GENDER DIFFERENTIAL ITEM FUNCTIONING EFFECTS ON VARIOUS ITEM RESPONSE FORMATS OF THE CES-D

by

MICHAELA NICOLE GELIN

B.A., University of British Columbia, 1998
Diploma in Guidance Studies, University of British Columbia, 1999

A THESIS SUBMITTED IN PARTIAL FULFILLMENT OF
THE REQUIREMENTS FOR THE DEGREE OF
MASTER OF ARTS

in

THE FACULTY OF GRADUATE STUDIES

DEPARTMENT OF EDUCATIONAL AND COUNSELING PSYCHOLOGY, AND SPECIAL EDUCATION

With Specialization in

MEASUREMENT, EVALUATION, AND RESEARCH METHODOLOGY

We accept this thesis as conforming to the required standard

THE UNIVERSITY OF BRITISH COLUMBIA

April 2001

©Michaela Nicole Gelin, 2001

Authorization Form

In presenting this thesis in partial fulfilment of the requirements for an advanced degree at the University of British Columbia, I agree that the Library shall make it freely available for reference and study. I further agree that permission for extensive copying of this thesis for scholarly purposes may be granted by the head of my department or by his or her representatives. It is understood that copying or publication of this thesis for financial gain shall not be allowed without my written permission.

Department of Educational and Counseling Psychology, and Special Education
The University of British Columbia
Vancouver, Canada

Date April 5, 2001

Abstract

The present study investigated potentially biased scale items on the Center for Epidemiologic Studies Depression (CES-D) scale in a sample of 600 community-dwelling adults between the ages of 17 and 87 years. The mean age was 46 years for males (N=310) and 42 years for females (N=290). The 20-item CES-D was scored using two binary methods (presence and persistence) and one ordinal method. Gender differential item functioning (DIF) was explored using Zumbo's (1999) ordinal logistic regression method with corresponding logistic regression effect size estimator with all three scoring methods. After statistically matching males and females on the underlying ability, gender DIF was found with the CES-D item *crying* for the ordinal and presence methods of scoring. The persistence scoring method identified two DIF items (*effort* and *hopeful*), however, this scoring method was of limited use due to low response rates on some items. Overall, the results indicate that the scoring method has an effect on DIF; thus DIF is a property of the item, scoring method, and purpose of the instrument.

Table Of Contents

List Of Tables

List Of Figures

Acknowledgements

I first wish to acknowledge my mentor and supervisor, Dr. Bruno Zumbo, for his guidance, statistical expertise, enthusiasm, and enduring support throughout all stages of this thesis. His assistance has been invaluable – from the earliest ideas to the final drafts. I would also like to acknowledge my committee member, Dr. Anita Hubley, for providing insightful and thoughtful suggestions throughout this thesis. I would also like to thank Dr. Beth Haverkamp for being the external examiner. Additional thanks are due especially to my parents for their unwavering support and confidence as I pursued both my undergraduate and graduate degrees.

Finally, I would like to acknowledge the Institute for Social Research and Evaluation at the University of Northern British Columbia for providing the data used in this thesis.

Chapter I

Introduction to the Problem

1.1 Introduction

The Center for Epidemiologic Studies Depression scale (CES-D; Radloff, 1977) is a widely used self-report measure developed for use in studies exploring the epidemiology of depressive symptomology in the general population. This scale has also been used in numerous studies to compare the prevalence of depressive symptomology in different groups such as (1) racial/ethnic groups (Aneshensel, Frerichs, & Huba, 1984; Snyder, Cervantes, & Padilla, 1990; Vera et al. 1991), (2) age groups (Gatz & Hurwicz, 1990; Hertzog, Van Alstine, Usala, Hultsch, & Dixon, 1990; Kessler, Foster, Webster, & House, 1992; Liang, Tran, Krause, & Markides, 1989), (3) nonpsychiatric (Devins et al., 1988) and psychiatric (Weissman, Sholomskas, Pottenger, Prusoff, & Locke, 1977) medically ill groups, and (4) between men and women (Clark, Aneshensel, Frerichs, & Morgan, 1981; Krause, 1986; Roberts, Andrews, Lewinsohn, & Hops, 1990; Snyder et al. 1990; Sonnenberg, Beekman, Deeg, & Van Tilburg, 2000; Stommel et al. 1993; Zunzunegui, Beland, Llacer, & Leon, 1998). The majority of these studies reported differences between the groups of interest by comparing the mean scale scores. However, the scale mean differences could indicate that the two groups have a different probability of endorsing the items or that the test items function differently for the two groups (e.g., males and females). Therefore, in order to accurately interpret these comparisons one needs to investigate whether scale score differences are correctly attributable to the construct of interest or whether they are being erroneously attributed to the construct of interest (hence being spurious). That is, if DIF results are erroneously attributable to the construct of interest, thereby inferring that there are

