-values for flagging DIF items. Detail analytic procedures for CFA and IRT are described below.

*CFA DIF detection*

CFA DIF analyses involving item loadings and intercepts were conducted using an analogous strategy with LISREL 8.8 (Jöreskog & Sörbom, 2006). Information on item loadings and intercepts for three groups (Whites, Blacks, and Mexican Americans) is shown in Table 1. Using the free baseline model, where only the parameters of the referent are constrained across groups, baseline and constrained models were run in succession and the chi-square difference statistics for the nested model comparisons were evaluated using a Bonferroni corrected critical $p$-value. When the observed chi-square difference was greater than the corresponding critical chi-square value (Bonferroni corrected, $\chi^2 = 11.88$ with 2 degrees of freedom), the item was flagged DIF.
Table 1. Study 1 – Item Parameter Estimates for Whites, Blacks, and Mexican Americans

<table>
<thead>
<tr>
<th>CES-D Item</th>
<th>Whites (N = 1,876)</th>
<th>Blacks (N = 464)</th>
<th>Mexican Americans (N = 2,623)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>CFA</td>
<td>IRT</td>
<td>CFA</td>
</tr>
<tr>
<td></td>
<td>λ</td>
<td>τ</td>
<td>a</td>
</tr>
<tr>
<td>1</td>
<td>1.00</td>
<td>0.44</td>
<td>1.39</td>
</tr>
<tr>
<td>2</td>
<td>0.78</td>
<td>0.35</td>
<td>1.13</td>
</tr>
<tr>
<td>3</td>
<td>1.21</td>
<td>0.33</td>
<td>2.30</td>
</tr>
<tr>
<td>4</td>
<td>0.57</td>
<td>0.29</td>
<td>0.89</td>
</tr>
<tr>
<td>5</td>
<td>0.80</td>
<td>0.40</td>
<td>1.16</td>
</tr>
<tr>
<td>6</td>
<td>1.51</td>
<td>0.50</td>
<td>2.71</td>
</tr>
<tr>
<td>7</td>
<td>1.30</td>
<td>0.51</td>
<td>1.74</td>
</tr>
<tr>
<td>8</td>
<td>1.19</td>
<td>0.93</td>
<td>0.97</td>
</tr>
<tr>
<td>9</td>
<td>0.65</td>
<td>0.20</td>
<td>1.47</td>
</tr>
<tr>
<td>10</td>
<td>0.72</td>
<td>0.27</td>
<td>1.48</td>
</tr>
<tr>
<td>11</td>
<td>1.16</td>
<td>0.59</td>
<td>1.25</td>
</tr>
<tr>
<td>12</td>
<td>1.49</td>
<td>0.62</td>
<td>1.70</td>
</tr>
<tr>
<td>13</td>
<td>0.76</td>
<td>0.35</td>
<td>1.10</td>
</tr>
<tr>
<td>14</td>
<td>1.39</td>
<td>0.50</td>
<td>1.95</td>
</tr>
<tr>
<td>15</td>
<td>0.43</td>
<td>0.20</td>
<td>1.03</td>
</tr>
<tr>
<td>16</td>
<td>1.32</td>
<td>0.46</td>
<td>1.74</td>
</tr>
<tr>
<td>17</td>
<td>0.74</td>
<td>0.21</td>
<td>1.79</td>
</tr>
<tr>
<td>18</td>
<td>1.27</td>
<td>0.44</td>
<td>2.40</td>
</tr>
<tr>
<td>19</td>
<td>0.30</td>
<td>0.11</td>
<td>1.14</td>
</tr>
<tr>
<td>20</td>
<td>0.93</td>
<td>0.34</td>
<td>1.61</td>
</tr>
</tbody>
</table>

Note. λ (lambda) = loading (=slope); τ (tau) = threshold (= intercept); a = discrimination parameter; b₁, b₂, and b₃ = location parameters
IRT DIF detection

Because the CES-D scale was polytomous, Samejima’s (1969) Graded Response model was chosen in this present study, using the MULTILOG computer program (Thissen, 1991). For this model, each four-category item has one discrimination parameter \(a\) and three location parameters \(b_1, b_2,\) and \(b_3\). The discrimination parameter reflects the extent to which an item differentiates between levels of underlying depression, and items with higher \(a\) are generally preferred because they are more informative in a psychometric sense. The location parameters refer to the point on the underlying depression scale in which the probability is 50% for endorsing the first category relative to the last 3 categories \((b_1 - 0\) vs. 1, 2, 3\), the first 2 categories relative to the last 2 categories \((b_2 - 0, 1\) vs. 2, 3\), and the first 3 categories relative to the fourth category \((b_3 - 0, 1, 2\) vs. 3\), respectively.

To assess model-data fit, the MODFIT subroutine was used. To determine good model-data fit, this study used adjusted chi-squares to degrees of freedom ratios for item singles, doubles, and triples, which may provide more advanced chi-square methods. Adjusted chi-squares to degrees of freedom ratios for item singles, doubles, and triples all showed less than 3, indicating good model-data fit (mean adjusted chi-squares/df for singles = 0.001; mean adjusted chi-squares/df for doubles = 2.508; and mean adjusted chi-squares/df for triples = 2.801) (Drasgow, Levine, Tsien, Williams, & Mead, 1995).

The concurrent calibration method was subsequently used to put the reference and focal group parameters on a common metric with Item 1 as an anchor item. In this step, Whites (in two cases of White-Black and White-Mexican American comparisons) and Mexican Americans (only in the case of Mexican American-Black comparisons) were
designated as the reference group, whose latent mean was set to zero. The Mexican Americans (only in the case of White-Mexican American comparisons) and Blacks (in two cases of White-Black and Mexican American-Black comparisons) were designated as the focal group; its latent mean was free to vary. Item parameter estimates are presented in Table 1. As described for the CFA DIF method, the free-baseline model strategy was also used for each CES-D item, and differences in relative goodness of fit were examined with respect to critical chi-square statistics. Each chi-square difference was compared to Bonferroni corrected $p$-values (corrected, $\chi^2 = 16.31$ with 4 degrees of freedom), and items exhibiting DIF were flagged.

Results

Descriptive Information of Sample

As shown in Table 2, the 1,876 Whites, 464 Blacks, and 2,623 Mexican Americans were significantly different in terms of their age and gender distribution. In terms of age distribution, 46.7% of the Whites and 32.3% of both the Blacks and Mexican Americans were 75 or older. Whites included more individuals who were aged 75 or older. More than half were female for all three groups (57% for Whites, 63% for Blacks, and 58% for Mexican Americans). At the same time, there were significant differences in gender distribution ($\chi^2 = 6.98, p < .05$). Black participants were more likely to be female than Whites and Mexican Americans. A study variable, the CES-D, showed significantly different mean scores across three racial/ethnic groups, showing higher scores of Mexican Americans than that of the Whites or Blacks ($F = 33.82, p < .001$). Moreover, using the standard cutoff of 16 on the CES-D for evidence of probable
depression, Mexican Americans consistently exhibited a greater likelihood for depression than levels reported for the Whites or Blacks. Specifically, 16% of the Whites, 14.4% of the Blacks, and 23.1% of the Mexican Americans fell into the probable depression category ($\chi^2 = 44.17, p < .01$).

Table 2. Study 1 – Descriptive Characteristics of the Sample

<table>
<thead>
<tr>
<th></th>
<th>New Haven EPESE (N = 2,340)</th>
<th>H-EPESE (N = 2,623)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Whites (N=1,876)</td>
<td>Blacks (N=464)</td>
</tr>
<tr>
<td>Age (≥75)</td>
<td>45.7%</td>
<td>32.3%</td>
</tr>
<tr>
<td>Female</td>
<td>56.6%</td>
<td>63.4%</td>
</tr>
<tr>
<td>CES-D</td>
<td>8.03/8.39</td>
<td>7.88/7.82</td>
</tr>
<tr>
<td>Probable depression (≥ 16)</td>
<td>16.0%</td>
<td>14.4%</td>
</tr>
</tbody>
</table>

Note. New Haven EPESE = New Haven Established Populations for Epidemiologic Studies of the Elderly; H-EPESE = Hispanic Established Populations for Epidemiologic Studies of the Elderly; SD = Standard Deviation; CES-D = Center for Epidemiological Studies-Depression

* $p < .05$. *** $p < .001$. 58
Descriptive Item Statistics

The twenty CES-D items were evaluated using classical test theory (CTT) statistics, which suggest that the first step for the DIF analyses is to check for mean differences. As shown in Table 3, mean scores on each item were compared across the three racial/ethnic groups using the ANOVA test. Significant mean differences were identified for twelve of twenty items (Items 3, 4, 6, 7, 9, 10, 12, 14, 15, 16, 17, 18, and 19). In each case, Mexican Americans consistently appeared to have higher mean score than other two groups, with an exception of two items (Items 15 and 19) showing higher means among Blacks. Table 4 also summarizes item responses to each response option of the 20 CES-D items.

DIF Results

Results of CFA and IRT DIF analyses are summarized in Table 5. Because the CFA and IRT methods did not always identify the same DIF items across any two groups, I decided to focus on DIF items detected by both methods, as a more conservative strategy that would reduce possible Type I error rates. This strategy, using two or more sets of statistical results and explaining consistent DIF items across different procedures, has been suggested by a number of researchers to detect DIF (Hambleton, 2006; Hambleton & Rogers, 1989). Although reporting commonly identified DIF items was the main focus, however, attention was also given to uncommon DIF items across two methods as well as DIF-free items.
Table 3. Study 1 – Mean and Standard Deviation (SD) of the CES-D in Whites, Blacks, and Mexican American

<table>
<thead>
<tr>
<th>CES-D Items</th>
<th>Whites (N = 1,876)</th>
<th>Blacks (N = 464)</th>
<th>Mexican Americans (N = 2,623)</th>
<th>F</th>
</tr>
</thead>
<tbody>
<tr>
<td>1. I was bothered by things that usually don’t bother me</td>
<td>.44 (.81)</td>
<td>.36 (.79)</td>
<td>.43 (.78)</td>
<td>1.65</td>
</tr>
<tr>
<td>2. I did not feel like eating; my appetite was poor</td>
<td>.35 (.78)</td>
<td>.43 (.85)</td>
<td>.35 (.71)</td>
<td>2.57ab,c</td>
</tr>
<tr>
<td>3. I felt that I could not shake off the blues even with help from my</td>
<td>.33 (.73)</td>
<td>.28 (.66)</td>
<td>.37 (.74)</td>
<td>4.37abc</td>
</tr>
<tr>
<td>family or friends</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>4. I felt that I was just as good as other people</td>
<td>.29 (.77)</td>
<td>.26 (.71)</td>
<td>.87 (1.21)</td>
<td>203.74abc</td>
</tr>
<tr>
<td>5. I had trouble keeping my mind on what I was doing</td>
<td>.40 (.76)</td>
<td>.34 (.67)</td>
<td>.43 (.75)</td>
<td>2.81c</td>
</tr>
<tr>
<td>6. I felt depressed</td>
<td>.50 (.81)</td>
<td>.44 (.76)</td>
<td>.54 (.81)</td>
<td>3.58abc</td>
</tr>
<tr>
<td>7. I felt everything was an effort</td>
<td>.51 (.90)</td>
<td>.55 (.96)</td>
<td>.61 (.92)</td>
<td>7.645abc</td>
</tr>
<tr>
<td>8. I felt hopeful about future</td>
<td>.93 (1.22)</td>
<td>.99 (1.23)</td>
<td>.94 (1.16)</td>
<td>.46</td>
</tr>
<tr>
<td>9. I thought my life had been a failure</td>
<td>.20 (.60)</td>
<td>.20 (.61)</td>
<td>.26 (.64)</td>
<td>5.318abc</td>
</tr>
<tr>
<td>10. I felt fearful</td>
<td>.27 (.63)</td>
<td>.29 (.68)</td>
<td>.32 (.66)</td>
<td>4.34abc</td>
</tr>
<tr>
<td>11. My sleep was restless</td>
<td>.59 (.98)</td>
<td>.50 (.86)</td>
<td>.55 (.90)</td>
<td>1.90</td>
</tr>
<tr>
<td>12. I was happy</td>
<td>.62 (.99)</td>
<td>.52 (.89)</td>
<td>.85 (1.07)</td>
<td>39.46abc</td>
</tr>
<tr>
<td>13. I talked less than usual</td>
<td>.35 (.79)</td>
<td>.36 (.79)</td>
<td>.39 (.75)</td>
<td>1.39</td>
</tr>
<tr>
<td>14. I felt lonely</td>
<td>.50 (.87)</td>
<td>.48 (.84)</td>
<td>.47 (.83)</td>
<td>.72</td>
</tr>
<tr>
<td>15. People were unfriendly</td>
<td>.20 (.59)</td>
<td>.37 (.79)</td>
<td>.25 (.68)</td>
<td>13.55abc</td>
</tr>
<tr>
<td>16. I enjoyed life</td>
<td>.46 (.91)</td>
<td>.34 (.75)</td>
<td>.91 (1.12)</td>
<td>139.04abc</td>
</tr>
<tr>
<td>17. I had crying spells</td>
<td>.21 (.58)</td>
<td>.14 (.50)</td>
<td>.42 (.78)</td>
<td>66.49abc</td>
</tr>
<tr>
<td>18. I felt sad</td>
<td>.44 (.72)</td>
<td>.43 (.73)</td>
<td>.54 (.83)</td>
<td>10.49abc</td>
</tr>
<tr>
<td>19. I felt people disliked me</td>
<td>.11 (.43)</td>
<td>.25 (.64)</td>
<td>.18 (.53)</td>
<td>18.97abc</td>
</tr>
<tr>
<td>20. I could not get going</td>
<td>.34 (.71)</td>
<td>.34 (.69)</td>
<td>.35 (.73)</td>
<td>1.26</td>
</tr>
</tbody>
</table>

Note. * Reverse-coded item. a A significant mean difference between Whites and Blacks was obtained at the .05 level. b A significant mean difference between Whites and Hispanics was obtained at the .05 level. c A significant mean difference between Blacks and Hispanics was obtained at the .05 level. * p < .05. ** p < .01 *** p < .001.
Table 4. Study 1 – Item Responses of the CES-D Scale in Whites, Blacks, and Mexican Americans

<table>
<thead>
<tr>
<th>CES-D Items</th>
<th>Rarely/none of the time</th>
<th>Some of the time</th>
<th>Much of the time</th>
<th>Most/all of the time</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>W</td>
<td>B</td>
<td>M</td>
<td>W</td>
</tr>
<tr>
<td>1. Bothered by things</td>
<td>70.7</td>
<td>76.7</td>
<td>71.7</td>
<td>20.5</td>
</tr>
<tr>
<td>2. Poor appetite</td>
<td>79.1</td>
<td>74.4</td>
<td>76.4</td>
<td>13.1</td>
</tr>
<tr>
<td>3. Could not shake blues</td>
<td>78.4</td>
<td>80.4</td>
<td>74.8</td>
<td>14.2</td>
</tr>
<tr>
<td>4. As good as other people</td>
<td>85.3</td>
<td>86.4</td>
<td>60.5</td>
<td>4.5</td>
</tr>
<tr>
<td>5. Trouble concentrating</td>
<td>72.1</td>
<td>73.9</td>
<td>70.1</td>
<td>19.9</td>
</tr>
<tr>
<td>6. Felt depressed</td>
<td>64.7</td>
<td>67.2</td>
<td>62.3</td>
<td>25.8</td>
</tr>
<tr>
<td>7. Everything an effort</td>
<td>69.3</td>
<td>68.5</td>
<td>61.7</td>
<td>18.8</td>
</tr>
<tr>
<td>8. Hopeful about future</td>
<td>59.3</td>
<td>56.7</td>
<td>50.0</td>
<td>6.9</td>
</tr>
<tr>
<td>9. Life a failure</td>
<td>87.0</td>
<td>87.9</td>
<td>82.8</td>
<td>8.7</td>
</tr>
<tr>
<td>10. Felt fearful</td>
<td>80.7</td>
<td>79.7</td>
<td>76.3</td>
<td>15.0</td>
</tr>
<tr>
<td>11. Restless sleep</td>
<td>66.4</td>
<td>67.0</td>
<td>66.8</td>
<td>19.2</td>
</tr>
<tr>
<td>12. Happy</td>
<td>68.0</td>
<td>70.5</td>
<td>52.8</td>
<td>9.4</td>
</tr>
<tr>
<td>13. Talked less</td>
<td>78.6</td>
<td>78.2</td>
<td>74.2</td>
<td>12.9</td>
</tr>
<tr>
<td>14. Felt lonely</td>
<td>68.2</td>
<td>67.9</td>
<td>69.5</td>
<td>20.9</td>
</tr>
<tr>
<td>15. People unfriendly</td>
<td>87.2</td>
<td>76.9</td>
<td>84.7</td>
<td>8.8</td>
</tr>
<tr>
<td>16. Enjoyed life</td>
<td>76.4</td>
<td>79.5</td>
<td>51.7</td>
<td>6.9</td>
</tr>
<tr>
<td>17. Crying spells</td>
<td>85.2</td>
<td>89.9</td>
<td>72.5</td>
<td>10.6</td>
</tr>
<tr>
<td>18. Felt sad</td>
<td>66.7</td>
<td>67.2</td>
<td>63.9</td>
<td>26.1</td>
</tr>
<tr>
<td>19. People disliked me</td>
<td>92.2</td>
<td>82.8</td>
<td>87.0</td>
<td>6.0</td>
</tr>
<tr>
<td>20. Could not get going</td>
<td>76.8</td>
<td>75.6</td>
<td>76.5</td>
<td>16.8</td>
</tr>
</tbody>
</table>

Note. W = Whites (N=1,876); B = Blacks (N=464); H = Mexican Americans (N=2,623). + Reverse-coded item. † A significant mean difference between Whites and Blacks was obtained at the .05 level. A significant mean difference between Whites and Mexican Americans was obtained at the .05 level. A significant mean difference between Blacks and Mexican Americans was obtained at the .05 level.
Table 5. Study 1 – DIF Results from CFA and IRT Methods

<table>
<thead>
<tr>
<th>Models</th>
<th>Whites vs. Blacks</th>
<th></th>
<th>Whites vs. Mexican Americans</th>
<th></th>
<th>Mexican Americans vs. Blacks</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>CFA(\Delta df = 2)</td>
<td>IRT(\Delta df = 4)</td>
<td>CFA(\Delta df = 2)</td>
<td>IRT(\Delta df = 4)</td>
<td>CFA(\Delta df = 2)</td>
<td>IRT(\Delta df = 4)</td>
</tr>
<tr>
<td></td>
<td>Chi-Square (Difference)</td>
<td></td>
<td>Chi-Square (Difference)</td>
<td></td>
<td>Chi-Square (Difference)</td>
<td></td>
</tr>
<tr>
<td>Baseline Model (Referent: Item1)</td>
<td>2033.85</td>
<td>27259.1</td>
<td>6344.63</td>
<td>58833.0</td>
<td>5495.77</td>
<td>44396.8</td>
</tr>
<tr>
<td>Comparison Models</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Constrained Item 2</td>
<td>(10.75)</td>
<td>(15.6)</td>
<td>(0.93)</td>
<td>(80.3)</td>
<td>(9.87)</td>
<td>(33.3)</td>
</tr>
<tr>
<td>Constrained Item 3</td>
<td>(8.39)</td>
<td>(13.4)</td>
<td>(18.24)</td>
<td>(18.4)</td>
<td>(19.51)</td>
<td>(25.8)</td>
</tr>
<tr>
<td>Constrained Item 4*</td>
<td>(8.83)</td>
<td>(3.5)</td>
<td>(893.04)</td>
<td>(346.2)</td>
<td>(307.16)</td>
<td>(124.0)</td>
</tr>
<tr>
<td>Constrained Item 5</td>
<td>(5.33)</td>
<td>(7.8)</td>
<td>(12.95)</td>
<td>(81.4)</td>
<td>(14.62)</td>
<td>(30.8)</td>
</tr>
<tr>
<td>Constrained Item 6</td>
<td>(6.82)</td>
<td>(7.2)</td>
<td>(17.57)</td>
<td>(39.7)</td>
<td>(13.57)</td>
<td>(29.6)</td>
</tr>
<tr>
<td>Constrained Item 7</td>
<td>(5.65)</td>
<td>(17.2)</td>
<td>(52.88)</td>
<td>(60.1)</td>
<td>(4.52)</td>
<td>(52.8)</td>
</tr>
<tr>
<td>Constrained Item 8*</td>
<td>(10.75)</td>
<td>(17.8)</td>
<td>(41.34)</td>
<td>(212.2)</td>
<td>(1.68)</td>
<td>(72.4)</td>
</tr>
<tr>
<td>Constrained Item 9</td>
<td>(4.89)</td>
<td>(5.0)</td>
<td>(32.60)</td>
<td>(32.2)</td>
<td>(7.75)</td>
<td>(6.7)</td>
</tr>
<tr>
<td>Constrained Item 10</td>
<td>(4.52)</td>
<td>(7.2)</td>
<td>(31.33)</td>
<td>(52.8)</td>
<td>(0.77)</td>
<td>(15.3)</td>
</tr>
<tr>
<td>Constrained Item 11</td>
<td>(9.62)</td>
<td>(10.1)</td>
<td>(6.03)</td>
<td>(80.8)</td>
<td>(1.99)</td>
<td>(37.3)</td>
</tr>
<tr>
<td>Constrained Item 12*</td>
<td>(14.83)</td>
<td>(5.8)</td>
<td>(214.43)</td>
<td>(225.4)</td>
<td>(177.44)</td>
<td>(53.9)</td>
</tr>
<tr>
<td>Constrained Item 13</td>
<td>(0.42)</td>
<td>(4.3)</td>
<td>(12.66)</td>
<td>(80.1)</td>
<td>(2.05)</td>
<td>(25.4)</td>
</tr>
<tr>
<td>Constrained Item 14</td>
<td>(1.20)</td>
<td>(6.2)</td>
<td>(6.98)</td>
<td>(29.9)</td>
<td>(1.77)</td>
<td>(18.2)</td>
</tr>
<tr>
<td>Constrained Item 15</td>
<td>(29.43)</td>
<td>(41.5)</td>
<td>(34.53)</td>
<td>(24.3)</td>
<td>(15.65)</td>
<td>(43.0)</td>
</tr>
<tr>
<td>Constrained Item 16*</td>
<td>(33.65)</td>
<td>(10.7)</td>
<td>(619.80)</td>
<td>(341.2)</td>
<td>(267.81)</td>
<td>(97.3)</td>
</tr>
<tr>
<td>Constrained Item 17</td>
<td>(21.96)</td>
<td>(5.0)</td>
<td>(383.23)</td>
<td>(105.7)</td>
<td>(294.00)</td>
<td>(49.1)</td>
</tr>
<tr>
<td>Constrained Item 18</td>
<td>(0.16)</td>
<td>(11.6)</td>
<td>(77.97)</td>
<td>(50.6)</td>
<td>(20.87)</td>
<td>(38.2)</td>
</tr>
<tr>
<td>Constrained Item 19</td>
<td>(47.72)</td>
<td>(48.5)</td>
<td>(146.12)</td>
<td>(72.6)</td>
<td>(10.83)</td>
<td>(38.5)</td>
</tr>
<tr>
<td>Constrained Item 20</td>
<td>(1.39)</td>
<td>(7.4)</td>
<td>(17.20)</td>
<td>(74.3)</td>
<td>(6.31)</td>
<td>(29.4)</td>
</tr>
<tr>
<td>Total # DIF Items</td>
<td>5</td>
<td>4</td>
<td>16</td>
<td>19</td>
<td>9</td>
<td>17</td>
</tr>
</tbody>
</table>

Note. Items in bold and underlined are common DIF items across CFA and IRT methods in each cross-racial/ethnic comparison.

\* Reverse-coded item. \(a\) In CFA, DIF flagged if \(\chi^2\) were > 11.88. \(b\) In IRT, DIF flagged if \(\chi^2\) were > 16.31.
In both CFA and IRT methods, this study used Item 1 as a referent and had nineteen model comparisons in which each of the CES-D items was constrained to be equal across groups in each model. Among all three group comparisons (e.g., Whites vs. Blacks; Whites vs. Mexican Americans; and Mexican Americans vs. Blacks), White-Mexican American group comparisons exhibited the greatest number of DIF items (16 common DIF items) and White-Black group comparisons flagged the fewest number of DIF items (2 common DIF items). Only one item, Item 15 (‘people were unfriendly’), consistently exhibited DIF across all three group comparisons, which suggest Item 15 functioned differently in each of the three groups. Regarding the responses to this interpersonal relation item (Item 15), Blacks were favored over Whites and Mexican Americans, while the latter were favored over Whites. These results clearly suggested that Blacks were more likely to endorse Item 15 (‘people were unfriendly’) than Whites and Mexican Americans.

Across all three group comparisons, sixteen of twenty items flagged DIF at least once. In other words, 80% of the twenty CES-D items functioned differently across at least two of the three groups. Among these sixteen DIF items, fourteen items (Items 2, 3, 4, 5, 6, 7, 9, 10, 12, 13, 16, 17, 18, and 20) favored Mexican Americans, indicating Mexican Americans had a greater likelihood to endorse these items than Whites or Blacks. Two items related to interpersonal problems (the previously mentioned Item 15, ‘people were unfriendly’ and Item 19, ‘people disliked me’) favored Blacks, suggesting Blacks were more likely to endorse them than Whites and Mexican Americans. Only four items (Items 1, ‘bothered by things’; 2, ‘poor appetite’; 11, ‘restless sleep’; and 14, ‘felt lonely’) showed no evidence of DIF, suggesting they were common depressive
symptom items that functioned equivalently across Whites, Blacks, and Mexican Americans. Three of these four items (Items 1, 2, and 11) are associated with somatic symptoms, while the fourth (Item 14) is usually found in Radloff’s depressive affect factor.

*Whites versus Blacks*

In the comparison of Whites and Blacks, CFA identified five DIF items (Items 12, 15, 16, 17, and 19) and IRT flagged four DIF items (Items 7, 8, 15 and 19). Two DIF items (Items 15, ‘people were unfriendly’ and 19, ‘people disliked me’) were identified in both CFA and IRT methods. The same findings have been previously observed in two DIF studies (Cole et al., 2000; Yang & Jones, in press) and the present study supported their findings using the same samples drawn from the New Haven EPESE but different DIF methods (Cole et al. used the Mantel-Haenszel (MH) adjustment; Yang & Jones used the multiple indicators, multiple causes (MIMIC) model). As was the case in these two studies, in the present analyses, Blacks were more likely than Whites to endorse the two interpersonal items.

*Whites versus Mexican Americans*

In the White-Mexican American comparison, CFA flagged sixteen items flagged as DIF and IRT identified nineteen DIF items. All sixteen of the CFA-flagged items were also identified by IRT: Items 3, 4, 5, 6, 7, 8, 9, 10, 12, 13, 15, 16, 17, 18, 19 and 20. In other words, 80% of the CES-D items were indicated to function differently across Whites and Mexican Americans. Notably, all four positive affect items (Items 4, ‘felt good as good as others’; 8, ‘felt hopeful about the future’; 12, ‘felt happy’; 16, ‘enjoyed life’) exhibited DIF in both approaches. All DIF items except four positive items favored
Mexican Americans, which means Mexican Americans were more likely to endorse these items compared to Whites. Responses to the four positive affect items, here actually representing low positive affect items since these four items were reverse-scored, favored Mexican Americans, indicating Mexican Americans were less likely to endorse all positive feeling items.

*Mexican Americans versus Blacks*

In the comparison of Mexican Americans and Blacks, CFA identified nine DIF items and IRT identified seventeen. There were nine common DIF items (Items 3, 4, 5, 6, 12, 15, 16, 17, and 18), suggesting that nearly half of the CES-D items functioned differently for Mexican Americans and Blacks. Three of four positive affect items (Items 4, ‘felt good as good as others’; 12, ‘felt happy’; 16, ‘enjoyed life’) showed DIF. Compared to Blacks, Mexican Americans were more likely to endorse all DIF items (including three low positive feeling items) except one interpersonal problem item (Item 15, ‘people were unfriendly’). Blacks were more likely to endorse Item 15 (‘people were unfriendly’) than were Mexican Americans. Three positive affect items favored Mexican Americans over Blacks, indicating Mexican Americans were more likely to endorse low positive feeling items. This finding parallels those for the comparison of Mexican Americans with Whites, and suggests that Mexican Americans may be less likely to report their positive feelings than Blacks.

*CFA versus IRT*

In the comparison of DIF items identified by CFA and IRT methods, common DIF items were found for two of seven items in the White-Black comparison, sixteen of nineteen items in the White-Mexican American comparison, and nine of seventeen items
in the Mexican American-Black comparison. Most of the uncommon DIF items identified by either CFA or IRT were items detected by the IRT method, suggesting that the IRT DIF method was more likely to detect DIF over CFA in the present study.

Discussion

The present study investigated the cultural equivalence of the CES-D across three racial/ethnic groups of older adults, including Whites and Blacks from the New Haven EPESE and Mexican Americans from the Hispanic EPESE. It is worth pointing out that the present study may be the first to examine item bias in the full version of the CES-D across racially/ethnically diverse elderly populations, and especially to include older Mexican Americans in the comparisons. The goal of this study was to identify items in the CES-D that function differentially across diverse racial/ethnic groups. The approach followed included two different analytic strategies, the CFA and the IRT, to identify DIF, and hence was able to examine items detected by one or the other, or both, strategies.

DIF analyses in Study 1 indicated that across all three racial/ethnic groups, sixteen of the twenty CES-D items displayed statistically significant DIF in both CFA and IRT methods. In other words, 80% of the CES-D items functioned differently across Whites, Blacks, and Mexican Americans. That left four items (Items 1, bothered by things; 2, poor appetite; 11, restless sleep; and 14, felt lonely) of the twenty CES-D items as being identified to function similarly across all three racial/ethnic groups of older adults. The bottom line is that all three groups clearly did not report their symptoms of depression on the CES-D equivalently, a finding that emphasizes the need for further study of measurement equivalence in at least this depression screening instrument.
In the comparison of Whites and Blacks, results supported previous findings from two published DIF studies on the CES-D items among older adults that used the same dataset employed in the present analysis (Cole et al., 2000; Yang & Jones, in press). Compared to Whites, Blacks consistently over-endorsed two interpersonal relation items (‘people were unfriendly’ and ‘people disliked me’), which may reflect perceptions of racial discrimination by Blacks. It has been well documented that Blacks generally experience more disadvantaged social conditions than Whites, and are more likely to report racial discrimination (e.g., Ren, Amick, & Williams, 1999; Williams, 2005). Disproportionate responses to two interpersonal problem items that may confound depressive symptoms with perceived racial prejudice have been observed in other published studies with Blacks (e.g., Blazer et al., 1998), although they were not testing DIF. These results suggest that researchers who are interested in investigating depressive symptoms among Blacks should pay careful attention to the CES-D items involving interpersonal relations.

No previous work has addressed item bias in the CES-D among older Mexican Americans. The most striking finding in the present DIF analyses was the general lack of measurement equivalence of the CES-D scale in the comparison of Mexican Americans to Whites and Blacks. Sixteen of the twenty CES-D items were shown to function differently in the comparison of Mexican Americans and Whites. Remarkably, Mexican Americans were predisposed to endorse all of the sixteen DIF items (here, including four low positive affect items) compared with their counterparts. The comparison with Blacks were nearly as dramatic, with nearly half of the items on the CES-D manifesting item
bias and with the results indicating higher levels of endorsement by the Mexican Americans.

A greater tendency to endorse depressive symptoms among Mexican Americans can be partially explained by research indicating that Mexican Americans are less hesitant to admit their symptoms of psychological distress (Haberman, 1970). This response style may lead to less underreporting symptoms of depression among Mexican Americans (McHorney & Fleishman, 2006), which would help to explain their relatively high depression scores. Overall, these results suggest that using the standard cut-off scores of 16 or higher (Andresen et al., 1994) with Mexican American elders may lead to misclassification, resulting in elevated false positive rates. Future work is warranted on the appropriate cut-off scores of the CES-D in Mexican American population.

One intriguing finding was that Mexican Americans appeared to be much less likely to endorse positive affect items than the other two groups. This finding parallels results reported in Iwata and colleagues’ (2002) two studies of young adults. Their studies found that Hispanics were more likely to report the absence of positive feelings than were Whites. In other words, the two studies showed that Hispanic young adults were less likely to endorse positive affect items compared to Whites, suggesting that Hispanics tend to inhibit at least the reporting—if not the actual experience—of positive feelings such as feeling good about oneself, feeling happy, feeling hopeful about the future and enjoying life. In addition, these positive item bias in the CES-D favoring Whites and Blacks may be also explained by relatively little hesitation to express positive feelings among Whites and Blacks. There is, in fact, some literature suggesting that positive feelings are prominent in mainstream American culture (Ying, 1989), and also
that life in mainstream American culture may generate more positive feelings in daily life (Iwata & Buka, 2002). Consistent with previous studies showing measurement equivalence of all positive affect items in the CES-D among older Whites and Blacks (e.g., Cole et al., 2000; Yang & Jones, in press), in the present study, Whites and Blacks showed a similar response tendency in expressing their positive feelings, indicating that they may share values, attitudes, and beliefs regarding the expression of positive feelings. There is, of course, also the clinically important possibility that the similarities and differences may reflect similarities and differences in the actual experience of emotions.

A related finding to the abovementioned response tendency to positive feeling items is that Mexican Americans showed greater endorsement of the extreme category on positive affect items, which directly affected item bias on positive feeling items. Compared with Whites and Blacks, Mexican Americans had a greater tendency to select ‘rarely/none of the time,’ one of the end points, when they were asked to report their positive feelings, such as feeling good about themselves (Item 4), feeling happy (Item 12), and enjoying life (Item 16). This disproportionate extreme response style has been previously reported in the literature (e.g., Clarke, 2000; Marin, Gamba, & Marin, 1992; Hui & Triandis, 1989), with a suggestion that Hispanics may prefer extreme responses because of a cultural value that associates extreme responses with sincerity and conviction. It is noteworthy that a greater tendency to select the extreme response style was found only for positive affect items, which suggest responses to positive feeling items among Mexican Americans may be associated with their inhibited endorsement of positive feelings as well as extreme response style. More importantly, this response bias
resulted in a higher total mean score of the CES-D and a significantly larger percentage of clinically depressed individuals.

There was also evidence that Hispanics reported somatic symptoms differently than did the White groups. The literature has suggested that Hispanics are more likely to somatize their psychological distress such as depressive symptoms (e.g., Angel & Guarnaccia, 1989; Fabrega, 1990). The present study found that four of seven somatic symptom items (Items 5, ‘trouble concentration’; 7, ‘everything is an effort’; 13, ‘talked less’; and 20, ‘could not get going’) exhibited DIF in the Mexican American-White comparison and one somatic item (Item 5, ‘trouble concentration’) in the Mexican American-Black comparison. In all cases there was a greater endorsement among Mexican Americans. These findings partially supported the interaction between ethnicity (favoring Mexican Americans) and somatic symptom DIF items in the CES-D.

Interestingly enough, however, this study found three of the four common items (i.e., no DIF items) found to function equivalently across all three groups were somatic symptom items. These no-DIF items were ‘bothered by things,’ ‘poor appetite,’ and ‘restless sleep,’ although ‘bothered by things’ was used as a referent. Given the fact that this study focused on only older adults and that older adults may tend to somatize their depressive symptoms (e.g., Norris, Arnau, Meagher, & Bramson, 2005), it seems that the interaction between DIF in somatic symptom items and race/ethnicity may be confounded with older age in this sample.

It should be emphasized that this was the first study using two popular DIF methods, both CFA and IRT methods, to detect common DIF with the CES-D. The use of the joint approaches has been suggested by a number of researchers as a way of
ensuring a comprehensive test of measurement equivalence (e.g., Hambleton, 2006; Hidalgo-Montesinos & Gomez-Benito, 2003; Wang and Russell, 2005). In fact, using multiple methods has been recommended for cross-cultural researchers to identify cultural differences more accurately (e.g., Schaffer, & Riordan, 2003). In addition to following their suggestion, this study used a unified strategy of DIF detection with CFA and IRT suggested by Stark et al. (2006), in order to provide higher power and lower Type I error rates, which is of particular concern in DIF detection method. Although CFA and IRT found some discrepant DIF items, the majority of DIF items were identified by both methods. Results suggest that the use of both methods may be helpful for detecting item bias. For example, when this study applied this strategy to detect and interpret DIF, results from this study in the comparison of Whites and Blacks showed perfect agreement with two previous studies (Cole et al., 2000; Yang & Jones, in press) that used the same sample as employed here. This suggests the combined use of the two statistical approaches increases the accuracy for testing measurement equivalence of depressive symptom inventories. However, results also raise questions about how to interpret DIF items that are only detected by one method. Careful attention and interpretation should be made to these uncommon DIF items across different methods for future research.