real differences when the differences are actually due to an irrelevant factor, item bias is evident. Conversely, if DIF results are correctly attributable to the construct of interest, thereby inferring that there are real differences between the groups, item impact is evident. The first step required to investigate whether there is item bias or item impact is to investigate differential item functioning (DIF). In this thesis, gender based DIF will be explored with the CES-D to identify test items that function differently between males and females. These comparison groups were chosen because of the large number of studies reporting that women are much more likely than men to report high levels of depressive symptoms.

The CES-D is comprised of 20 items that reflect various aspects of depressive symptomology. Furthermore, the CES-D is used with three types of scoring: ordinal, presence, and persistence. The ordinal scoring method uses a Likert type format that is intended to identify the presence and severity of depressive symptoms (Radloff & Locke, 1986). Alternatively, the presence and persistence scoring methods use a binary scoring format. The essential difference between these two methods is that the latter requires that the symptomology be present longer than the former. DIF should be investigated for each of these scoring methods because it is unclear whether the scoring methods affect DIF. In essence, if an item is found to be DIF with the ordinal scoring method, will it still perform differentially with one of the binary scoring methods or will re-scoring the measure remove the DIF? Furthermore, if scoring methods affect DIF, then inferences derived from DIF results based on one scoring method will not be appropriate or valid for a different scoring method. No previous study has compared DIF for various scoring methods with the same instrument.

The purpose of this thesis is to investigate whether, for each scoring method, any CES-D scale items exhibit gender DIF. The DIF will be explored using Zumbo's (1999) ordinal logistic regression method along with his corresponding effect size estimator that are used in combination to help identify differentially functioning items. In particular, this technique is used to investigate whether or not males and females have a different probability of endorsing the items on the CES-D. If DIF is found with some items, then further investigations are needed to determine if the DIF is because of item bias or item impact. Determining whether an item displays bias or impact has a number of significant implications for researchers, selection personnel, test takers, and policy makers. The primary issue is one of consequential matters of test fairness and equity. That is, there should be a level playing field where men and women have equal opportunities (e.g., in the personnel selection context) and being treated equitably (e.g., in the screening for depression it is inappropriate to portray women as being more depressed than men if it is an artifact of the measurement process).

To this end, the following section will briefly review differential item functioning (DIF), the relevant terminology found in the literature on bias, and Zumbo's ordinal logistic regression method for the detection of DIF. In the next chapter, the few studies that have explored gender biased items of the CES-D will be reviewed. Next, chapter III will describe the study methodology. This will include the study participants, a description of the CES-D measure and the different scoring methods. It will also include the analyses steps, criteria for determining a DIF item, and a brief discussion on the odds ratio statistic used to determine the direction of responding. The results of the study will be reported in Chapter IV. The final chapter will discuss the study results in terms of the research questions: (a) gender DIF

for the CES-D items and (b) the effect of different scoring methods on DIF. Next, implications for research and practice are considered, including what to do with an item that displays DIF. This is followed by limitations of the study and future directions. The thesis concludes with comments on the contribution of this study for measuring depression with the CES-D, gender differences on depression, and for DIF analyses with various scoring methods of the same instrument.