From a methodological point of view, Study 1 reinforces the need for careful evaluation of measurement equivalence across diverse groups. Of particular interest were the apparent response inhibition for positive affect items evident among Mexican American elders and the two interpersonal problem items where Blacks appeared to have a greater predisposition for endorsement. When instruments are used to screen and assess
for depression in diverse racial/ethnic populations, researchers and practitioners should be aware of the risk that individuals from different cultural backgrounds may tend to be misclassified, such as false positives, false negatives, or both, leading directly to under- or over-diagnosis for depression. Use of inaccurate measures could also lead to misguided public policies.

Although the findings of Study 1 hold implications to research, practice, and public policies, limitations should be noted. One factor that was not controlled in the study was the potential influence of historical time and cohort differences between the samples from the New Haven EPESE and the H-EPESE. The New Haven EPESE was collected in 1981-1982, whereas the H-EPESE was collected in 1993-4. The over ten year differences between those two samples may have led to differential response patterns. In addition, this study included a relatively small sample of Blacks. Both limitations underscore the importance of appropriate nationally representative datasets that can provide enough information to capture racial/ethnic disparities in health.

In summary, Study 1 highlights the importance of considering symptoms of depression that may be experienced and expressed differently by diverse cultural groups. Mexican American elders, in particular, were found to differ substantially from White elders in their predisposition to respond. Black elders, in general, were much more likely to respond in patterns similar to those of the White elders. The reasons underlying the differences, as well as the similarities, are at present unknown. Clearly, more work remains to be done, especially with regard to understanding potential sources of DIF, such as sociodemographic characteristics. Ultimately, this avenue of research may lead
to the development of a screening tool that is as free of item bias as possible across diverse racial/ethnic groups.
CHAPTER THREE: STUDY 2

SOCIODEMOGRAPHIC-RELATED MEASUREMENT BIAS

Introduction

Responding to the national commitment to reduce racial/ethnic disparities in health and health care (e.g., U.S. Department of Health and Human Services, 2005), this study addressed one important but little-studied area in health disparities research: measurement equivalence of mental health screening tools. While a number of studies have addressed the question of measurement equivalence, most have simply compared factor structures across diverse racial/ethnic groups with no consideration of potential item bias. Measurement equivalence is of particular importance in health disparities research because if items on a measure have differential meanings or validity across diverse groups, group comparisons may be misleading and the prevalence estimates may be inaccurate. Especially when self-report screening tools are applied in diverse racial/ethnic groups, particular attention should be paid to whether item response levels are systematically inflated or deflated by factors, such as cultural values and gender role that are unrelated to the a targeted construct such as depressive symptoms (Stewart & Nápoles-Springer, 2003). In cases where these potentially biasing factors operate differentially across diverse racial/ethnic groups, apparent group differences or similarities assessed by the self-report instrument could be the result of response bias —what has been called differential item functioning (DIF) — rather than true group differences.
Two analytic approaches have been used for testing measurement equivalence, confirmatory factor analysis (CFA) and item response theory (IRT). Most studies examining DIF have only relied on results from one of these methods. The two methods, however, often produce different results, since they differ in underlying approaches (Raju, Laffitte, & Byrne, 2002; Stark, Chernyshenko, & Drasgow, 2006; Teresi, 2006). CFA is based on a linear model, tests invariance of loading parameter first and intercepts later, and uses a free baseline model for hypothesis testing. In contrast, IRT is based on a nonlinear model, tests invariance of discrimination and location parameters simultaneously, and uses a constrained baseline model strategy for hypothesis testing. The implications of these differences, for item level comparisons, are presently unknown; it is only recently that simulation studies are being run to compare the two approaches to DIF detection (e.g., Meade & Lautenschlager, 2004; Raju et al., 2002; Stark et al., 2006). Until such studies are completed some psychometricians are recommending applying multiple DIF detection approaches for more accurate DIF results and for more definitive information concerning which items are showing DIF and which ones are not (e.g., Hambleton, 2006; Wang & Russell, 2005). According to Hambleton (2006), it may be useful to focus on the items that reveal consistent DIF across different methods. Taking the abovementioned guidelines into account, one of the objectives of the present study was to apply these two DIF methods as a means of identifying items that are consistently classified as DIF for more accurate results and stronger conclusions.

A related, and also understudied area in measurement equivalence is the influence of sociodemographic characteristics (Stewart & Nápoles-Springer, 2003). Unlike sociodemographic-related measurement bias, race/ethnicity DIF on self-report screening
tools has been addressed in previous research (e.g., Kim, Chiriboga, & Jang, 2007). Recent literature on U.S. racial/ethnic populations documents the substantial sociodemographic diversity among America’s racial/ethnic group (Williams, 2005). A number of studies have suggested that sociodemographic differences may play a causal role in racial/ethnic disparities in health (e.g., Alwin & Wray, 2005; Mirowsky & Ross, 2003; Williams, 2005). However, little is known about how such characteristics may also influence the reporting of health symptoms and the completion of instruments. Sociodemographic characteristics may reflect fundamental differences in the experience of symptoms that give rise to the lack of measurement equivalence (McHorney & Fleishman, 2006).

These sociodemographic characteristics in fact are fundamental to shaping the different experiences and lived realities of people (e.g., Wray, Alwin, & McCammon, 2005). In turn these differences in experiences may influence people’s values, perceptions, and views and could, among other things, systematically inflated or deflated item response levels. More importantly, even shared sociodemographic conditions may not confer similarity across diverse racial/ethnic groups, which suggests the possibility of interactions between race/ethnicity and sociodemographic characteristics regarding measurement equivalence. For these reasons, this Study 2 was interested in not only the overall effect of sociodemographic strata on the measurement equivalence but also the differential effect of sociodemographic factors on the measurement equivalence in diverse racial/ethnic groups.

The instrument I selected to evaluate measurement equivalence was the Center for Epidemiological Studies –Depression Scale (CES-D; Radloff, 1977). Since its initial
development in 1977, the 20-item CES-D has been used in a substantial number of studies. Despite its wide use in diverse populations and confirmed general usefulness (Mui, Burnette, & Chen, 2002), there has been little attempt at the systematic assessment of the CES-D items that function differentially across diverse groups.

A few CES-D studies have examined the effect of sociodemographic characteristics (e.g., age, gender, or educational attainment) on measurement equivalence (e.g., Cole, Kawachi, Maller, & Berkman, 2000; Yang & Jones, in press). Cole and colleagues (2000) studied item-level biases in the CES-D across groups varying in age, gender, and race/ethnicity (Whites vs. Blacks). They found one gender biased item (‘crying’) and two race/ethnicity biased items (‘people were unfriendly’ and ‘people disliked me’). Using a different DIF detection method (the multiple indicators, multiple causes [MIMIC] model), Yang and Jones (in press) also successfully replicated findings from Cole and colleagues (2000). In both studies, two interpersonal relation items had a higher predisposition for Blacks and the ‘crying’ item had higher endorsement among women. In their studies, age and gender item biases within each racial/ethnic group were not analyzed in ways that might have captured more of the fundamental differences that give rise to item bias. Neither study investigated the effect of educational attainment on the measurement equivalence of the CES-D among older adults.

Perhaps more importantly, none of the previous work has fully considered the sociodemographic-related measurement bias in the CES-D within—as opposed to across—different racial/ethnic groups. Each racial/ethnic group has its unique sociodemographic profiles which may underlie within group differences (Alwin & Wray, 2005), and which may also in themselves result in differing predispositions to respond
depressive symptom items (Nguyen, Kitner-Triolo, Evans, & Zonderman, 2004). Therefore, it may be meaningful to examine sociodemographic-related item bias in the CES-D scale within each racial/ethnic group. By identifying sociodemographic-related item bias in the CES-D within each racial/ethnic group, future research may be in a better position to explain why diverse racial/ethnic groups respond differentially to the certain items of the CES-D scale.

The purpose of this Study 2 was to examine the sociodemographic-related item bias of the CES-D in the total sample as well as three racial/ethnic elderly groups: Whites, Blacks, and Mexican Americans. Due to the relatively large sample sizes, this investigation had a unique opportunity to investigate possible interaction effects of sociodemographic variables and race/ethnicity (Whites, Blacks, and Mexican Americans). The use of two DIF methods to examine item bias in the CES-D also represents a potential contribution to the field. With respect to findings from this dual use approach, this study took a conservative approach to identifying DIF by recognizing DIF only when the same item was identified by both the CFA and IRT methods.

Methods

Sample

Two national datasets were used for this study. The New Haven Established Populations for Epidemiologic Studies of the Elderly (EPESE) provided the White and Black samples and the Hispanic Established Populations for Epidemiologic Studies of the Elderly (H-EPESE) provided the Mexican American sample. These two datasets were selected because of (1) their inclusion of older adults aged 65 or older; (2) their use of the original 20-item CES-D with a four-point rating scale; and (3) their inclusion of Whites...
and Blacks in New Haven EPESE and Mexican Americans in the H-EPESE. The New Haven EPESE is a longitudinal study of community-dwelling participants aged 65 or older, and is one of four similar studies (the other sites collected data in East Boston, Iowa, and North Carolina). The H-EPESE, also a longitudinal study, included Mexican American participants aged 65 or older and living in Texas, New Mexico, Colorado, Arizona, and California. The H-EPESE was modeled after the design of the EPESE studies in order to compare with other populations in 1993-4.

Using the first waves of the New Haven EPESE and H-EPESE, subjects were excluded if they had any missing data on 20 CES-D items. This resulted in the listwise deletion of 407 Whites (17.8 % of the total Whites), 65 Blacks (12.3 % of the total Blacks), and 427 Mexican Americans (14.0 % of the total Mexican Americans). Further 41 subjects (32 Whites and 9 Blacks) were excluded due to missing data on their educational attainment. No missing data was found with regard to age and gender. The remaining sample comprised 1,844 Whites, 455 Blacks, and 2,623 Mexican Americans. Removed participants in each group had similar characteristics in terms of gender distribution and educational attainment, but were more likely to be older for all three racial/ethnic groups.

Measures

*CES-D scale*

Study 2 also used the CES-D as a target instrument. Both the New Haven EPESE and H-EPESE used the original 20-item version of the CES-D (Radloff, 1977). Respondents in both datasets were asked to report how often each symptom was experienced during the past week, and their symptoms were rated on a 4-point likert
scale, with categories presented in the following order: “rarely or none of the time (coded as 0),” “some or a little of the time (coded as 1),” “much of the time (coded as 2),” and “most or all of the time (coded as 3).” The four positive items were reverse-coded and scale scores were computed by summing across twenty items to produce total scores ranging from 0 (no depressive symptoms) to 60 (severe depressive symptoms). Scores of 16 or higher are typically viewed as evidence of probable depression (Andresen, Carter, Malmgren, & Patrick, 1994). Internal consistency of the CES-D was satisfactory in the present sample: $\alpha = .87$ for the total sample, $\alpha = .86$ for Whites, $\alpha = .84$ for Blacks, and $\alpha = .88$ for Mexican Americans.

**Statistical Analysis**

Study 2 used the same methodology used in Study 1. Because the DIF detection methods used in this investigation assume that a single dominant factor underlies CES-D item responses (Stark et al., 2006), the unidimensionality assumption was first evaluated using a principal component analysis (PCA) via SPSS and a confirmatory factor analysis (CFA) via LISREL 8.8 (Jöreskog & Sörbom, 2006). This was done within each sociodemographic characteristic group (e.g., younger vs. older, men vs. women, more or less educated) within each of the three racial/ethnic groups as well as for the total sample. Twenty four runs for each PCA and CFA were performed.

After verifying the underlying unidimensionality of the CES-D, the application of IRT and CFA DIF detection using the likelihood ratio tests proceeded. In both methods, this study followed the general approach to hypothesis testing described by Stark et al. (2006), since the latter approach showed high power and low Type I error rates across a wide variety of simulation conditions. In essence, the authors suggest testing for DIF by
a common strategy that can be implemented in both CFA and IRT. The strategy is called
free baseline with Bonferroni correction. A fully free baseline model (with the exception
of a single referent item) was used as the basis for subsequent nineteen nested model
comparisons where one item at a time was constrained to be equal across groups. Item
parameters were compared simultaneously (discrimination−loadings and
locations−intercepts), using Bonferroni corrected $p$-values for flagging DIF items.
Detailed analytic procedures for CFA and IRT are described below.

**DIF Detection**

*CFA DIF detection.* CFA DIF analyses involving item loadings and intercepts
were conducted using an analogous strategy with LISREL 8.8. Using the free baseline
model, where only the parameters of the referent are constrained across groups, baseline
and constrained models were run in succession. The chi-square difference statistics for
the nested model comparisons were evaluated using a Bonferroni corrected critical $p$-
value. When the observed chi-square difference was greater than the corresponding
critical chi-square value (Bonferroni corrected, $\chi^2 = 11.88$ with 2 degrees of freedom), the
item was flagged DIF.

*IRT DIF detection.* Because the CES-D scale is polytomous, the Graded
Response Model (GRM; Samejima, 1969) was estimated using the MULTILOG
computer program (Thissen, 1991). For the GRM, each four-category item has one
discrimination parameter ($a$) and three location parameters ($b_1$, $b_2$ and $b_3$). The
discrimination parameter reflects the extent to which an item differentiates between
levels of underlying depression; items with higher $a$ are generally preferred because they
are more informative in a psychometric sense. The location parameters refer to the point
on the underlying depression scale in which the probability is 50% for endorsing the first category relative to the last 3 categories \((b_1 - 0 \text{ vs. } 1, 2, 3)\), the first 2 categories relative to the last 2 categories \((b_2 - 0 \text{ vs. } 1, 2, 3)\), and the first 3 categories relative to the fourth category \((b_3 - 0, 1, 2 \text{ vs. } 3)\), respectively. To determine whether the GRM used for parameter estimation adequately fits the data, chi-square fit statistics were assessed using the MODFIT program. Adjusted chi-squares to degrees of freedom ratios for each item were all less than 3, indicating a good model-data fit (Drasgow, Levine, Tsien, Williams, & Mead, 1995).

The concurrent calibration method was subsequently used to put the reference and focal group parameters on a common metric with Item 1 as an anchor item. In this step, groups consisting of those who were younger (aged 65 to 74), female, and more educated (8th grade or more) were designated as the reference group, whose latent mean was set to zero. Groups with older age (75 or older), male gender, and lower education (less than 8th grade) were designated as the focal group; focal group latent mean were free to vary. As described for the CFA DIF method, the free-baseline model strategy was also used for each CES-D item, and differences in relative goodness of fit were examined with respect to critical chi-square statistics. Each chi-square difference was compared to Bonferroni corrected \(p\)-values (corrected, \(\chi^2 = 16.31\) with 4 degrees of freedom), and items exhibiting DIF were flagged.
Results

Sample Characteristics

As shown in Table 6, Whites, Blacks, and Mexican Americans were significantly different in terms of their age, gender, and educational attainment. The White sample included more individuals who were aged 75 or older: 46.7% for Whites and 32.3% for both the Blacks and Mexican Americans. In all three groups well over half were female (57% for Whites, 63% for Blacks, and 58% for Mexican Americans). Black participants were more likely to be female than Whites and Mexican Americans. In terms of educational attainment, there were more Mexican Americans with less than an eight grade level of education (77%) than was the case for Whites, among whom the majority had more than an 8th grade level of education (40.5% for 8th-11th grade, 20.1% for 12th grade, and 15.8% for more than 12th grade). A study variable, the CES-D, also demonstrated significantly different mean scores across three racial/ethnic groups, showing higher scores in Mexican American sample than in the White or Black samples ($F = 33.82, p < .001$). Moreover, using the standard cutoff of 16 on the CES-D for evidence of probable depression, Mexican Americans consistently exhibited a greater likelihood for probable depression than was reported for the Whites or Blacks. Specifically, 16% of the Whites, 14.4% of the Blacks, and 23.1% of the Mexican Americans fell into the probable depression category ($\chi^2 = 44.17, p < .01$).
Table 6. Study 2 – Descriptive Characteristics of the Sample

<table>
<thead>
<tr>
<th>Variables</th>
<th>New Haven EPESE (N = 2,299)</th>
<th>H-EPESE (N = 2,623)</th>
<th>Mexican Americans (N = 2,623)</th>
<th>F or χ²</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Whites (N = 1,844)</td>
<td>Blacks (N = 455)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Age (≥75)</td>
<td>45.7%</td>
<td>32.3%</td>
<td>32.3%</td>
<td>89.58***</td>
</tr>
<tr>
<td>Female</td>
<td>56.6%</td>
<td>63.4%</td>
<td>58.1%</td>
<td>6.98*</td>
</tr>
<tr>
<td>Educational attainment</td>
<td></td>
<td></td>
<td></td>
<td>1279.26***</td>
</tr>
<tr>
<td>Less than 8th grade</td>
<td>23.6%</td>
<td>50.5%</td>
<td>77.0%</td>
<td></td>
</tr>
<tr>
<td>8th-11th grade</td>
<td>40.5%</td>
<td>31.9%</td>
<td>13.3%</td>
<td></td>
</tr>
<tr>
<td>12th grade</td>
<td>20.1%</td>
<td>13.2%</td>
<td>6.7%</td>
<td></td>
</tr>
<tr>
<td>More than 12th grade</td>
<td>15.8%</td>
<td>4.4%</td>
<td>3.1%</td>
<td></td>
</tr>
<tr>
<td>CES-D Scale</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Mean (SD)</td>
<td>8.03 (8.39)</td>
<td>7.88 (7.82)</td>
<td>10.06 (9.31)</td>
<td>33.82***</td>
</tr>
<tr>
<td>Probable depression (≥ 16)</td>
<td>16.0%</td>
<td>14.4%</td>
<td>23.1%</td>
<td>44.17***</td>
</tr>
<tr>
<td>Reliability (α)</td>
<td>.86</td>
<td>.84</td>
<td>.88</td>
<td></td>
</tr>
</tbody>
</table>

Note. New Haven EPESE = New Haven Established Populations for Epidemiologic Studies of the Elderly; H-EPESE = Hispanic Established Populations for Epidemiologic Studies of the Elderly; SD = Standard Deviation; CES-D = Center for Epidemiological Studies-Depression

* p < .05. *** p < .001.
Dimensionality of the CES-D

Dimensionality was examined in each sociodemographic group by three racial/ethnic groups and by the total sample, making for a total of twenty four dimensionality tests. The PCA results indicated that while the several analyses produced three to six factors, the ratio of first to second eigenvalue ranged from 3.01 (low educated Mexican Americans: 1\textsuperscript{st} factor = 6.90; 2\textsuperscript{nd} factor = 2.29; 3\textsuperscript{rd} factor = 1.14) to 4.63 (high educated Whites: 1\textsuperscript{st} factor = 5.89; 2\textsuperscript{nd} factor = 1.27; 3\textsuperscript{rd} factor = 1.24; 4\textsuperscript{th} factor = 1.12), respectively, which suggests that the data are essentially unidimensional (Lord, 1980; Stout, 1990). This result was supported by one-factor confirmatory factor analyses, where goodness-of-fit indices exceeded .90 (except for two .89 fit indices for low educated Mexican Americans). Among twenty four CFA analyses, high educated Whites showed the best goodness-of-fit indices (CFI = .95, NFI = .94, NNFI = .95), while low educated Mexican Americans showed the lowest goodness-of-fit indices (CFI = .90, NFI = .89, NNFI = .89). These results confirmed the overall unidimensionality of the CES-D for the present analyses.