1.2 What is Differential Item Functioning (DIF)?

Differential item functioning (DIF) is a statistical technique that is used to identify differential item response patterns between groups of test-takers (e.g., male versus female, Caucasian versus African American). In assessing response patterns, the comparison groups, conceptualized in this study as men and women, are first statistically matched on the underlying construct of interest (e.g., depressive symptomology), then the DIF methods evaluate the response patterns to individual test items. Thus, as Zumbo (1999) states, DIF occurs when examinees with the same underlying ability on the construct measured by the test, but who are from different groups, have a different probability of endorsing (or correctly answering) the item. He continues with a conceptualization of the basic principle of DIF: "If different groups of test-takers (e.g., males and females) have roughly the same level of something (e.g., knowledge), then they should perform similarly on individual test items regardless of group membership" (p. 5).

In this thesis, DIF matches males and females on depressive symptomology, measured by the CES-D total score. This study could have also matched males and females on a different criterion measure for the latent variable, such as a medical diagnosis of depression from a clinician. This matching alternative would be useful if, for example, the

CES-D total score was found to be misleading or inappropriate, such as the case if DIF were found for many CES-D items.

DIF is different than previous classical test theory techniques used to assess bias because DIF *matches* the groups of interest on the *latent* variable of interest; previous bias studies compared mean scores either without any matching technique or simply compared the factor structure for the groups of interest. Previous studies that found group differences on observed scores, such as group comparisons of scale means, may be misleading because respondents are not first matched on the construct of interest. Thus, matching groups on the variable measured by the test is important for determining whether item responses are equally valid for different groups. However, it should be noted that DIF is a statistical method to flag potentially problematic items. Therefore, it is the first step in determining whether there is item bias or item impact. Further study would be needed by content experts to determine whether one has bias or impact.

Item bias is a value judgment with social, political, and ethical implications, and thus, takes into account the purpose of the test. Specifically, item bias requires that the source of the differential functioning of the item is *irrelevant* to the purpose of the test and/or interpretation of the measure. In essence, item bias is an artifact of the testing procedure. That is, item bias would occur if one group of test-takers (e.g., males) were less likely to endorse the item than the comparison group of test-takers (e.g., females) because the item is tapping a factor over-and-above the factor of interest. For example, if females were less likely to endorse an item from an achievement test of mathematical ability than men because the question required prior knowledge of basketball scores (assuming females do not know the point system used in basketball and males do) then the item is biased. Thus, for item bias

to occur, DIF must be apparent; however, as Zumbo (1999) reminds us, "DIF is a necessary, but not sufficient, condition for item bias" (p. 12).

Item impact is evident when one group of examinees is found to endorse the item more than the other group of examinees because the two groups truly differ on the underlying ability or factor being measured by the test. That is, item impact occurs when the item measures a *relevant* characteristic of the test, and 'real' differences between the two groups of interest are found. For example, if females were less likely to endorse an item from an achievement test of mathematical ability than men matched on mathematical ability, and men and women truly differed on mathematical aptitude, item impact is present.

The distinction between whether the group differences are based on irrelevant or relevant characteristics of the measure is really a question about the purpose of the measure. Therefore, one needs to be clear about the purpose of the test before conducting the analysis. As well, it is important to note, that if an item is flagged as displaying DIF, it does not mean that the item should be automatically omitted from the scale. Rather, items that are flagged as displaying DIF should be carefully analyzed by experts in the appropriate area. For example, if a CES-D item were flagged as displaying DIF then depression researchers should carefully analyze why the item was flagged.

Taken as a whole, methodologically, the DIF analysis is computed by initially matching the two groups of interest on their underlying ability as determined by their overall performance on the test (i.e. the total test score). Next, the DIF statistic is computed and this indicates the extent to which members of one group perform differently from members of some other group who are of comparable overall ability. If DIF is found, further analyses are needed to determine if the DIF is because of item bias or item impact. This thesis does not

continue with the investigation of the source of item bias or impact because it would take experts in the area of depression to conduct that study. Moreover, a study of item bias or impact with differential functioning items of the CES-D for a general population has not been reported in the literature.