DIF Results

Results of age, gender, and educational attainment DIF analyses are summarized in Table 7, Table 8 and Table 9, respectively. In both the CFA and IRT methods, this study used Item 1 as a referent and had nineteen model comparisons in which each of the CES-D items was constrained to be equal across groups in each model. Following suggestions by a number of psychometricians (Hambleton, 2006; Hambleton & Rogers, 1989), this study reported DIF items that showed up consistently across the two methods as a more conservative strategy that would reduce possible Type I error rates. Although
jointly identified DIF items were the main focus, attention was also given to DIF items detected by only one method as well as DIF-free items.

As an overview, six of the CES-D items were shown to have statistically significant sociodemographic-related DIF in both CFA and IRT methods. Looking at another way, 70% of the CES-D items (i.e., fourteen items) were relatively free of item bias associated with sociodemographic characteristics. Different effects of sociodemographic characteristics on the CES-D item bias were found in the three racial/ethnic groups. Using the joint CFA/IRT identification criterion, no evidence of item bias by sociodemographic characteristics was observed among Whites. Among Blacks, one item was observed to be biased by educational attainment (Item #17, ‘crying’). Among Mexican Americans, the same item (Item #17, ‘crying’) was biased by gender. Item #17 was the only item confounded with race/ethnicity and sociodemographic characteristics (here, educational attainment and gender).

Age-DIF (Younger than 75 vs. 75 or older)

Table 7 summarizes age DIF results from CFA and IRT. In the comparison of those 65 to 74 years and those 75 and older, no consistent DIF item was detected by both CFA and IRT methods, suggesting no evidence of age-related item bias among older adults. The same results were found in two DIF studies of Black and White elders that used the same New Haven EPESE sample as used in the present study (Cole et al., 2000; Yang & Jones, in press). However, there were six DIF items detected by only one method: CFA found three age-DIF items (Items #7, 10, 16) in the total sample; IRT found two age-DIF items (Items #6, 12) among Whites, one (Item #17) among Blacks,
Table 7. Study 2 – Age DIF Results from CFA and IRT Methods

<table>
<thead>
<tr>
<th>Models</th>
<th>CFA&lt;sup&gt;a&lt;/sup&gt; (Δ df = 2)</th>
<th>IRT&lt;sup&gt;b&lt;/sup&gt; (Δ df = 4)</th>
<th>CFA&lt;sup&gt;a&lt;/sup&gt; (Δ df = 2)</th>
<th>IRT&lt;sup&gt;b&lt;/sup&gt; (Δ df = 4)</th>
<th>CFA&lt;sup&gt;a&lt;/sup&gt; (Δ df = 2)</th>
<th>IRT&lt;sup&gt;b&lt;/sup&gt; (Δ df = 4)</th>
<th>CFA&lt;sup&gt;a&lt;/sup&gt; (Δ df = 2)</th>
<th>IRT&lt;sup&gt;b&lt;/sup&gt; (Δ df = 4)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Total Chi-Square (Difference)</td>
<td>White Chi-Square (Difference)</td>
<td>Black Chi-Square (Difference)</td>
<td>Mexican Americans Chi-Square (Difference)</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Baseline Model (Referent: Item 1)</td>
<td>6452.68</td>
<td>65125.5</td>
<td>1619.35</td>
<td>23247.5</td>
<td>849.32</td>
<td>6885.6</td>
<td>5253.01</td>
<td>41133.0</td>
</tr>
<tr>
<td>Comparison Models</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Item 2, ‘appetite’</td>
<td>(0.65)</td>
<td>(3.1)</td>
<td>(0)</td>
<td>(7.3)</td>
<td>(0.20)</td>
<td>(0.7)</td>
<td>(1.99)</td>
<td>(2.8)</td>
</tr>
<tr>
<td>Item 3, ‘blues’</td>
<td>(6.50)</td>
<td>(2.9)</td>
<td>(0.31)</td>
<td>(7.3)</td>
<td>(1.62)</td>
<td>(2.4)</td>
<td>(6.24)</td>
<td>(3.3)</td>
</tr>
<tr>
<td>Item 4, ‘good’</td>
<td>(7.28)</td>
<td>(14.2)</td>
<td>(1.72)</td>
<td>(8.7)</td>
<td>(10.03)</td>
<td>(9.0)</td>
<td>(9.02)</td>
<td>(12.2)</td>
</tr>
<tr>
<td>Item 5, ‘mind’</td>
<td>(1.22)</td>
<td>(12.1)</td>
<td>(0.02)</td>
<td>(13.9)</td>
<td>(3.29)</td>
<td>(0.4)</td>
<td>(4.89)</td>
<td>(1.9)</td>
</tr>
<tr>
<td>Item 6, ‘depressed’</td>
<td>(3.46)</td>
<td>(1.8)</td>
<td>(0.31)</td>
<td>(16.4)</td>
<td>(8.59)</td>
<td>(2.9)</td>
<td>(4.05)</td>
<td>(5.6)</td>
</tr>
<tr>
<td>Item 7, ‘effort’</td>
<td>(22.40)</td>
<td>(5.8)</td>
<td>(1.40)</td>
<td>(9.2)</td>
<td>(7.59)</td>
<td>(2.7)</td>
<td>(1.86)</td>
<td>(3.7)</td>
</tr>
<tr>
<td>Item 8, ‘hopeful’</td>
<td>(6.85)</td>
<td>(1.0)</td>
<td>(0.37)</td>
<td>(8.1)</td>
<td>(7.66)</td>
<td>(0.9)</td>
<td>(6.46)</td>
<td>(0.7)</td>
</tr>
<tr>
<td>Item 9, ‘failure’</td>
<td>(3.60)</td>
<td>(1.2)</td>
<td>(0.33)</td>
<td>(11.3)</td>
<td>(7.91)</td>
<td>(6.8)</td>
<td>(11.07)</td>
<td>(6.0)</td>
</tr>
<tr>
<td>Item 10, ‘fearful’</td>
<td>(20.95)</td>
<td>(4.6)</td>
<td>(10.97)</td>
<td>(3.6)</td>
<td>(0.62)</td>
<td>(5.6)</td>
<td>(10.83)</td>
<td>(0.7)</td>
</tr>
<tr>
<td>Item 11, ‘sleep’</td>
<td>(1.21)</td>
<td>(5.7)</td>
<td>(1.71)</td>
<td>(2.6)</td>
<td>(2.04)</td>
<td>(0.3)</td>
<td>(0.07)</td>
<td>(11.7)</td>
</tr>
<tr>
<td>Item 12, ‘happy’</td>
<td>(8.99)</td>
<td>(1.8)</td>
<td>(0.20)</td>
<td>(21.9)</td>
<td>(6.33)</td>
<td>(2.2)</td>
<td>(8.42)</td>
<td>(14.7)</td>
</tr>
<tr>
<td>Item 13, ‘talked less’</td>
<td>(0.32)</td>
<td>(1.1)</td>
<td>(0.58)</td>
<td>(5.3)</td>
<td>(1.52)</td>
<td>(4.4)</td>
<td>(0.61)</td>
<td>(11.1)</td>
</tr>
<tr>
<td>Item 14, ‘lonely’</td>
<td>(11.02)</td>
<td>(2.4)</td>
<td>(3.70)</td>
<td>(6.4)</td>
<td>(6.01)</td>
<td>(5.4)</td>
<td>(5.44)</td>
<td>(5.9)</td>
</tr>
<tr>
<td>Item 15, ‘unfriendly’</td>
<td>(1.86)</td>
<td>(3.3)</td>
<td>(2.05)</td>
<td>(3.2)</td>
<td>(3.79)</td>
<td>(6.3)</td>
<td>(0.32)</td>
<td>(12.6)</td>
</tr>
<tr>
<td>Item 16, ‘enjoyed’</td>
<td>(13.44)</td>
<td>(1.9)</td>
<td>(0.53)</td>
<td>(4.5)</td>
<td>(0.17)</td>
<td>(3.8)</td>
<td>(9.75)</td>
<td>(17.4)</td>
</tr>
<tr>
<td>Item 17, ‘crying’</td>
<td>(0.70)</td>
<td>(1.1)</td>
<td>(0.12)</td>
<td>(5.7)</td>
<td>(10.24)</td>
<td>(30.5)</td>
<td>(3.83)</td>
<td>(5.3)</td>
</tr>
<tr>
<td>Item 18, ‘sad’</td>
<td>(7.67)</td>
<td>(8.5)</td>
<td>(0.56)</td>
<td>(10.0)</td>
<td>(6.47)</td>
<td>(1.0)</td>
<td>(7.74)</td>
<td>(2.6)</td>
</tr>
<tr>
<td>Item 19, ‘disliked’</td>
<td>(0)</td>
<td>(2.7)</td>
<td>(1.58)</td>
<td>(2.3)</td>
<td>(2.11)</td>
<td>(3.1)</td>
<td>(1.65)</td>
<td>(8.2)</td>
</tr>
<tr>
<td>Item 20, ‘get going’</td>
<td>(5.62)</td>
<td>(9.6)</td>
<td>(0.91)</td>
<td>(1.8)</td>
<td>(3.12)</td>
<td>(5.9)</td>
<td>(6.81)</td>
<td>(1.3)</td>
</tr>
</tbody>
</table>

Note. Items in bold and underlined are common DIF items across CFA and IRT methods in each cross-racial/ethnic comparison.

<sup>a</sup> Reverse-coded item.  <sup>b</sup> In CFA, DIF flagged if \( \Delta \chi^2 \) were > 11.88.  <sup>c</sup> In IRT, DIF flagged if \( \Delta \chi^2 \) were > 16.31.
and one (Item #16) among Mexican Americans, men were more likely than women to endorse the ‘crying’ item.

*Gender-DIF (Male vs. Female)*

As shown in Table 8, two gender biased items (Items #15, ‘people were unfriendly’ and #17, ‘I had crying spells’) were identified by both CFA and IRT. Compared to women, men showed a greater tendency to endorse the ‘unfriendly’ item at all levels of depression. Women had a greater propensity than men to endorse ‘crying’ item at lower to mid-higher depressions scores (theta = -3 to +2.4). However, at more severe levels of depressive symptoms (theta ≥ +2.5), men were more likely than women to endorse the ‘crying’ item.

Gender was confounded with Mexican American race/ethnicity on the ‘crying’ item. As shown in Figure 3, compared to Mexican American men, women were more likely to report the ‘crying’ item at higher levels of depressive symptoms (theta ≥ +.7). However, no gender differences on the ‘crying’ item were observed at lower levels of depressive symptoms (theta < +.7) as shown in Figure 3. This finding indicates that with crying spells, greater gender DIF occurred at higher levels of depression. When compared to the total female sample, Mexican American women were more likely to endorse the ‘crying’ item at higher levels of depressive symptoms (theta ≥ +1.2), whereas at lower levels of depressive symptoms (theta < 1.2), they were less likely to endorse the ‘crying’ item. Mexican American men showed a response pattern similar to that of men in the total sample, although their scores on the ‘crying’ item were higher than that of the total male sample across all levels of depression.
Table 8. Study 2 – Gender DIF Results from CFA and IRT Methods

<table>
<thead>
<tr>
<th>Models</th>
<th>Total</th>
<th>Gender DIF (Male vs. Female)</th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>CFA(^a)</td>
<td>IRT(^b)</td>
<td>CFA(^a)</td>
<td>IRT(^b)</td>
<td>CFA(^a)</td>
</tr>
<tr>
<td></td>
<td></td>
<td>((\Delta df = 2))</td>
<td>((\Delta df = 4))</td>
<td>((\Delta df = 2))</td>
<td>((\Delta df = 4))</td>
<td>((\Delta df = 2))</td>
</tr>
<tr>
<td>Baseline Model</td>
<td>(Referent: Item1)</td>
<td>Chi-Square (Difference)</td>
<td>Chi-Square (Difference)</td>
<td>Chi-Square (Difference)</td>
<td>Chi-Square (Difference)</td>
<td>Chi-Square (Difference)</td>
</tr>
<tr>
<td>Item 2, ‘appetite’</td>
<td></td>
<td>(1.66)</td>
<td>(7.4)</td>
<td>(2.26)</td>
<td>(1.0)</td>
<td>(0.58)</td>
</tr>
<tr>
<td>Item 3, ‘blues’</td>
<td></td>
<td>(0.25)</td>
<td>(1.9)</td>
<td>(3.09)</td>
<td>(0)</td>
<td>(0.01)</td>
</tr>
<tr>
<td>Item 4, ‘good’</td>
<td></td>
<td>(0.62)</td>
<td>(5.0)</td>
<td>(2.87)</td>
<td>(2.6)</td>
<td>(5.17)</td>
</tr>
<tr>
<td>Item 5, ‘mind’</td>
<td></td>
<td>(2.20)</td>
<td>(2.1)</td>
<td>(2.86)</td>
<td>(3.3)</td>
<td>(4.29)</td>
</tr>
<tr>
<td>Item 6, ‘depressed’</td>
<td></td>
<td>(0.16)</td>
<td>(3.1)</td>
<td>(3.96)</td>
<td>(11.0)</td>
<td>(0.27)</td>
</tr>
<tr>
<td>Item 7, ‘effort’</td>
<td></td>
<td>(0.02)</td>
<td>(11.9)</td>
<td>(1.74)</td>
<td>(5.2)</td>
<td>(0.53)</td>
</tr>
<tr>
<td>Item 8, ‘hopeful’</td>
<td></td>
<td>(4.58)</td>
<td>(7.6)</td>
<td>(3.18)</td>
<td>(5.6)</td>
<td>(0.04)</td>
</tr>
<tr>
<td>Item 9, ‘failure’</td>
<td></td>
<td>(1.41)</td>
<td>(7.1)</td>
<td>(4.49)</td>
<td>(6.1)</td>
<td>(0.05)</td>
</tr>
<tr>
<td>Item 10, ‘fearful’</td>
<td></td>
<td>(4.33)</td>
<td>(2.8)</td>
<td>(2.16)</td>
<td>(3.9)</td>
<td>(0.27)</td>
</tr>
<tr>
<td>Item 11, ‘sleep’</td>
<td></td>
<td>(0.33)</td>
<td>(6.1)</td>
<td>(4.58)</td>
<td>(1.8)</td>
<td>(3.47)</td>
</tr>
<tr>
<td>Item 12, ‘happy’</td>
<td></td>
<td>(1.12)</td>
<td>(1.8)</td>
<td>(8.32)</td>
<td>(2.0)</td>
<td>(4.49)</td>
</tr>
<tr>
<td>Item 13, ‘talked less’</td>
<td></td>
<td>(1.69)</td>
<td>(6.1)</td>
<td>(1.81)</td>
<td>(6.3)</td>
<td>(0.27)</td>
</tr>
<tr>
<td>Item 14, ‘lonely’</td>
<td></td>
<td>(0.11)</td>
<td>(4.9)</td>
<td>(7.24)</td>
<td>(5.6)</td>
<td>(0.89)</td>
</tr>
<tr>
<td>Item 15, ‘unfriendly’</td>
<td></td>
<td>(27.16)(^{\text{DIFF}})</td>
<td>(17.0)(^{\text{DIFF}})</td>
<td>(10.69)</td>
<td>(8.1)</td>
<td>(1.05)</td>
</tr>
<tr>
<td>Item 16, ‘enjoyed’</td>
<td></td>
<td>(1.85)</td>
<td>(10.5)</td>
<td>(6.61)</td>
<td>(2.1)</td>
<td>(0.76)</td>
</tr>
<tr>
<td>Item 17, ‘crying’</td>
<td></td>
<td>(27.77)(^{\text{DIFF}})</td>
<td>(32.0)(^{\text{DIFF}})</td>
<td>(9.32)</td>
<td>(10.3)</td>
<td>(4.49)</td>
</tr>
<tr>
<td>Item 18, ‘sad’</td>
<td></td>
<td>(4.71)</td>
<td>(3.7)</td>
<td>(9.50)</td>
<td>(0.8)</td>
<td>(0.07)</td>
</tr>
<tr>
<td>Item 19, ‘disliked’</td>
<td></td>
<td>(0.77)</td>
<td>(6.7)</td>
<td>(3.68)</td>
<td>(4.5)</td>
<td>(0.27)</td>
</tr>
<tr>
<td>Item 20, ‘get going’</td>
<td></td>
<td>(2.56)</td>
<td>(0.3)</td>
<td>(1.06)</td>
<td>(2.0)</td>
<td>(0.24)</td>
</tr>
<tr>
<td>Total # DIF Items</td>
<td></td>
<td>2</td>
<td>2</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
</tbody>
</table>

Note. Items in bold and underlined are common DIF items across CFA and IRT methods in each cross-racial/ethnic comparison.

\(^{a}\) Reverse-coded item. \(^{a}\) In CFA, DIF flagged if \(\Delta \chi^2\) were > 11.88. \(^{b}\) In IRT, DIF flagged if \(\Delta \chi^2\) were > 16.31.
Educational attainment DIF results are summarized in Table 9. In the comparison of those with less than an 8th grade education and 8th grade or more, five DIF items (Items #8, ‘hopeful,’ #12, ‘happy,’ #16, ‘enjoyed,’ #17, ‘crying,’ and #19, ‘people disliked me’) were identified in both CFA and IRT methods. Overall, those with lower educational attainment had greater propensities to endorse all five DIF items (recall that Items #8, 12, and 16 had been reverse-coded, so greater propensity to endorse these three items represents greater propensity to report low positive affect) compared to those with higher educational attainment. However, different response patterns between the two groups were not observed at lower levels of depressive symptoms for all five DIF items. Three low positive affect items (Items #8, ‘hopeful,’ #12, ‘happy,’ and #16, ‘enjoyed,’)

Figure 3. Item information function for Item 17 (‘I had crying spells’) showing Gender-DIF in the total sample and Mexican American race/ethnicity.

Educational Attainment-DIF (Less than 8th grade vs. 8th grade or more)

Educational attainment DIF results are summarized in Table 9. In the comparison of those with less than an 8th grade education and 8th grade or more, five DIF items (Items #8, ‘hopeful,’ #12, ‘happy,’ #16, ‘enjoyed,’ #17, ‘crying,’ and #19, ‘people disliked me’) were identified in both CFA and IRT methods. Overall, those with lower educational attainment had greater propensities to endorse all five DIF items (recall that Items #8, 12, and 16 had been reverse-coded, so greater propensity to endorse these three items represents greater propensity to report low positive affect) compared to those with higher educational attainment. However, different response patterns between the two groups were not observed at lower levels of depressive symptoms for all five DIF items. Three low positive affect items (Items #8, ‘hopeful,’ #12, ‘happy,’ and #16, ‘enjoyed,’)

90
<table>
<thead>
<tr>
<th>Models</th>
<th>Total</th>
<th>Whites</th>
<th>Blacks</th>
<th>Mexican Americans</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>CFA&lt;sup&gt;a&lt;/sup&gt;</td>
<td>IRT&lt;sup&gt;b&lt;/sup&gt;</td>
<td>CFA&lt;sup&gt;a&lt;/sup&gt;</td>
<td>IRT&lt;sup&gt;b&lt;/sup&gt;</td>
</tr>
<tr>
<td></td>
<td>(Δ df = 2)</td>
<td>(Δ df = 4)</td>
<td>(Δ df = 2)</td>
<td>(Δ df = 4)</td>
</tr>
<tr>
<td></td>
<td>Chi-Square (Difference)</td>
<td>Chi-Square (Difference)</td>
<td>Chi-Square (Difference)</td>
<td>Chi-Square (Difference)</td>
</tr>
<tr>
<td>Baseline Model (Referent: Item1)</td>
<td>6359.75</td>
<td>64037.9</td>
<td>1623.15</td>
<td>22222.1</td>
</tr>
<tr>
<td>Comparison Models</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Item 2, ‘appetite’</td>
<td>(1.65)</td>
<td>(26.0)&lt;sup&gt;DIF&lt;/sup&gt;</td>
<td>(0.96)</td>
<td>(2.3)</td>
</tr>
<tr>
<td>Item 3, ‘blues’</td>
<td>(5.72)</td>
<td>(16.6)&lt;sup&gt;DIF&lt;/sup&gt;</td>
<td>(2.84)</td>
<td>(7.6)</td>
</tr>
<tr>
<td>Item 4, ‘good’</td>
<td>(4.24)</td>
<td>(211.9)&lt;sup&gt;DIF&lt;/sup&gt;</td>
<td>(2.03)</td>
<td>(18.6)&lt;sup&gt;DIF&lt;/sup&gt;</td>
</tr>
<tr>
<td>Item 5, ‘mind’</td>
<td>(6.91)</td>
<td>(25.6)&lt;sup&gt;DIF&lt;/sup&gt;</td>
<td>(0.07)</td>
<td>(17.0)&lt;sup&gt;DIF&lt;/sup&gt;</td>
</tr>
<tr>
<td>Item 6, ‘depressed’</td>
<td>(0.14)</td>
<td>(0.8)</td>
<td>(0.03)</td>
<td>(13.0)</td>
</tr>
<tr>
<td>Item 7, ‘effort’</td>
<td>(0.49)</td>
<td>(41.0)&lt;sup&gt;DIF&lt;/sup&gt;</td>
<td>(0.20)</td>
<td>(13.0)</td>
</tr>
<tr>
<td>Item 8, ‘hopeful’</td>
<td>(24.56)&lt;sup&gt;DIF&lt;/sup&gt;</td>
<td>(98.3)&lt;sup&gt;DIF&lt;/sup&gt;</td>
<td>(0.20)</td>
<td>(4.7)</td>
</tr>
<tr>
<td>Item 9, ‘failure’</td>
<td>(6.76)</td>
<td>(27.8)&lt;sup&gt;DIF&lt;/sup&gt;</td>
<td>(0.74)</td>
<td>(6.6)</td>
</tr>
<tr>
<td>Item 10, ‘fearful’</td>
<td>(1.60)</td>
<td>(9.4)</td>
<td>(4.06)</td>
<td>(2.9)</td>
</tr>
<tr>
<td>Item 11, ‘sleep’</td>
<td>(0.19)</td>
<td>(19.4)&lt;sup&gt;DIF&lt;/sup&gt;</td>
<td>(0.02)</td>
<td>(1.7)</td>
</tr>
<tr>
<td>Item 12, ‘happy’</td>
<td>(21.26)&lt;sup&gt;DIF&lt;/sup&gt;</td>
<td>(94.1)&lt;sup&gt;DIF&lt;/sup&gt;</td>
<td>(0.01)</td>
<td>(2.8)</td>
</tr>
<tr>
<td>Item 13, ‘talked less’</td>
<td>(3.01)</td>
<td>(14.0)</td>
<td>(1.02)</td>
<td>(6.4)</td>
</tr>
<tr>
<td>Item 14, ‘lonely’</td>
<td>(0)</td>
<td>(6.6)</td>
<td>(2.10)</td>
<td>(3.7)</td>
</tr>
<tr>
<td>Item 15, ‘unfriendly’</td>
<td>(19.03)&lt;sup&gt;DIF&lt;/sup&gt;</td>
<td>(14.2)</td>
<td>(9.54)</td>
<td>(13.9)</td>
</tr>
<tr>
<td>Item 16, ‘enjoyed’</td>
<td>(18.04)&lt;sup&gt;DIF&lt;/sup&gt;</td>
<td>(185.6)&lt;sup&gt;DIF&lt;/sup&gt;</td>
<td>(0.05)</td>
<td>(3.1)</td>
</tr>
<tr>
<td>Item 17, ‘crying’</td>
<td>(42.06)&lt;sup&gt;DIF&lt;/sup&gt;</td>
<td>(60.8)&lt;sup&gt;DIF&lt;/sup&gt;</td>
<td>(7.58)</td>
<td>(7.6)</td>
</tr>
<tr>
<td>Item 18, ‘sad’</td>
<td>(1.60)</td>
<td>(14.7)</td>
<td>(1.04)</td>
<td>(2.5)</td>
</tr>
<tr>
<td>Item 19, ‘disliked’</td>
<td>(53.59)&lt;sup&gt;DIF&lt;/sup&gt;</td>
<td>(48.2)&lt;sup&gt;DIF&lt;/sup&gt;</td>
<td>(9.80)</td>
<td>(11.6)</td>
</tr>
<tr>
<td>Item 20, ‘get going’</td>
<td>(1.98)</td>
<td>(30.3)&lt;sup&gt;DIF&lt;/sup&gt;</td>
<td>(0.86)</td>
<td>(1.8)</td>
</tr>
<tr>
<td>Total # DIF Items</td>
<td>6</td>
<td>13</td>
<td>0</td>
<td>2</td>
</tr>
</tbody>
</table>

<sup>Note</sup>. Items in bold and underlined are common DIF items across CFA and IRT methods in each cross-racial/ethnic comparison.  
<sup>a</sup> Reverse-coded item.  
<sup>b</sup> In CFA, DIF flagged if Δχ² were > 11.88.  
<sup>c</sup> In IRT, DIF flagged if ΔΔχ² were > 16.31.
favored those with lower educational attainment, indicating it was harder for lower educated people to respond to the positive side of these items.

Figure 4. Item information function for Item 17 (‘I had crying spells’) showing Education-DIF in the total sample and Black race/ethnicity.

Educational attainment was confounded with Black race/ethnicity on the ‘crying’ item. As shown in Figure 4, low educated Blacks were more likely than high educated Blacks to endorse the ‘crying’ item at lower to fairly high levels of depressive symptoms (theta = -3 to +2.1). Interestingly, however, at the highest levels of depression scores (theta ≥ +2.2), Blacks with higher educational attainment showed a greater tendency to endorse the ‘crying’ item than Blacks with lower educational attainment. Similar response patterns were observed among total lower and higher education groups: those
with low educational attainment were generally more likely than those with high educational attainment to endorse the ‘crying’ item (theta = -3 to +2.7), whereas at very severe levels of depression (theta ≥ +2.8), those with high educational attainment were more likely than those with low educational attainment to endorse the ‘crying’ item. Compared to the total sample, both low and high educated Blacks were less likely to respond the ‘crying’ item across all levels of depressive symptom scores.

Discussion

With the increasing recognition of the importance of understanding mental health not only across racial/ethnic groups, but within such groups, has come a corresponding recognition of the importance of investigating the measurement equivalence of popular screening tools such as the CES-D. Few if any studies have examined item bias in the CES-D across sociodemographic strata in racially/ethnically diverse elderly populations. This Study 2 evaluated sociodemographic-related measurement equivalence of the CES-D in three racial/ethnic elderly groups drawn from two national datasets, the New Haven and Hispanic EPESE. In addition, given the importance of using multiple analytic strategies for measurement equivalence research (Hambleton, 2006; Wang & Russell, 2005), the present dissertation study used two different analytic methods, CFA and IRT, to detect item bias on the CES-D more accurately.

Across all sociodemographic subgroups of age, gender, educational attainment, six of the twenty CES-D items revealed statistically significant DIF in both CFA and IRT methods: two showed gender bias (‘people were unfriendly’ and ‘crying spells’) and five showed bias associated with educational attainment (‘hopeful,’ ‘happy,’ ‘enjoyed,’ ‘crying spells,’ and ‘people disliked me’). Notably, the ‘crying’ item was not only biased
by both gender and educational attainment cases, but there were also interactions between race/ethnicity and both gender and educational attainment. This clearly suggests that ‘crying’ was the most biased item across all the sociodemographic strata in diverse racial/ethnic elderly groups. The six DIF items included the two interpersonal relation items, three of the four positive feeling items, and one depressive affect item. When using both CFA and IRT methods, no evidence of age DIF was observed in the present samples, all of which included only persons aged 65 and over.

With regard to gender bias in the ‘crying spells’ item, the present finding supported previous item bias studies on the CES-D (e.g., Cole et al., 2000; Stommel et al., 1993) showing a greater propensity to endorse the ‘crying spells’ item among females. This can be explained by the concept of women’s permission to cry in society (e.g., Hammen & Padeskym, 1977), which suggests that crying may be viewed as socially acceptable behavior for women in many circumstances. Similarly, it also has been suggested that greater tendencies to report mental health issues among women are consistent with socialization patterns that allow women to report discomfort more readily and to appear less stoical than men (Mechanic, 1978; Nathanson, 1975).

One novel finding related to the ‘crying spells’ item was that at the highest level of depressive symptoms (here, at 2.5 or higher theta), men were more likely than women to endorse this item. Unlike previous studies that has only identified women’s higher endorsement on this item (e.g., Cole et al., 2000; Stommel et al., 1993), the unique response pattern reported here clearly indicates that men were more likely to cry than women when their depressive symptoms were severe. At lower to moderately high depression scores (here, at theta ranged -3 to 2.4), men had a tendency to inhibit their
reporting of the ‘crying spells,’ perhaps as a result of social desirability. A number of previous studies have noted that crying does not indicate depressed mood for men (e.g., Ross & Mirowsky, 1984). This finding suggests that this latter conclusion should be revisited and that further investigation on the ‘crying’ item with different samples is warranted.

No clear evidence of an age bias in the CES-D items was found in the present study, in agreement with two published DIF studies on the CES-D items among older adults (Cole et al., 2000; Yang & Jones, in press). As suggested by Cole and colleagues (2000), this study also suspect that the absence of an age bias may result either from an actual lack of bias due to age or my use of a restricted age range. Given the fact that the present DIF analyses showed six age biased items in the CES-D that were detected by one, but not both, of the two DIF methods, the results suggest future research should include younger or middle-aged adults for comparisons with older adults.

The crucial role of educational attainment on the CES-D items was identified in this dissertation study of older adults. This may be the first study to evaluate educational bias in the CES-D items among racially/ethnically diverse older adults. Five items (three reverse-coded positive affect items – ‘hopeful,’ ‘happy,’ ‘enjoyed,’ one depressive affect item – ‘crying,’ and one interpersonal problem item – ‘people disliked me’) revealed consistently higher endorsement by the lower educated group. Educational attainment may be also a proxy for more fundamental differences that lead to DIF (McHorney & Fleishman, 2006). That is, higher educational attainment generally provides more resources and choices (Krause, 2007) and generally provides more opportunities for experiences that promote general well-being (Mirowsky & Ross, 2003). It may be
therefore possible for individuals with higher levels of education to express themselves in a more positive way, which in turn may be associated with their lower endorsement of depressive symptom items.

One of the most intriguing findings was that highly educated individuals had a greater propensity to endorse positive affect items (‘hopeful,’ ‘happy,’ and ‘enjoyed,’). In other words, compared to those with lower educational attainment, it was easier for those with higher educational attainment to report such positive feelings as feeling good about themselves, feeling hopeful about the future, and feeling happy. This is compatible with research suggesting that higher educational level is associated with greater self-expressiveness (Krause, 2007), which is connected to a greater positive affect and greater satisfaction with life (Bettencourt & Sheldon, 2001). This may eventually lead to lower endorsement of depressive symptoms.

It should be emphasized that the ‘crying’ item was the only one biased in the subgroups divided by gender and by educational attainment, as well as the only one confounded with race/ethnicity. Results indicated two interaction effects: 1) gender and Mexican American race/ethnicity and 2) educational attainment and Black race/ethnicity. In essence, Mexican American women and lower educated Blacks had greater propensities to endorse the ‘crying’ item compared with their counterparts (Mexican American men and higher educated Blacks, respectively).

In terms of the interaction effect of gender, Mexican American women were more likely than Mexican American men to endorse the ‘crying spells’ item, especially at higher levels of depressive symptom scores (here, at .7 or higher theta). The abovementioned greater propensity to report ‘crying’ item among female respondents
was clearly confounded with Mexican American culture, where for women crying may be a more acceptable behavior than it is for Whites and Blacks (Azocar, Areán, Miranda, & Muñoz, 2001; Golding, Aneshensel, & Hough, 1991). Compared with other White and Blacks women, Mexican American women showed the highest score on the ‘crying spells’ item only at higher depressive symptom scores (at theta 1.2 or higher).

With regard to the interaction between educational level and Black race/ethnicity, less educated Blacks in general were more likely than more educated Blacks to endorse the ‘crying’ item at lower to fairly high levels of depression. This finding also shows the effect of educational attainment on self-expressiveness among lower educated Blacks, suggesting the connections between a lower self-expressiveness and a lower positive affect, which in turn may lead to higher levels of depressive symptoms (Bettencourt & Sheldon, 2001; Krause, 2007).

Another interesting response pattern was observed in this interaction effect: Blacks with higher educational attainment showed a greater propensity to express their crying spells when they had severe depressive symptoms (at theta 2.2 or higher). However, when their levels of depression were not severe, higher educated Blacks appeared to inhibit their expression or at least the reporting of the ‘crying spells’ item. It is also worth mentioning that Blacks of both lower and higher education groups had the lowest scores on the ‘crying’ item compared with all other groups, which clearly suggests Blacks are underreporting on this item. Some survey researchers have also found Blacks to underreport socially stigmatizing behaviors and viewed this response pattern as a part of displaying their social desirability (e.g., Johnson & van de Vijver, 2003).
Noteworthily, both of the interpersonal relation items in the CES-D showed bias: ‘people were unfriendly’ was gender biased and ‘people dislike me’ was education biased. Compared with women and people with higher educational attainment, men and people with lower educational attainment had greater propensities to endorse both interpersonal problem items across all levels of respondents’ depressive symptoms. This may reflect self-perceptions of discrimination experienced by those with male gender and lower educational attainment. In fact, there is some evidence that groups with lower power, as reflected by lower educational attainment, are likely to view the world as chaotic and catastrophic and to distrust the world outside family and friends (e.g., Briones et al., 1990; Hoppe & Heller, 1975), which may in turn lead to their self-perceptions of discrimination against them in everyday life. In fact, the two interpersonal problem items have revealed race/ethnicity-related item bias among Mexican Americans (e.g., Kim et al., 2007) as well as Blacks (e.g., Blazer et al., 1998; Cole et al., 2000).

From a methodological standpoint, using multiple analytic strategies to detect DIF was of particular interest in this dissertation study. Following the suggestion of researchers (e.g., Hambleton, 2006; Hambleton & Rogers, 1989; Wang & Russell, 2005), this dissertation study applied two of the most common DIF methods, CFA and IRT. DIF items that showed up consistently across the two procedures were reported in order to render conservative conclusions as to which items showed DIF and which ones did not. In addition, this study followed a unified strategy of DIF detection suggested by Stark and colleagues (2006), which can provide higher power and lower Type I error rates. CFA and IRT methods yielded similar results, although CFA identified more DIF items than did IRT. The results from CFA and IRT gave us a great amount of information on
sociodemographic-related DIF on the CES-D, which also provided methodological and practical implications for future research. However, from a practical point of view, questions still remain as to how to deal with DIF once it is identified, as well as how to interpret uncommonly detected DIF items. Clearly, more work needs to be done in health disparities and measurement research to develop clear guidelines to deal with DIF, such as removing consistent DIF items and adjusting cut-scores.

Overall the results from Study 2 suggest that when self-report instruments are used to screen and assess for depression in diverse sociodemographic populations, researchers and practitioners should be aware of the risk that individuals from different sociodemographic and cultural backgrounds may tend to be misclassified, which can directly lead to misdiagnosis as well as mistreatment for depression. Use of inaccurate measures could also lead to misguided public policies. Therefore, in light of the abovementioned consequences of using nonequivalent measures, researchers should pay careful attention to making measures more reliable and socioculturally appropriate, as well as to establishing measurement equivalence of the existing depression measures. The latter is the first and crucial step before diverse sociodemographic and racial/ethnic groups can be compared.
CHAPTER FOUR: STUDY 3

ACCULTURATION- AND LANGUAGE-RELATED MEASUREMENT BIAS

Introduction

Study 3 focuses on Mexican Americans, the largest subgroup of Hispanics in the United States. Mexican Americans are themselves a culturally diverse group. A number of researchers have shown that Mexican Americans have different characteristics depending on their place of birth (e.g., Chiriboga, 2004; Sundquist & Winkleby, 2000), socioeconomic status (e.g., Krause & Markides, 1985), and the level of acculturation (e.g., Chiriboga, Jang, Banks, & Kim, 2007; González, Haan, & Hinton, 2001). Of particular interest in the present investigation are differences in levels of acculturation because acculturation is considered one of the key dimensions for understanding health disparities in diverse populations (Stewart & Nápoles-Springer, 2003).