1.3 Ordinal Logistic Regression Method

In order to calculate DIF in binary and ordinal scored items, this study used Zumbo's (1999) ordinal logistic regression method. To date, Zumbo's ordinal logistic regression method and corresponding measure of effect size is the only method available for both binary *and* ordinal scored items. Thus, one advantage of this method is that it allows for a direct comparison of the results from binary and ordinal scored items because only this one method is required. That is, statistical method effects do not influence the results. In addition, this method has a corresponding effect size estimator that can be used with binary and ordinal items to help determine the magnitude of DIF. The effect size estimator is extremely important for this particular study because DIF is based on a large sample size and, without an examination of the effect size, trivial effects may appear to be statistically significant.

The ordinal logistic regression method of DIF will be used for the items on the CES-D that are scored using the ordinal method and it will be repeated for the items that are scored using the two binary methods. As described in Zumbo's (1999) handbook, this procedure uses the item response as the dependent variable, with the grouping variable (characterized as variable GRP), total scale score for each examinee (characterized as variable TOTAL) and a group by total interaction as independent variables. This can be expressed as a linear regression of predictor variables on a latent continuously distributed random variable, y*.

The ordinal logistic regression equation is

$$y^* = b_0 + b_1 TOTAL + b_2 GRP + b_3 TOTAL*GRP_i + \varepsilon_i$$

Zumbo's (1999) ordinal logistic regression method provides a test of DIF that measures the effect of group and the interaction, over-and-above the total scale score while, at the same time, statistically matching on the total scale score. This DIF method has a natural hierarchy of entering variables into the model in which the conditioning variable (i.e. the total score) is entered first. Next, the grouping variable (e.g., gender) is entered. This step measures the effect of the grouping variable while holding constant the effect of the conditioning variable. Finally, the interaction term (e.g., TOTAL*GENDER) is entered into the equation which describes whether the difference between the group responses on an item varies over that latent variable continuum. Each of these steps provides a Chi-squared statistic which is used in the statistical test of DIF.

The DIF computation is basically the difference between the Chi-squared value for Step #3 and the Chi-squared value for Step #1. That is, the Chi-squared value for Step #1 is subtracted from the Chi-squared value for Step #3 giving a resultant two degrees of freedom Chi-squared value. The two degrees of freedom arises because it is the difference between the three degrees of freedom at Step #3 and the one degree of freedom at Step #1. Next, the p-value for this resultant two degrees of freedom Chi-squared test is determined by using a Chi-squared probability table that is found in most statistical textbooks.

Just as a Chi-squared statistic is computed for each step in Zumbo's (1999) ordinal logistic regression method, the corresponding effect size estimator is computed for each step. This corresponding effect size value is calculated as an R-squared which can be applied to

both binary and ordinal items. Using these R-squared values, the magnitude of DIF can be computed by subtracting the R-squared value for Step #1 from that for Step #3.

Lastly, in order to classify an item as displaying DIF, one must consider both the two degrees of freedom Chi-squared test of DIF and Zumbo's corresponding effect size measure. Zumbo (1999) proposed two criteria that must be met for an item to be classified as displaying DIF. First, the two degrees of freedom Chi-squared test for DIF must have a p-value less than or equal to 0.01. Second, the corresponding effect size measure must have a R-squared (R^2) value of at least 0.130. However, Jodoin and Gierl's (in press) investigation of DIF effect size measures suggests that this R^2 value is very conservative, and thus they propose a more liberal R^2 value for detecting DIF. Specifically, Jodoin and Gierl propose R^2 values below 0.035 for negligible DIF, between 0.035 and 0.070 for moderate DIF, and above 0.070 for large DIF. Taken together, this thesis will require that (1) an item must have a p-value less than or equal to 0.01 with the two degrees of freedom Chi-square test, and (2) the corresponding R^2 must be greater than or equal to 0.035 for an item to be classified as displaying DIF. If both of these criteria are met, Jodoin and Gierl's effect size criteria will be used to quantify the magnitude of DIF.