A dynamic and ongoing cultural process, acculturation has been referred to as the degree to which people change when faced with the challenge of living in a cultural context differing from their own (Trimble, 2003). Studies examining the link between level of acculturation and health outcomes have shown mixed results, with some reporting positive connections (e.g., Berry & Kim, 1989; Chiriboga, Black, Aranda, & Markides, 2002; González et al., 2001) and some reporting negative (e.g., Krause & Goldenhar, 1992; Sundquist & Winkleby, 2000). Despite the mixed results, what most health disparities researchers agree upon is that the level of acculturation may influence people’s life styles, behaviors, attitudes, and general experiences to the host culture (e.g.,
Stewart & Nápoles-Springer, 2003). Therefore, acculturation may be an important construct that can explain social and cultural differences within Mexican Americans.

With respect to the last point, Mexican Americans who are more acculturated may be more likely to adopt ways of thinking and feeling that characterize the host culture. In contrast, Mexican Americans who are less acculturated may be less able to accept or adopt the new ways of thinking and expressing themselves, and instead hold onto the values and behaviors that reflect their culture of origin. These differences in acculturation may influence the ways in which psychological symptoms are organized and expressed, as well as holding implications for the levels of measurement bias evident in mental health instruments such as screening tools for depressive symptoms in here.

The most frequently used depression screening tool in the United States is the Center for Epidemiological Studies – Depression Scale (CES-D; Radloff, 1977). Originally developed for European American populations, good internal consistency and general usefulness to assess depressive symptoms have been observed when the CES-D is applied to Mexican Americans (e.g., González et al., 2001; Liang, Van Tran, Krause, & Markides, 1989). The Spanish version of the CES-D has been also widely applied to Mexican Americans and shows good comparability to the English version of the CES-D (e.g., Perczek, Caver, Price, & Pozo-Kaderman, 2000). However, one missing piece of psychometric information is whether item responses to the CES-D among Mexican Americans are systematically influenced by factors, such as level of acculturation and language of assessment that are unrelated to depression. If the items in the CES-D do not function equivalently across subgroups of Mexican Americans, the CES-D may fail to capture depressive symptoms in certain subgroups of Mexican Americans. Under this
condition, estimates of prevalence may be inaccurate and subgroup comparisons within Mexican Americans may be misleading.

The effect of acculturation on the measurement equivalence of the CES-D has not been fully determined in samples of older Mexican Americans, or indeed any other group of Hispanics. One study of older Mexican Americans found that the pattern of factor loadings in the CES-D was different in high and low acculturated groups, suggesting an association between the level of acculturation and item endorsement (Chiriboga et al., 2007). To date only a single study has used an appropriate differential item functioning (DIF) method to examine item bias. Using a sample of pregnant Hispanic women, Nguyen and colleagues (2007) found that responses to the CES-D differed by acculturation and that the low acculturated group was less likely to endorse somatic symptoms but more likely to endorse positive items than the acculturated group. In their study, the term acculturation was measured with respondents’ language preference, and the total sample was divided into two acculturation groups: people who preferred English were considered acculturated and those who preferred Spanish were considered unacculturated. Although language of preference has been used a proxy for acculturation (e.g., Cabassa, 2003; Zane & Mak, 2003), the latter is in fact a complex construct that include more than simple language ability. In addition, there has been research showing that feelings reported in one’s primary language may be expressed with more emotion than those expressed in a second language (Cuellar & Roberts, 1984; Roberts, Vernon, & Rhoades, 1989). None of the existing studies have paid attention to the effects of both the level of acculturation and language on the item bias in the CES-D.
An important issue for the present study of measurement equivalence was to find a methodology that can distinguish a lack of measurement equivalence (i.e., DIF) from true differences in the trait distributions (i.e., impact) more accurately. Researchers have used two popular methods to detect DIF: confirmatory factor analysis (CFA) and item response theory (IRT). Most studies have used only one of these two methods to identify DIF items. Only a few simulation studies have compared these two DIF methods (e.g., Meade & Lautenschlager, 2004; Raju, Laffitte, & Byrne, 2002).

For more accurate results and firmly-grounded conclusions as to which items are classified as DIF, a number of researchers recommended using multiple DIF analytic strategies (e.g., Hambleton, 2006; Wang & Russell, 2005). In addition, one suggested strategy to reduce Type I error rates in DIF results has been to focus on the items that show up consistently across different DIF methods (Hambleton, 2006). Recently, Stark, Chernyshenko, and Drasgow (2006) proposed and tested a common strategy for detecting DIF with CFA and IRT based on the likelihood ratio (LR) test. Their method, which involved comparing statistically correct free-baseline models with a series of constrained models that simultaneously examined item loadings (for CFA)/discrimination parameters (for IRT) and intercepts (for CFA)/location parameters (for IRT), showed higher power and low Type I error rates across simulation conditions for both CFA and IRT. A recent study using this common strategy found it to be a useful approach for detecting DIF items (e.g., Kim, Chiriboga, & Jang, 2007).

One of my foci in the present dissertation study is to use the Stark and colleagues’ (2006) common strategy and compare the results across both CFA and IRT. The intent was to identify the effects of acculturation and instrument language on the measurement
equivalence of the CES-D in older Mexican Americans. It was expected that item responses to the CES-D would differ depending on the level of acculturation and language used in responding to the questions.

Methods

Sample

Data were drawn from the first wave of the Hispanic Established Populations for Epidemiologic Studies of the Elderly (H-EPESE). The H-EPESE is a longitudinal study of Mexican Americans aged 65 and older who live in Texas, New Mexico, Colorado, Arizona, and California. In 1993-4, baseline interviews were conducted with 3,050 subjects using English and Spanish versions of the interview. Subjects were included in the analyses if they responded to all 20 CES-D items (N = 2,623). From this listwise deletion, 427 participants (14.0 % of the total sample) were excluded. Those excluded had similar characteristics to those included in terms of gender distribution and educational attainment, but were more likely to be older ($t = 9.42, p < .001$).

Measures

CES-D scale

The 20 item Center for Epidemiological Studies-Depression Scale (CES-D; Radloff, 1977) was used in this study and contains sixteen negative items and four positive items. The response format was a 4-point Likert scale, with categories presented in the following order: “rarely or none of the time (coded as 0),” “some or a little of the time (coded as 1),” “occasionally (coded as 2),” and “most or all of the time (coded as 3).” The four positive items (Items #4, 8, 12, and 16) was reverse-scored and scale scores
were computed by summing across twenty items to produce total scores ranging from 0 (no depressive symptoms) to 60 (severe depressive symptoms). Scores of 16 or higher are typically viewed as evidence of probable depression (Andresen, Carter, Malmgren, & Patrick, 1994). A measure with this cutpoint was calculated and reliability in the present sample was satisfactory: $\alpha = .88$.

**Acculturation**

Level of acculturation was measured with thirteen items drawn from Hazuda et al. (1988) and Cuellar, Harris, and Jasso (1980). The thirteen items assessed linguistic acculturation including the self-reported ability to read, write, and understand English, and language used in conversations with family (spouse, children, and parents), friends, neighbors, and coworkers. Previous research also noted that linguistic acculturation has accounted for a substantial portion of acculturation status (e.g., Chiriboga et al., 2007; Jang, Kim, Chiriboga, & King-Kallimanis, 2007). Since several items had relatively high levels of missing data as a result of vacated or never-occupied roles (e.g., language currently spoken with parents – in most cases the parent was deceased), imputation via the Solas statistical program (Statistical Solutions, 2001) was used as a data substitution method. Detailed information on the imputation technique is described in papers by Chiriboga (2004) and Chiriboga et al. (2007). Principal component analysis with oblimin and varimax rotation yielded one factor, and internal consistency based on the thirteen items was shown to be good ($\alpha = .98$). Due to the special interest in the effects of level of acculturation, the total sample was divided into two acculturation groups using a median score on the linguistic ability factor: low acculturated Mexican Americans ($N = 1,283$) and high acculturated Mexican Americans ($N = 1340$).
**Instrument Language**

All participants in the H-EPESE study were given the choice of being interviewed in English or Spanish. Approximately 78% of older Mexican Americans chose to be interviewed in Spanish. For the purpose of this study, the total sample was divided into two instrument language groups: Mexican Americans interviewed in English (N = 561) and Spanish (N = 2,062).

**Acculturation × Language**

To determine whether item bias in the CES-D was more likely to be associated with acculturation or language, Mexican Americans were divided into four acculturation (high/low) × language (English/Spanish) groups: high acculturated Mexican Americans interviewed in English (i.e., High-English: N = 503); high acculturated Mexican Americans interviewed in Spanish (i.e., High-Spanish: N = 837); low acculturated Mexican Americans interviewed in English (i.e., Low-English: N = 58); and low acculturated Mexican Americans interviewed in English (i.e., Low-English: N = 1,225). The Low-English group was excluded in the analyses due to the small sample size.

**Statistical Analysis**

**Dimensionality test**

Since the DIF detection methods used in this investigation assume the unidimensionality of the CES-D (Stark et al., 2006), the dimensionality was evaluated first using a principal component analysis (PCA) via SPSS and confirmatory factor analysis (CFA) via LISREL 8.8. This was done within each acculturation (high vs. low), language (English vs. Spanish), and acculturation x language (high-English vs. high-Spanish vs. low-Spanish) group. The PCA results indicated that while the analyses did in
fact produce either three or four factors, the ratio of first to second eigenvalue ranged
from 3.00 (high acculturated Mexican Americans interviewed in English: 1st factor = 6.36, % variance = 31.79; 2nd factor = 2.13, % variance = 10.74; 3rd factor = 1.17, %
variance = 5.88) to 3.19 (low acculturated Mexican Americans: 1st factor = 7.23, %
variance = 36.15; 2nd factor = 2.27, % variance = 11.38; 3rd factor = 1.07, % variance =
5.38), which suggests that the data are essentially unidimensional (Lord, 1980; Stout,
1990). This result was supported by one-factor confirmatory factor analyses, where
goodness-of-fit indices exceeded .90 (except for one .89 fit index for Mexican American
group interviewed in Spanish). The low acculturated group showed slightly better
goodness-of-fit indices (CFI = .93, NFI = .92, NNFI = .92), while the Mexican American
group interviewed in Spanish showed the lowest goodness-of-fit indices (CFI =.91, NFI =
.89, NNFI = .90). These results confirmed the overall unidimensionality of the CES-D
for the present analyses.

DIF analyses

After verifying the essential unidimensionality of the CES-D, the application of
IRT and CFA DIF detection using the likelihood ratio tests proceeded. For both methods
this study followed the general approach to hypothesis testing described by Stark et al.
(2006), since the latter showed high power and low Type I error rates across a wide
variety of simulation conditions. In essence, the authors suggest testing for DIF by using
free baseline with Bonferroni correction that can be implemented in both CFA and IRT.
A fully free baseline model (with the exception of a single referent item) was used as the
basis for subsequent nineteen nested model comparisons where one item at a time was
constrained to be equal across groups. Item parameters were compared simultaneously
(discrimination–loadings and locations–intercepts), using Bonferroni corrected $p$-values for flagging DIF items. Detailed analytic procedures for CFA and IRT are described below.

**CFA DIF Detection.** CFA DIF analyses involving item loadings and intercepts were conducted using an analogous strategy with LISREL 8.8. Using the free baseline model, where only the parameters of the referent (Item 1) are constrained across groups, baseline and constrained models were run in succession. The chi-square difference statistics for the nested model comparisons were evaluated using a Bonferroni corrected critical $p$-value. When the observed chi-square difference was greater than the corresponding critical chi-square value (Bonferroni corrected, $\chi^2 = 11.88$ with 2 degrees of freedom), the item was flagged DIF.

**IRT DIF Detection.** With regard to IRT DIF detection, the Graded Response Model (GRM; Samejima, 1969) was estimated using the MULTILOG program because the CES-D scale is polytomous. For the GRM, each four-category item has one discrimination parameter ($a$) and three location parameters ($b_1$, $b_2$ and $b_3$). The discrimination parameter reflects the extent to which an item differentiates between levels of underlying depression, and items with higher $a$ are generally preferred because they are more informative in a psychometric sense. The location parameters refer to the point on the underlying depression scale in which the probability is 50% for endorsing the first category relative to the last 3 categories ($b_1 - 0$ vs. 1, 2, 3), the first 2 categories relative to the last 2 categories ($b_2 - 0$, 1 vs. 2, 3), and the first 3 categories relative to the fourth category ($b_3 - 0$, 1, 2 vs. 3), respectively. To determine whether the GRM used for parameter estimation adequately fits the data, chi-square fit statistics were assessed using
the MODFIT program. Adjusted chi-squares to degrees of freedom ratios for each item were all less than 3, indicating good model-data fit (Drasgow, Levine, Tsien, Williams, & Mead, 1995).

In IRT DIF detection, the concurrent calibration method was subsequently used to put the reference and focal group parameters on a common metric with Item 1 as an anchor item. The designated focal and reference groups are presented in Table 2 and Table 3. In this step, the latent mean of the designated reference group was set to zero, whereas the latent mean of the designated focal group was free to vary. As described for the CFA DIF method, the free-baseline model strategy was also used for each CES-D item, and differences in relative goodness of fit were examined with respect to critical chi-square statistics. Each chi-square difference was compared to Bonferroni corrected \( p \)-values (corrected, \( \chi^2 = 16.31 \) with 4 degrees of freedom), and items exhibiting DIF were flagged.

Results

Sample Characteristics

Descriptive characteristics of the two acculturation groups are summarized in Table 10. Compared to the low acculturated, the high acculturated Mexican American elders were likely to be younger (\( t = 3.96, p < .001 \)), male (\( \chi^2 = 14.48, p < .001 \)), more educated (\( t = -30.48, p < .001 \)), and born in the U.S. (\( \chi^2 = 399.26, p < .001 \)). As I expected, high acculturated elders were more likely than low acculturated to be interviewed in English (\( \chi^2 = 424.95, p < .001 \)). That low acculturated subjects might be interviewed in English may appear contradictory; it resulted from the fact that this study
defined high or low acculturation on the basis of a median split: those closer to the median could function reasonably well in the alternative language.

Table 10. *Study 3 – Descriptive Information of the High and Low Acculturated Mexican Americans*

<table>
<thead>
<tr>
<th>Variable</th>
<th>High Acculturated Mexican Americans (N = 1,340)</th>
<th>Low Acculturated Mexican Americans (N = 1,283)</th>
<th>t (χ²)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Age</td>
<td>72.06/5.93</td>
<td>73.05/6.73</td>
<td>3.96***</td>
</tr>
<tr>
<td>Gender</td>
<td>(54.6%)</td>
<td>(61.9%)</td>
<td>(14.48***)</td>
</tr>
<tr>
<td>Educational Attainment</td>
<td>6.82/3.88</td>
<td>2.88/2.58</td>
<td>-30.48***</td>
</tr>
<tr>
<td>Born in the U.S.</td>
<td>(76.9%)</td>
<td>(38.3%)</td>
<td>(399.26***)</td>
</tr>
<tr>
<td>Instrument Language</td>
<td></td>
<td></td>
<td>(424.95***)</td>
</tr>
<tr>
<td>English</td>
<td>(37.5%)</td>
<td>(4.5%)</td>
<td></td>
</tr>
<tr>
<td>Spanish</td>
<td>(62.5%)</td>
<td>(92.5%)</td>
<td></td>
</tr>
<tr>
<td>CES-D</td>
<td>9.26/8.67</td>
<td>10.89/9.86</td>
<td>4.51***</td>
</tr>
<tr>
<td>Probable Depression (≥16)</td>
<td>(21.4%)</td>
<td>(26.5%)</td>
<td>(8.83**)</td>
</tr>
</tbody>
</table>

**p < .01, ***p < .001

Of particular relevance is that high acculturated Mexican American elders were significantly less likely to report symptoms on the CES-D scale (*t* = 4.57, *p* < .001).
Using the standard cutoff of 16 on the CES-D for evidence of probable depression, however, both high and low acculturated groups exhibited a greater likelihood for depression than levels previously reported for the non-Hispanic White population (e.g., Bromberger et al., 2004; Cornoni-Huntley, Blazer, Lafferty, Everett, Brock, & Farmer, 1990; Swenson et al., 2000). Significant group differences did persist: 26.5% of the low acculturated and 21.4% of the high acculturated fell into the probable depression category ($\chi^2 = 8.83, p < .01$). These differences between low and high acculturated groups have been also reported in a number of previous studies of older Hispanic populations (e.g., Gonzalez et al., 2001; Mills & Henretta, 2001).

**DIF Results**

Table 11 and Table 12 summarize results of CFA and IRT DIF analyses. In both CFA and IRT methods, this study used Item 1 as a referent and had nineteen model comparisons in which each of the CES-D items was constrained to be equal across groups in each model. Following suggestions by a number of researchers (Hambleton, 2006; Hambleton & Rogers, 1989), this study reported DIF items that revealed consistently across the two methods as a more conservative strategy that would reduce possible Type I error rates, which is of particular concern in DIF detection.