Furthermore, if DIF exists for an item, the steps computed in the calculation of DIF using Zumbo's (1999) ordinal logistic regression will be reviewed to determine if the DIF is uniform or non-uniform. Uniform DIF occurs when there is no interaction between the probability of endorsing an item and the group membership being tested. That is, DIF functions in a uniform fashion across the latent continuum of variation (i.e., depression). That is, uniform DIF may occur when the DIF is attributable to differences in item difficulty only. This can be determined by comparing the R-squared values between steps #2 and #1 "to

measure the unique variation attributable to the group differences over-and-above the conditioning variable (the total score) (Zumbo 1999, p. 26). Uniform DIF can also be graphically illustrated as two nonlinear regression lines (one for each group) with a substantial area between the two curves that do not cross over each other. The regression lines typically characterize the probability of endorsing an item as a function of an underlying construct. If uniform DIF is found, the odds ratio will be used to interpret the direction of the DIF (i.e., are females or males more likely to respond?). For the ordinal scoring method, the odds ratio is computed from Step #2 of Zumbo's (1999) ordinal logistic regression (i.e. the regression model adding uniform DIF to the model). Next, a $\chi^2(1)$ test (Step #2 – Step #1) with corresponding p-value and the R^2 effect size values are computed and the odds ratio is computed from the regression coefficient (more technically, the odds ratio is computed as the exponentiation of the regression coefficient).

Conversely, non-uniform DIF occurs when there is an interaction between group membership and the criterion variable (CES-D total score). Non-uniform DIF reflects a situation in which an item might differentially favor a group of respondents (e.g., males) at one end of the latent continuum and disfavor the comparison group (e.g., females) at the other end of the spectrum. In terms of the ordinal logistic regression used for detecting DIF in this thesis, non-uniform DIF can be determined by comparing the R-squared values at step #3 to the R-squared values at step #2. An item is considered uniform DIF if the difference between steps #2 and #3 is statistically non-significant and has a trivial effect size. Non-uniform DIF can also be graphically illustrated as two nonlinear regression lines (one for each group) that cross over each other and characterize the probability of endorsing an item

as a function of an underlying construct. This graph would look similar to an interaction plot from an ANOVA.

Chapter II

Literature Review

2.1 Gender Differences with the CES-D

Since the introduction of the CES-D in 1977, numerous studies have documented that women tend to have higher scale scores than men on the CES-D. That is, women tend to endorse depressive symptoms more than men (e.g., Callahan & Wolinsky, 1994; Clark et al., 1981; Krause, 1986; Sonnenberg et al., 2000). Furthermore, several studies exploring general depression have reported that women experience depression twice as frequently as men whether one looks at depressive symptoms or depressive disorders, and whether referred or non-referred samples are used (e.g., Culbertson, 1997; Leon, Klerman, & Wickramaratne, 1993; Nolen-Hoeksema, 1987). Moreover, this 2:1 prevalence ratio frequently has been found with individuals in the age range between late adolescence and approximately 64 years of age (Nolen-Hoeksema, 1990).

Item level gender differences in depression self-report measures have also been documented. A number of studies report problematic items on the CES-D by comparing mean score differences or the factor structure between males and females. For example, Roberts et al. (1990) found the CES-D items "crying" and "appetite" had different factor loadings for males and females in a sample of adolescents. Unfortunately, these differences often are documented as a form of bias (e.g., gender bias) but the groups are not first matched. However, as mentioned previously, comparing mean score differences on items or total scores between two groups provide uninterpretable results. As Santor et al. (1994) express, "Finding an overall mean difference between two groups does not demonstrate bias, nor does failing to find a difference preclude the possibility of bias" (p. 256).

To date only one study (Cole, Kawachi, Maller, & Berkman, 2000) has presented item-level gender DIF with the CES-D and this study only explored the CES-D with the ordinal scoring method. Although Cole et al. label their methodology as an extension of the Mantel-Haenszel method of detecting DIF, a closer look reveals that they are using Zumbo's method of modelling DIF through ordinal logistic regression, except that they do not use the R^2 effect size method. Instead, they report on the odd-ratio as an effect size. Using this technique with a sample of 2340 community dwelling adults 65 years of age or older, Cole et al. found that the CES-D item "crying" functioned differently; the proportional odds of women responding higher on the "crying" item were 2.14 times that of men matched on overall depressive symptoms. That is, women were more than twice as likely to endorse the "crying" item than men.