**Acculturation-bias**

As shown in the first data column of Tables 11 and 12, in the comparison of high and low acculturated groups, CFA flagged eight DIF items (Items # 3, 7, 11, 14, 16, 17, 18, and 19) and IRT identified three DIF items (Items # 4, 16, and 17). Two DIF items (Items #16, the reverse-scored ‘I enjoyed life’ and #17, ‘I had crying spells’) were identified in both CFA and IRT. The (not) ‘enjoyed’ item favored the low acculturated 111
<table>
<thead>
<tr>
<th>Models</th>
<th>Acculturation-DIF</th>
<th>Language-DIF</th>
<th>Acculturation × Language-DIF</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>High vs. Low</td>
<td>English vs. Spanish</td>
<td>High-English vs. High-Spanish</td>
</tr>
<tr>
<td>Baseline Model (Referent: Item1)</td>
<td>$\chi^2 (\Delta \chi^2)$</td>
<td>$\chi^2 (\Delta \chi^2)$</td>
<td>$\chi^2 (\Delta \chi^2)$</td>
</tr>
<tr>
<td>Item 2, ‘appetite’</td>
<td>(7.02)</td>
<td>(0.38)</td>
<td>(4.87)</td>
</tr>
<tr>
<td>Item 3, ‘blues’</td>
<td>(43.83)</td>
<td>(19.54)</td>
<td>(4.87)</td>
</tr>
<tr>
<td>Item 4, ‘good’</td>
<td>(4.09)</td>
<td>(10.5)</td>
<td>(4.32)</td>
</tr>
<tr>
<td>Item 5, ‘mind’</td>
<td>(3.38)</td>
<td>(4.16)</td>
<td>(5.66)</td>
</tr>
<tr>
<td>Item 6, ‘depressed’</td>
<td>(10.20)</td>
<td>(2.57)</td>
<td>(0.70)</td>
</tr>
<tr>
<td>Item 7, ‘effort’</td>
<td>(14.20)</td>
<td>(0.77)</td>
<td>(4.04)</td>
</tr>
<tr>
<td>Item 8, ‘hopeful’</td>
<td>(6.00)</td>
<td>(1.66)</td>
<td>(1.42)</td>
</tr>
<tr>
<td>Item 9, ‘failure’</td>
<td>(11.68)</td>
<td>(7.54)</td>
<td>(0.08)</td>
</tr>
<tr>
<td>Item 10, ‘fearful’</td>
<td>(7.47)</td>
<td>(1.57)</td>
<td>(9.23)</td>
</tr>
<tr>
<td>Item 11, ‘sleep’</td>
<td>(14.81)</td>
<td>(19.01)</td>
<td>(56.76)</td>
</tr>
<tr>
<td>Item 12, ‘happy’</td>
<td>(11.18)</td>
<td>(67.05)</td>
<td>(21.99)</td>
</tr>
<tr>
<td>Item 13, ‘talked less’</td>
<td>(0.80)</td>
<td>(12.67)</td>
<td>(19.01)</td>
</tr>
<tr>
<td>Item 14, ‘lonely’</td>
<td>(20.13)</td>
<td>(1.61)</td>
<td>(16.25)</td>
</tr>
<tr>
<td>Item 15, ‘unfriendly’</td>
<td>(8.89)</td>
<td>(25.84)</td>
<td>(7.32)</td>
</tr>
<tr>
<td>Item 16, ‘enjoyed’</td>
<td>(12.37)</td>
<td>(127.47)</td>
<td>(114.08)</td>
</tr>
<tr>
<td>Item 17, ‘crying’</td>
<td>(53.50)</td>
<td>(55.57)</td>
<td>(1.59)</td>
</tr>
<tr>
<td>Item 18, ‘sad’</td>
<td>(14.88)</td>
<td>(1.49)</td>
<td>(20.34)</td>
</tr>
<tr>
<td>Item 19, ‘disliked’</td>
<td>(27.65)</td>
<td>(20.73)</td>
<td>(1.09)</td>
</tr>
<tr>
<td>Item 20, ‘get going’</td>
<td>(4.14)</td>
<td>(0.91)</td>
<td>(0.62)</td>
</tr>
</tbody>
</table>

Total # DIF Items: 8, 8, 6, 1, 7

Note. High = high acculturated group; Low = low acculturated group; English = group interviewed in English; Spanish = group interviewed in Spanish. Items in bold and underlined are common DIF items detected by both CFA and IRT methods. In CFA, DIF flagged if $\Delta \chi^2$ were > 11.88 with 2 degree of freedom. † Reverse-coded item. *Reference group. ‡ Focal group.
Table 12. Study 3 – DIF Results from Item Response Theory (df = 157, Δ df = 4)

<table>
<thead>
<tr>
<th>Models</th>
<th>Acculturation-DIF</th>
<th>Language-DIF</th>
<th>Acculturation × Language-DIF</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>High vs. Low</td>
<td>English vs. Spanish</td>
<td>High-English vs. High-Spanish</td>
</tr>
<tr>
<td>Baseline Model</td>
<td>$\chi^2$ ($\Delta \chi^2$)</td>
<td>$\chi^2$ ($\Delta \chi^2$)</td>
<td>$\chi^2$ ($\Delta \chi^2$)</td>
</tr>
<tr>
<td>(Referent: Item 1)</td>
<td>41432.5</td>
<td>40341.7</td>
<td>21282.7</td>
</tr>
<tr>
<td>Comparison Models</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Item 2, ‘appetite’</td>
<td>(3.0)</td>
<td>(3.7)</td>
<td>(5.5)</td>
</tr>
<tr>
<td>Item 3, ‘blues’</td>
<td>(3.0)</td>
<td>(15.1)</td>
<td>(6.0)</td>
</tr>
<tr>
<td>Item 4, ‘good’†</td>
<td>(17.4)</td>
<td>(14.0)</td>
<td>(10.9)</td>
</tr>
<tr>
<td>Item 5, ‘mind’</td>
<td>(5.0)</td>
<td>(2.8)</td>
<td>(3.2)</td>
</tr>
<tr>
<td>Item 6, ‘depressed’</td>
<td>(12.0)</td>
<td>(6.2)</td>
<td>(1.6)</td>
</tr>
<tr>
<td>Item 7, ‘effort’</td>
<td>(4.8)</td>
<td>(17.4)</td>
<td>(20.1)</td>
</tr>
<tr>
<td>Item 8, ‘hopeful’†</td>
<td>(6.7)</td>
<td>(39.3)</td>
<td>(40.0)</td>
</tr>
<tr>
<td>Item 9, ‘failure’</td>
<td>(7.3)</td>
<td>(10.9)</td>
<td>(7.7)</td>
</tr>
<tr>
<td>Item 10, ‘fearful’</td>
<td>(3.6)</td>
<td>(11.6)</td>
<td>(11.2)</td>
</tr>
<tr>
<td>Item 11, ‘sleep’</td>
<td>(7.5)</td>
<td>(5.9)</td>
<td>(10.0)</td>
</tr>
<tr>
<td>Item 12, ‘happy’†</td>
<td>(14.9)</td>
<td>(37.7)</td>
<td>(21.8)</td>
</tr>
<tr>
<td>Item 13, ‘talked less’</td>
<td>(2.2)</td>
<td>(4.7)</td>
<td>(4.8)</td>
</tr>
<tr>
<td>Item 14, ‘lonely’</td>
<td>(5.6)</td>
<td>(3.0)</td>
<td>(4.3)</td>
</tr>
<tr>
<td>Item 15, ‘unfriendly’</td>
<td>(7.5)</td>
<td>(15.1)</td>
<td>(7.4)</td>
</tr>
<tr>
<td>Item 16, ‘enjoyed’†</td>
<td>(29.4)</td>
<td>(88.1)</td>
<td>(61.9)</td>
</tr>
<tr>
<td>Item 17, ‘crying’</td>
<td>(17.4)</td>
<td>(14.1)</td>
<td>(3.8)</td>
</tr>
<tr>
<td>Item 18, ‘sad’</td>
<td>(8.7)</td>
<td>(3.8)</td>
<td>(4.3)</td>
</tr>
<tr>
<td>Item 19, ‘disliked’</td>
<td>(15.9)</td>
<td>(25.6)</td>
<td>(13.1)</td>
</tr>
<tr>
<td>Item 20, ‘get going’</td>
<td>(5.7)</td>
<td>(7.6)</td>
<td>(4.9)</td>
</tr>
<tr>
<td>Total # DIF Items</td>
<td>3</td>
<td>5</td>
<td>4</td>
</tr>
</tbody>
</table>

*Note.* High = High Acculturated Group; Low = Low Acculturated Group.

Items in bold and underlined are common DIF items detected by both CFA and IRT methods. In IRT, DIF flagged if $\Delta \chi^2$ were $> 16.31$ with 4 degree of freedom. † Reverse-coded item.  robber group.  b Focal group
group over the high acculturated group, suggesting a greater endorsement of low enjoyment among low acculturated subjects. In other words, high acculturated Mexican Americans were more likely than low acculturated to endorse the positive side of the ‘enjoyed’ item. For the ‘crying’ item, low acculturated elders showed a greater tendency to endorse in general, but at severe levels of depressive symptoms among Mexican Americans, high acculturated elders showed a greater endorsement.

Language-bias

As presented in the second data column of Table 11 and 12, in the comparison of groups interviewed in English and Spanish, CFA identified eight DIF items (Items # 3, 11, 12, 13, 15, 16, 17, and 19) and IRT displayed five DIF items (Items # 7, 8, 12, 16, and 19). There were three common DIF items (Items # 12, ‘I was happy,’ #16, ‘I enjoyed life’ and #19, ‘people disliked me’) favoring the group interviewed in Spanish, which indicates that those interviewed in Spanish had greater propensities to endorse the three DIF items than those interviewed in English. It should be noted that because of reverse-scoring, responses to the two positive affect items (Items #12, ‘happy’ and #16, ‘enjoyed’) indicate that those interviewed in Spanish were less likely to endorse the ‘happy’ and ‘enjoyed life’ end of the items than were those interviewed in English.

Acculturation × Language-bias

Three sets of acculturation × language comparisons were made to evaluate whether item bias was more likely to be associated with acculturation or language: 1) to examine how instrument language (English/Spanish) was associated with item bias within the same level of acculturation (i.e., controlling the acculturation effect), high acculturated Mexican Americans interviewed in English (i.e., High-English) and high
acculturated Mexican Americans interviewed in Spanish (i.e., High-Spanish) were compared; 2) to examine how acculturation affected item bias on the CES-D within the same language use condition (i.e., controlling the language effect), high acculturated Mexican Americans interviewed in Spanish (i.e., High-Spanish) and low acculturated Mexican Americans interviewed in Spanish (i.e., Low-Spanish) were compared; and 3) to examine possible interaction effects of acculturation and instrument language, high acculturated Mexican Americans interviewed in English (i.e., High-English) and low acculturated Mexican Americans interviewed in Spanish (i.e., Low-Spanish) were compared.

First, in the comparison of High-English and High-Spanish combinations, the amount of instrument language DIF was slightly decreased by controlling acculturation effects (Table 11 and Table 12, third data column). Compared with the language DIF results, both CFA and IRT flagged slightly less DIF items: CFA identified six DIF items (Items # 11, 12, 13, 14, 16, and 18) and IRT flagged four DIF items (Items # 7, 8, 12, and 16). Two common DIF items (Items #12, ‘I was happy’ and #16, ‘I enjoyed life’) were identified in both CFA and IRT. Both were positive affect items and favored the High-Spanish group. Since the two positive items were reverse-coded, this indicates that those who were higher in acculturation but interviewed in Spanish were less likely to report positive feelings than were those in the High-English group. Results from this comparison suggest that instrument language affected the measurement equivalence of the CES-D even when the level of acculturation was conditioned at the same level, which also indicates that acculturation was not an important factor to explain language DIF.
Second, in the comparison of High-Spanish and Low-Spanish (Tables 11 and 12, fourth data column), the amount of acculturation bias was reduced after controlling the effects of language. The two common DIF items identified from the acculturation DIF results (Items #16 and #17) disappeared in this comparison after the effect of language was controlled. One DIF item (Item #11) was flagged only in CFA, whereas eight DIF items in CFA and three DIF items in IRT were identified from the acculturation DIF results. Results clearly suggest that the acculturation-related item bias in the CES-D was explained by language differences among Mexican Americans.

Lastly, in the comparison of High-English and Low-Spanish (Tables 11 and 12, last data column), more common DIF items were identified compared to any other comparisons, suggesting possible interaction effects of acculturation and instrument language. Five common DIF items were identified (Items #3, ‘I could not shake off the blues’, #4, ‘feeling good about myself’, #12, ‘I was happy’, #16, ‘I enjoyed life’, and #19, ‘people disliked me’). All five favored Low-Spanish, suggesting that Low-Spanish was more likely to endorse the five items than was High-English (recall that Items # 4, 12, and 16 had been reverse-coded, so these three items represent low positive affect). With regard to three positive items, it was easier for High-English to respond to the positive side of these items.

Discussion

The primary purpose of Study 3 was to examine how the level of acculturation and instrument language influences measurement equivalence of the CES-D among Mexican American elders. The H-EPESE dataset used in the analyses was unusual in that relatively large numbers of English proficient subjects were interviewed in Spanish.
The data thus provided a unique opportunity to investigate whether responses to the CES-D items were more likely to be associated with acculturation or instrument language by comparing groups after conditioning acculturation and language at the same level. Study 3 anticipated that item responses to the CES-D would vary by the level of acculturation and language of interview. In addition, the use of two analytic approaches, CFA and IRT, was designed to provide stronger evidence for the potential item bias.

One major finding was that the level of acculturation and instrument language independently affected measurement bias in the CES-D among older Mexican Americans. It is worth mentioning that this may be the first work in gerontology to differentiate the effects of instrument language and acculturation on the CES-D. CFA and IRT identified two acculturation biased (‘I enjoyed life’ and ‘I had crying spells’) and three language biased items (‘I was happy,’ ‘I enjoyed life’ and ‘people disliked me’). One item (‘I enjoyed life’) was biased by both the level of acculturation and language, which will be discussed later. These results supported previous studies showing different patterns of depressive symptoms associated with acculturation (e.g., Jang, Kim, & Chiriboga, 2005; Nguyen et al., 2007) and language (e.g., Guarnaccia et al., 1989). Findings indicate individual differences in responding to the items of the CES-D even within the same racial/ethnic group, which may also reflect different ways of perceive, feel, and express their symptoms depending upon how much people are acculturated and what language people are using.

One novel finding was that the effects of acculturation on the measurement equivalence of the CES-D were entirely explained by language differences. The identified effects of acculturation on item bias in the CES-D were eliminated when the
same language condition was given to high and low acculturated Mexican Americans. In contrast, the identified effects of language on the CES-D were diminished but still persisted even when the same level of acculturation was given. These results clearly suggest the greater effects of language over acculturation on measurement bias in the CES-D. One possible explanation for these results is that instrument language may explain more fundamental differences that lead to response biases to depressive symptoms and therefore within the same language group (those interviewed in Spanish in the present study), differences between high and low acculturated groups may be diminished.

Another possible explanation may be that people who report in their native language of Spanish are more likely to report their symptoms emotionally (e.g., Cuellar & Roberts, 1984) and under this condition, their degree of acculturation may not influence their responses to depressive symptoms because of their elevated symptom expression. However, it is unfortunate that this Study 3 could not compare high and low acculturated Mexican Americans among those interviewed in English due to the small sample size of low acculturated Mexican Americans who responded in English (N = 58). Since no studies have addressed this issue before, further investigation is needed to understand this phenomenon.

A final explanation is that the biases that resulted from language of interview could simply stem from the difficulties involved in finding culturally equivalent wording in the different languages. While the translations may be technically correct, even with respect to colloquial expressions, the extended connotations of the words used in the
English and Spanish versions may exert a subtle bias in terms of predisposition to respond.

Study 3 found four meaningful DIF items biased by the level of acculturation and language, which may reflect sociocultural differences. First, two positive feeling items were biased. ‘I enjoyed life’ was the only item biased by both the level of acculturation and language of interview: Mexican Americans who were highly acculturated and interviewed in English had a greater propensity to report the positive side of this item than their counterparts. Another positive item biased by language of interview (‘I was happy’) also showed a greater endorsement of the positive side of this item among subjects interviewed in English. These response patterns may reflect mainstream American culture learned by high acculturated Mexican Americans and by those interviewed in English—but it should be recalled that only 58 low acculturated subjects were interviewed in English. There has been research suggesting that positive feelings are prominent in mainstream American culture (Ying, 1989) or at least a greater willingness to report positive feelings (Iwata & Buka, 2002). However, it should be noted that findings in this dissertation Study 3 showed disagreement with findings from Nguyen et al (2007), which indicated that low acculturated Hispanic women were predisposed to endorse their positive feelings. These contradictory findings may be due to the use of different samples: their sample included young Mexican American women, whereas the present sample included older Mexican Americans. Further investigation with different Mexican American samples, and with a wider age range, is needed.

Second, consistent with a recent acculturation DIF study among Hispanic women (Nguyen et al., 2007), this study found ‘crying spells’ to be acculturation-biased. The
‘crying spells’ item was over-endorsed by low acculturated Mexican Americans except at the highest depression scores, which may reflect their adherence to Mexican culture, where crying may be an acceptable behavior reflecting suffering (Azocar, Arcán, Miranda, & Muñoz, 2001; Golding, Aneshensel, & Hough, 1991). Interestingly, however, high acculturated elders were more likely to endorse the ‘crying’ item in the situation where their level of depressive symptomatology was high. Nguyen et al. (2007) has also suggested this nonuniform DIF for the ‘crying spells’ item across high and low acculturated groups, although their study did not address the question of whether one group was more favored to respond.

Lastly, with regard to the ‘people disliked me’ item biased by language, the present study identified a greater endorsement among those interviewed in Spanish. Previous studies have suggested this interpersonal problem item was more likely to be biased by race/ethnicity, especially among Blacks (Cole, Kawachi, Maller, & Berkman, 2000) and Mexican Americans (Kim, Chiriboga, & Jang, 2007). I suspect self-perceptions of discrimination experienced by Mexican Americans interviewed in Spanish as a possible source of their predisposition to endorse this item. This response pattern is not surprising given research suggesting that those who are more acculturated (i.e., those with greater usage of English) were less likely to experience discrimination than those less acculturated (i.e., those with greater usage of Spanish) (Finch, Kolody, & Vega, 2000).

Findings from this Study 3 help to explain cultural differences in responses to depressive symptoms among Mexican Americans. The identified differences in response patterns call attention to limitations of current screening for depression using the CES-D
across different cultural groups even within the same racial/ethnic population. The cut-off point of 16 for probable depression has been used since CES-D’s initial development on European Americans (Andresen et al., 1994; Radloff, 1977) and has not changed. Even though we know that some racial/ethnic groups score higher than non-Hispanic Whites, there has been no serious effort at determining what might be an optimal cut-off point for each diverse group. The results of this paper suggest that researchers and clinicians should carefully consider how the CES-D is used. In addition, the results suggest at least one possible approach that might be used in future work: using DIF findings to provide weighting systems that vary by age, gender, race/ethnicity, and—my particular interest in the present study—the level of acculturation as well as language preference as a means of adjusting cut-off scores.

There are some limitations of this Study 3 that warrant consideration. As mentioned earlier, since the present dissertation study included a small sample of low acculturated Mexican Americans interviewed in English (N = 58), the effects of acculturation within the English-speaking group was not considered. The examination of differences in the responses to the CES-D also needs to extend beyond the Mexican American population and beyond acculturation and language as a sorting factor. The issue of how to deal with uncommonly identified DIF items by CFA and IRT remain unsolved at present. Further investigation will be needed to find a source of these discrepant items, and I hope that psychometricians can develop an analytic framework that will ultimately lead to a clear answer to this question.