While the study by Cole et al. (2000) is the only item-level study of gender differences with the CES-D, Stommel et al. (1993) assessed item bias by gender on the CES-D using factor analysis. Using a series of multi-sample confirmatory factor analysis models on a sample of 1212 subjects (708 cancer patients between the ages of 19-89, average age of 61; 504 caregivers of chronically ill elderly between the ages of 18-88, average age of 63), Stommel et al. (1993) found that the items "crying" and "talked less" were gender biased. Females were more likely to endorse the item "crying spells" compared to males, while females were less likely to endorse the item "talked less" compared to males matched on overall depression. The technique they used was an ordinary least-squares multiple regression with: (a) the item of interest as the dependent variable, (b) gender and the remaining items as predictors, and (c) the t-test of the gender variable was examined to see if it is statistically significant. This approach is a primitive form of DIF analysis that only

allows for uniform DIF and treats the dependent variable (i.e., the item of interest) as a continuous variable.

Although the exclusive use of scale-level methods of factor analysis and reliability, such as the method used by Stommel et al. (1993), are commonly used to investigate item bias, a recent paper by Zumbo (in press) demonstrates that item-level DIF does not manifest itself in scale-level methods. In other words, factor analysis by itself will not necessarily detect DIF. Accordingly, Zumbo (in press) recommends that one must do item-by-item analysis to identify differentially functioning items. He also states that factor analysis can still be conducted; however, it answers a different question. It can only confirm that the test is measuring the same thing in both groups.

However, differential item functioning by gender with the CES-D has not been explored using a sample with a broad age range from the general population, nor has it been explored using the various scoring methods of the CES-D. The Cole et al. (2000) study only used the ordinal scoring. In fact, the study reported in this thesis is the first to compare DIF across various scoring methods – ordinal and binary.

2.2 Previous Applications of DIF to Depression Measures

A broader literature review on the use of DIF methods (e.g., Rasch, Mantel-Haenszel, and logistic regression methods) applied to other depression measures such as the Beck Depression Inventory (BDI: Beck, Ward, Mendelson, Mock, & Erbaugh, 1961), Geriatric Depression Scale (GDS: Yesavage et al., 1983), Hamilton Rating Scale for Depression (HRSD: Hamilton, 1960), Hudson's Generalized Contentment Scale (GCS: Hudson, 1982) and/or Zung's Self-rating Depression Scale (SDS; Zung, 1965) was carried out. This

extensive literature search revealed that DIF techniques have been used only with the BDI and HRSD.

Using nonparametric item response modelling, Santor, Ramsay, and Zuroff (1994) examined gender item bias on the BDI with a sample of depressed outpatient (N=648) and nonpatient college (N=1182) individuals. Each of the 21 items on the BDI consisted of four graded statements that were scored from 0 to 3, and a total depression score was computed by summing all of the scaled responses. Examining gender bias as a function of the severity of depression with the depressed outpatient sample resulted in three BDI items demonstrating DIF: Item 6 (sense of punishment), item 10 (crying), and item 14 (distortion of body image). "Overall bias was defined as the weighted squared difference between the [option characteristic] curves for men and women" (Santor et al., p.261). A similar analysis was conducted with the college sample. Results from this sample also revealed that item 14 (distortion of body image) demonstrated the greatest amount of gender bias in expected item score; however, no gender bias was found for items 6 (sense of punishment) or 10 (crying).

In a more recent study by Santor and Ramsay (1998), item 1 (depressed mood) from the HRSD was used as an example to illustrate DIF between depressed (N=418) and nondepressed (N=238) individuals. After matching individuals from the two groups on depressive severity (e.g., total HRSD score using an ordinal scoring method), clinically depressed individuals were more likely to endorse item 1 (depressed mood) than nondepressed individuals. The other items on the HRSD have not been examined for DIF.

2.3 Explanations for Gender Differences

The prevalence of findings on gender differences have led to several proposed explanations designed to account for the apparent differences. Such explanations include genetic causes (i.e., depression may be genetically transmitted), social and personality factors, as well as the role of endocrine factors. Another possibility is that the prevalence rates of depression are, in fact, equal in men and women, and the apparent gender differences are believed to reflect an artifact as opposed to true differences. That is, the apparent gender differences do not reflect differences in depression per se, but rather reflect differences in the way men and women express depressive symptoms (e.g., Winokur & Clayton, 1967), recall depressive symptoms (e.g., Angst & Dobler-Mikola, 1984), are willing to report depressive symptoms (e.g., Frank, Carpenter, & Kupfer, 1988), and/or even diagnostic biases among mental health professionals (e.g., Lopez, 1989; Potts, Burnam, & Wells, 1991; Wrobel, 1993).