In sum, the present dissertation Study 3 highlights the importance of understanding differences in responses to depressive symptoms within the same
racial/ethnic group. When established measures are used to screen and assess for depression in diverse racial/ethnic groups, researchers should recognize the risk that people from different cultural background may tend to be misclassified due to their different responses. Culturally appropriate screens for depression should be a high priority in health disparities research, and this line of research should be extended to other mental health screening tools to generate meaningful comparisons across diverse groups.
CHAPTER FIVE: DISCUSSION

Summary of the Study

A critical factor in cross-cultural or cross-national research is the comparability of instruments across the diverse groups being studied. This factor becomes even more critical when the intent of an instrument is to detect problem behaviors that can be mitigated through interventions. The purpose of this dissertation was to examine how culture—defined broadly, in the case of the present study, to include racial/ethnic, sociodemographic, acculturation and language variations—influences measurement properties of the CES-D, one of the most commonly used screening tools for the detection of depressive symptomatology and of probable depression.

Specifically, this dissertation research focused on identifying race/ethnicity-, sociodemographic-, and acculturation and instrument language-related measurement bias in the CES-D. The samples, consisting of Whites, Blacks, and Mexican Americans, were drawn from two relatively large multistage studies. One of these studies, the New Haven EPESE, was funded as part of a set of four studies known collectively as the Established Populations for Epidemiologic Studies of the Elderly (EPESE). The other was a study of Mexican Americans that is often referred to as the Hispanic EPESE (H-EPESE) since it included nearly all of the questions asked in the four EPESE studies, as well as questions specific to a Hispanic population.

Overview of Findings

A series of three substudies were conducted in this dissertation. These studies successively examined issues of measurement bias in greater and greater detail. Thus, in
the first study, the focus was simply on the extent to which there were general issues of measurement bias when the three racial/ethnic groups (i.e., Whites, Blacks, and Mexican Americans) were considered. In the second study, the focus was specifically on how sociodemographic characteristics influence measurement bias within each racial/ethnic group. Effects of age, gender, and educational attainments on measurement bias in the CES-D were tested in Whites, Blacks, and Mexican Americans separately. In the third study, the focus was on Mexican American elders to investigate issues of measurement bias in the CES-D when the level of acculturation and instrument language were considered.

Results demonstrated the utility of the research design. In essence, Study 1 found a lack of measurement equivalence of the CES-D among Mexican Americans in the comparison with both Whites and Blacks. Race/ethnicity-specific items were also identified in Study 1: two interpersonal relation items in Blacks and four positive affect items in Mexican Americans. Study 2 identified the crucial role of gender and educational attainment on item bias in the CES-D. The interaction between gender and educational level and race/ethnicity was also found in Study 2: Mexican American women and lower educated Blacks had a greater predisposition to endorse the ‘crying’ item. In Study 3, acculturation and instrument language-biased items were identified in Mexican Americans. Study 3 also suggested that acculturation-bias was entirely explained by whether the CES-D was administered in the English or the Spanish versions.
Implications

Methodological Implications

To put these results in context, it may be helpful to review some of the major questions facing the field of differential item functioning (DIF). There are at least three major issues needing to be addressed in DIF research. As already addressed in this dissertation research, perhaps the most important issue in DIF research may be how to identify DIF accurately. A second issue in DIF research may be how to explain the source of DIF with regard to meaningful psychological constructs such as cultural values, gender role, and social desirability. A third issue in DIF research may be how to deal with DIF once it is identified.

Detecting DIF

Regarding the accurate identification of DIF, this dissertation study used multiple analytic strategies to detect DIF more accurately. A few studies have addressed the importance of using multiple procedures to detect DIF in previous research (e.g., Hambleton, 2006; Hambleton, & Rogers, 1989). Using two of the most common DIF methods, CFA and IRT, three substudies reported DIF items that showed up consistently across the two analytic procedures in order to make conclusions stronger as to which items showed DIF and which ones did not. In all three substudies, results from CFA and IRT gave us a great amount of information to explain identified DIF items in the CES-D. These results imply that this dissertation makes a significant contribution to DIF detection strategies in the field of gerontology and measurement research.

Explaining DIF
With regard to the explanation of DIF, each study in this dissertation was able to at least partially explain the identified DIF items in terms of sociocultural and psychological constructs. In Study 1, over-endorsement of interpersonal relation items among Blacks were in part explained by perceptions of racial discrimination by Blacks. The differential responses to depressive symptoms among Mexican Americans, when compared to Whites and Blacks, were mostly explained by cultural differences such as less hesitation to admit psychological distress and extreme response style on positive affect items. In Study 2, sociocultural differences between groups such as gender role, self-expressiveness related to educational experience, and social desirability explained the identified gender and educational bias in the CES-D. In Study 3, sociocultural differences within Mexican Americans explained the identified acculturation and instrument language DIF. Most importantly, Study 3 found that acculturation-bias was mostly explained by differences in instrument language (English or Spanish), which may reflect flaws in the instrument translation process or inadequate item formulations (e.g., complex wording or differences in connotations and social desirability).

*Dealing with DIF*

Perhaps one of the most critical issues for future DIF research may be to develop optimal procedures to deal with DIF once it is identified. In other words, what are implications for research if DIF is found? McHorney and Fleishman (2006) contrasted two phases of research. In the measurement development phase, instruments can be easily modified by removing items that manifest DIF and replacing them with different items. However, in a later phase of research, researchers may have limited opportunities to modify instruments.
Especially when secondary data are used to detect DIF, replacing items may not be possible. Removing DIF items, while a perfectly acceptable strategy, may have potential negative consequences. Item removal may sacrifice content validity (McHorney & Fleishman, 2006). For example, in Study 1, sixteen items functioned differentially across Mexican Americans and Whites. In other words, only four of the twenty CES-D items were identified to function similarly across the two groups. It is not clear whether it is possible to fully capture depressive symptoms using four items when the array of symptomatology is great. In this context, removing DIF items may also adversely affect other psychometric properties such as reliability.

In cases that researchers identify DIF items but cannot remove these DIF items, adjusting cutoff scores can be used. For example, the magnitude of DIF in addition to the statistical significance of DIF could be identified first in order to estimate parameters that can provide information on effect size. It should be noted that only some DIF procedures such as DFIT, MIMIC modeling, and logistic regression can estimate parameters that provide information on effect size (McHorney & Fleishman, 2006). This dissertation study could not provide adjusted cutoff scores due to the DIF procedures used in the present investigation (i.e., CFA and IRT-Likelihood Ratio test). Much more work needs to be done with regard to establishing general guidelines for gauging the magnitude of DIF effects, which may eventually link to adjusting cutoff scores. In addition, as suggested by McHorney and Fleishman (2006), as research investigating DIF continues to accumulate, it will be more important to develop recommendations or guidelines for how to proceed once DIF has been detected.
Practice Implications

This line of research has important implications with regard to how we screen for depression in different cultural groups, within the United States and cross-nationally. Basically, this line of research shows that although core depressive symptoms exist across all cultures, different cultures conceptualize the problem of depressive symptoms in different ways. In certain non-Western cultures, there may be no equivalent concepts for depression, but that does not mean depression is absent (Marsella, Sartorius, Jablensky, & Fenton, 1985). Depressive symptoms may be experienced, expressed, and responded to in different manners. Thus, this line of research assumes that cultural differences may be found in conceptualization, meaning, and symptom expression of depression across different cultural groups. Therefore, it is important that researchers, clinicians, and practitioners know that depression may present differently across different ethnic groups (Minsky, Vega, Miskimen, Gara, & Escobar, 2003). According to the Diagnostic and Statistical Manual of Mental Disorders, Fourth Edition, “Culture can influence the experience and communication of symptoms of depression. Underdiagnosis or misdiagnosis can be reduced by being alert to ethnic and cultural specificity in the presenting complaints of a major depressive episode.” (American Psychiatric Association, 1994, p.324).

This line of research argues that within the context of cultural variations in the conceptualization, expression, and experience of depressive disorders, screening instruments might be unable to identify depressive disorder for specific cultural groups, and thus adequate measures are needed. In other words, reliable and culturally valid screening instruments to establish equivalence across cultures will be an essential
component of culturally competent clinical practice, especially when differences in the lay conceptualization of depression are found within a given culture. Culturally appropriate or equivalent measures will help improve interventions for depressive disorders among racially or ethnically diverse older adults. It may not be absolutely necessary to have a different scale for each cultural group. However, some modifications of instruments may increase specificity and sensitivity in detecting depression for each cultural group. Also, when the CES-D is used to screen for depression, researchers and clinicians must recognize the risk that people from racially or ethnically diverse groups are more likely to be misclassified in epidemiological studies.

When we screen for depression in different racial/ethnic groups, a combined use of core depressive symptoms and culture-specific symptoms may be useful not only for understanding unique cultural phenomena in specific contexts, but also for enabling comparisons across cultures. For example, in an epidemiologic survey in Puerto Rico (Guarnaccia, Canino, Rubio-Stipec, & Bravo, 1993), one item was added to the Diagnostic Interview Schedule (DIS) asking whether the respondent had ever experienced *ataque de nervios* (nerves attack). The item was introduced in order to obtain information on an idiom of distress indigenous to Puerto Rican culture. The prevalence of this item showed that 13.8% of the sample reported having had a nervous attack at least once in their lifetime and 63% of these met criteria for one of the DIS diagnoses tested, usually depressive disorder. This approach may be particularly useful because of the following reasons: 1) the core DIS remained unchanged, permitting cross-cultural comparisons with basically the same instrument and 2) the additions or modifications to the algorithms that were introduced did not alter the original algorithms,
thus allowing the evaluation of depressive disorders according to the original DIS
procedures as well as according to the modifications introduced in Puerto Rico. This
example shows how epidemiological methods can be augmented with culturally specific
research strategies without abandoning the basic epidemiological goal of across-group
generalizability (Canino et al., 1997).

When some cultural groups appear to experience different depressive symptom
clusters than others, clinicians and therapists working with those groups may need to
adjust their own concepts of depression to permit appropriate diagnosis and treatment.
As noted before, they may also need to be aware that depression may present differently
across different racial/ethnic groups. We all need to view depression as a “fuzzy
concept” or a family of overlapping concepts rather than as a single disorder that presents
in a uniform way (Crockett et al., 2005). Cultural competence of clinicians and therapists
is of particular importance because culturally aware therapists and clinicians know that
body language, goal setting, decision-making styles, and assessment tools are all
culturally laden. Sue and colleagues (1991), for example, found ethnic match of provider
and patient showed longer duration of mental health treatment, as well as better patient
response to treatment. Researchers and clinicians need to become more culturally
sensitive, and culturally informed researchers and clinicians may incorporate some
cultural concepts, such as self-orientation, values, family structure, and individualism-
collectivism orientations.

This line of research may provide useful insights on how we improve the
detection of depression among racially or ethnically diverse older adults. Mui and
colleagues (2002) also suggest that practitioners should be aware that biases stemming
from poor equivalence may produce erroneous estimates of symptoms, and that adjustments such as detection of culturally inappropriate items and changes in cut-off scores, particularly if false positives are a concern, may be warranted. Bilingual or bicultural practitioners may be well suited to attend to such issues.

Limitations and Future Research

Despite the abovementioned implications to research and practice, limitations and directions for future research should be noted. As mentioned briefly, one major question is how to deal with the difference in results yielded from the CFA and IRT methods. Throughout the three studies for dissertation, CFA and IRT did not yield identical results. There may be two possible reasons for these differences. First, given that it is more difficult to detect small amount of DIF than large amount of DIF, the different results may be more likely when there is relatively small amount of DIF rather than large amounts. Another possible explanation for the different results is that observed score differences on the CES-D items across diverse cultural subgroups may play less of a role in the calculation for one of the two methods. However, questions still remain as to how to interpret uncommonly detected DIF items. Clearly, more work needs to be done in health disparities and measurement research to develop clear guidelines to interpret and deal with the different results from various DIF procedures.

One challenging part in DIF research is that none of the DIF analytic procedures produce identical results (e.g., Hambleton, 2006). Hambleton also suggested that the more different the procedures, the more likely they will be to produce different results. For example, Crane and colleagues (2006) summarized the various DIF results with the Mini-Mental State Examination (MMSE) and found some differences with the same DIF
procedure using interchangeable software programs (e.g., MULTILOG and Parscale).
This is why a number of psychometricians are recommending applying multiple DIF
detection approaches for more accurate DIF results and for more definitive information
concerning which items are showing DIF and which ones are not (e.g., Hambleton, 2006;
Wang & Russell, 2005). As mentioned earlier, it has been suggested to use the items that
reveal consistent DIF across different methods (Hambleton, 2006). Based on the results
from this dissertation study, I strongly believe that researchers should also apply multiple
analytic procedures and compare results for future DIF research in order to provide
confidence for findings. It may be also helpful to replicate the DIF findings from one
study to another with different DIF analytic procedures.

Another limitation is that this dissertation (Study 1 and Study 2) did not control
the potential time and cohort differences between the samples from the New Haven
EPESE (collected in 1981-1982) and Hispanic EPESE (collected in 1993-4). The over
ten year differences between those two samples may have led to differential response
patterns. In addition, Study 1 and 2 included a relatively small sample of Blacks. Both
limitations underscore the importance of appropriate nationally representative datasets
that can provide enough information to capture racial/ethnic disparities in health for
future research.

Future research should pay more attention to identifying and explaining
interaction effects between DIF items in depression screening tools and various
exogenous factors such as gender, race/ethnicity, and educational attainment. These
interaction effects have not been considered in previous research. Results from Study 2
and Study 3 showed several nonuniform DIF items in the CES-D suggesting interaction
effects. For example, in Study 2, an interaction effect between a ‘crying spells’ item and gender was found. Unlike previous findings showing women’s higher endorsement on the ‘crying spells’ item, this interaction effect in Study 2 showed evidence that men were more likely to cry than women when their depressive symptoms were severe. Although this finding should be revisited and further investigation on the ‘crying’ item with different samples is warranted, this is very meaningful because it is the first to be addressed in depression research using two sophisticated modern statistical methods (CFA and IRT). Given that identified interaction effects in this dissertation provided research and practice implications, future research should focus on replicating these current findings to ensure and identifying other meaningful interaction effects between DIF items in depression screening tools and other unknown exogenous variables.

Most importantly, future research should focus on the optimal depression screen and optimal cut-score for the CES-D. Although this dissertation study identified culture-biased items in the CES-D, optimal cutoff scores for depression screening tools have not been identified for racially or ethnically diverse older adults, or indeed for individuals of any age or gender. Because cut-off scores are not 100% sensitive with respect to the gold standard criterion variable against which they were validated, there may be misclassification errors (Teresi & Holmes, 2002). These errors can be compounded in diverse racial/ethnic samples. Previous research has reported that when using the same cut-off scores of the CES-D, prevalence rates across racial/ethnic groups varied dramatically, ranging from 3.5% (Germans) to more than 30% (Korean Americans). These findings show that because culturally non-sensitive items may result in more false positives as well as false negatives, cut-off scores for the CES-D will result in higher or
lower prevalence estimates of depression. Thus, at a minimum optimal cut-off scores for the CES-D should be identified to screen racially/ethnically diverse older adults with clinically significant depression.

Final Thoughts

Finally, when we consider culture and depressive symptoms, the following questions are helpful to refrain from imposing diagnostic categories and criteria developed in one culture on another (Tanaka-Matsumi, 2001): 1) How is depression defined by the profession and by the indigenous culture?; 2) What words and concepts are used to describe depression?; 3) Are different words and concepts equivalent?; 4) What aspects are known to be culturally similar and variable?; 5) What would account for cultural similarities and differences?; and 6) How does one communicate depression to others in the same culture? These questions immediately call for testing and establishing the cultural validity of diagnostic categories and their criteria. Application of culturally sensitive and valid assessment of depression will produce culturally competent prevention and treatment for depression, and this will be eventually tailored to meet the needs of specific ethnic and cultural groups. The field of cross-cultural depression is in great need of evaluating the utility of culture-accommodating assessment and treatment practice.

In sum, we must be careful in making comparisons of depressive symptoms across diverse racial/ethnic groups. We learned that there are universal core depressive symptoms similar across all cultures, but depressive symptoms are experienced, expressed, and responded to in different manners across cultures. If we want to optimally understand, assess, and diagnose depression, we need to take into account the cultural
contexts in which it operates. It is impossible for us to create culture-free depression measures now. But, by improving existing depression measures, we can make them more culturally informed or culturally appropriate in different racial/ethnic groups, and that should be our goal.
REFERENCES


Jones, R. N. (2006). Identification of measurement differences between English and Spanish language versions of the Mini-Mental State Examination: Detecting differential item functioning using MIMIC modeling. Medical Care, 44(Suppl. 3), S124-S133.


American respondents in the New Haven and the Hispanic EPESE. Manuscript submitted for publication.


(Eds.), *Multicultural measurement in older populations* (pp. 11-21). New York: Springer Publishing Company.


Advances in theory, measurement, and applied research (pp. 163-185).


Radloff, L. (1977). The CES-D Scale: a self-report depression scale for research in the


of methods based on confirmatory factor analysis and item response theory.
*Journal of Applied Psychology, 87*, 517-529.

health disparities research. *Health Services Research, 40*, 1640-1657.

The interplay between discrimination and socioeconomic status. *Ethnic
Disparities, 9*, 151-165.

Roberts, R. E. (1980). Reliability of the CES-D scale in different ethnic contexts.
*Psychiatry Research, 2*, 125-134.

screen for depression and anxiety: Effects of language and ethnic status.
*Psychiatry Research, 31*, 69-83.

status on reliability and validity of the Center for Epidemiologic Studies-
Depression Scale with psychiatric patients. *Journal of Nervous and Mental
Disease, 177*, 581-592.

Depression in later-life Puerto Rican primary care patients: The role of illness,
stress, social integration, and religiosity. *International Psychogeriatrics, 15*, 239-251


Yang, F. M., & Jones, R. N. (in press). Center for Epidemiologic Studies-Depression scale (CES-D) item response bias found with Mantel-Haenszel method was successfully replicated using latent variable modeling. *Journal of Clinical Epidemiology*.


ABOUT THE AUTHOR

Giyeon Kim received a Bachelor’s Degree in Human Development and Family Studies from Duksung Women’s University in 2000 and a M.A. in Human Development from Ewha Womans University in 2002. She entered the School of Aging Studies at the University of South Florida for her Ph.D. degree in 2003.

While in the Ph.D. program at the University of South Florida, Ms. Kim was involved in a number of health disparities studies as a Principal Investigator/co-Principal Investigator and has been devoted to developing an understanding of advanced analytic models in the mental health and minority aging research. She was a coauthor of thirteen peer-reviewed articles and received The Southern Gerontological Society’s Student Paper Award for her work on measurement issues in multi racial/ethnic society.

She is currently working as a Post-Doctoral Fellow at Temple University and conducts research on racial/ethnic disparities in health and healthcare utilization among older adults.