However, there is no clear evidence that men and women express depressive symptoms differently (Nolen-Hoeksema, 1987). Similarly, there are inconclusive findings that gender differences are accounted for by a greater tendency for women to remember symptoms (e.g., Coryell, Endicott, & Keller, 1992; Fennig, Schwartz, & Bromet, 1994; Wilhelm & Parker, 1994) and/or admit to experiencing depressive symptoms (e.g., King & Buchwald, 1982; Tousignant, Brosseau, & Tremblay, 1987). On the other hand, a number of studies have supported the possibility that mental health practitioners tend to overdiagnose depression in women and underdiagnose depression in men (Lopez, 1989; Wrobel, 1993; Loring & Powell, 1988; Potts et al. 1991), thereby contributing to artificial prevalence rates.

What is unclear from this literature is whether the differences being found are: (a) real differences -- impact, (b) differences due to a measurement artifact – bias, or (c) real differences have been minimized or exaggerated by measurement artifacts.

2.4 Research Questions

Given that (a) few studies have explored gender DIF in depression measures, (b) only one study has investigated gender DIF for the CES-D and that this study (Cole et al., 2000) focused on one sample of seniors 65 years of age or older, and (c) no study has compared DIF for ordinal and binary item formats on the same scale, the present study is needed and will contribute to the literature on depression and gender differences, the CES-D, and the psychometrics of DIF. The research questions investigated in this thesis are:

i) Does gender DIF exist for the CES-D for the ordinal, presence, and persistence scoring formats?

ii) Are any CES-D items found as DIF irrespective of the scoring method (i.e., for all the scoring methods)?

iii) Are any CES-D items found to be DIF for only some of the scoring methods?

Therefore, the two purposes of this thesis are (a) to investigate gender DIF for the CES-D items, and (b) to investigate whether different scoring methods affect the DIF results. Given Cole et al.'s (2000) findings with a sample of seniors, I expect that the *crying* item will demonstrate gender DIF for the ordinal scoring method in a general population. However, given that there has been no empirical or theoretical work comparing the effect of scoring method, it is unclear whether the different scoring methods have an effect on DIF.

Chapter Three

Methodology

3.1 Participants

Individuals who were included in this study were obtained from the Health and Health Care Survey carried out by the Institute for Social Research and Evaluation (ISRE) at the University of Northern British Columbia, Canada, in the fall of 1998. The sample comprised of 600 community-dwelling adults living in Northern British Columbia: 290 females and 310 males, who were drawn randomly from the Dominion phone list. The mean age of female participants was 42 years (SD = 13.4, range = 18 to 87 years), and the mean age of male participants was 46 years (SD = 12.1, range = 17 to 82 years).

3.2 Measure

The Center for Epidemiologic Studies – Depression (CES-D) scale used in this study is a 20 item self-administered instrument originally introduced by Lenore Radloff (1977). This scale was designed to measure the current feelings of depression in the general population. Although this scale has been applied to various clinical samples (e.g., Craig & Van Natta, 1976; Weissman et al., 1977), it was never designed to be used as a screening tool for identifying clinical depression (e.g., within standardized systems such as DSM-IV; American Psychiatric Association, 1994) or for discriminating among subtypes of depression. The CES-D has also been translated into many different foreign languages (e.g., Caetano, 1987) and it has been validated for use with a number of different ethnic groups (e.g., Roberts, 1980), as well as for specific age groups such as children (e.g., Weissman, Orvaschel, & Padian, 1980), adolescents, and the elderly (e.g., DeForge & Sobal, 1988; Gatz & Hurwicz, 1990).

3.2.1 CES-D Item and Total Scoring

The CES-D, reproduced in Figure 1, asks respondents to indicate the frequency/duration with which they have experienced a specific symptom associated with depression (e.g., My sleep was restless) during the previous week. Each item has four options that have specific anchors which correspond to the frequency that each of the 20 symptoms was experienced. These anchors are intended to reflect the differences in the presence and severity of depressive symptoms and are usually labelled as: Option 0, *rarely or none of the time / less than 1 day*; Option 1, *some or a little of the time / 1-2 days*; Option 2, *occasionally or a moderate amount of the time / 3-4 days*; and Option 3, *most of the time / 5-7 days*. Using this response scale, the CES-D can be scored in three different ways: an ordinal scoring format and two dichotomous scoring formats (see Figure 2).

Originally, the four options are scored 0, 1, 2, or 3, respectively, which may be termed the "ordinal" method of scoring. Next, the scoring of the positively worded items (item 4, 8, 12 and 16) is reversed, and then all 20 scaled responses are summed for a possible range of scores from 0 to 60.

Alternatively, the options may be scored dichotomously with respect to a specific threshold. The most popular dichotomous scoring method is termed the "presence" method of scoring. This method refers to a respondent's report of having experienced the symptom at least some of the time during the preceding week (i.e. for 1 to 7 days). This method is used when researchers are interested only in the presence or absence of any depressive symptomology. In this case, Option 0 is assigned a score of 0, indicating no depression, and all other response options (Options 1, 2, and 3) are assigned a score of 1, indicating depression. Then the positively worded items are reverse scored (ones are recoded as zeros,

and vice versa), and lastly each scaled response is summed for a possible total scale score of 20.

An alternative dichotomous format, termed the "persistence" method, is used when researchers are interested only in whether an individual is likely to be depressed. The "persistence" of a symptom usually refers to the respondent's report of having experienced the symptom for 3-7 days during the preceding week. For this method, Option 0 and Option 1 are scored as 0, and Option 2 and Option 3 are scored as 1. Next, the positively worded items are reverse scored, and then each scaled response is added for a possible range of scores between 0 and 20.[1] Studies using dichotomous formats have been presented by Clark et al. (1981), Craig and Van Natta (1976), Myers and Weissman (1980), Roberts and Vernon (1983), and Santor and Coyne (1997).

[1] An extreme version of the "persistence" method of scoring is also used with the CES-D, and refers to the respondent's report of having experienced the symptom for 5-7 days during the preceding week. This is extreme because an individual must endorse Option 3 for a symptom to be scored as indicating depression. This scoring method was not used in this thesis because there was not enough variability in the item responses for the ordinal logistic regression to be computed. That is, as expected in a general population survey, few respondents select 5-7 days.

Figure 1. Center for Epidemiologic Studies Depression (CES-D) Scale.

Center for Epidemiologic Studies Depression (CES-D)
Scale: Format for Self-Administered Use

INSTRUCTIONS: Using the scale below, please circle the number for each statement that best describes how often you felt or behaved this way during the past week.

0 = Rarely or none of the time (less than 1 day)
1 = Some or a little of the time (1-2 days)
2 = Occasionally or a moderate amount of time (3-4 days)
3 = Most or all of the time (5-7 days)

DURING THE PAST WEEK:	Less than 1 day	1-2 days	3-4 days	5-7 days
1. I was bothered by things that usually don't bother me.	0	1	2	3
2. I did not feel like eating; my appetite was poor.	0	1	2	3
3. I felt that I could not shake off the blues even with help from my family or friends.	0	1	2	3
4. I felt that I was just as good as other people.	0	1	2	3
5. I had trouble keeping my mind on what I was doing.	0	1	2	3
6. I felt depressed.	0	1	2	3
7. I felt that everything I did was an effort.	0	1	2	3
8. I felt hopeful about the future.	0	1	2	3
9. I thought my life had been a failure.	0	1	2	3
10. I felt fearful.	0	1	2	3
11. My sleep was restless.	0	1	2	3
12. I was happy.	0	1	2	3
13. I talked less than usual.	0	1	2	3
14. I felt lonely.	0	1	2	3
15. People were unfriendly.	0	1	2	3
16. I enjoyed life.	0	1	2	3
17. I had crying spells.	0	1	2	3
18. I felt sad.	0	1	2	3
19. I felt that people dislike me.	0	1	2	3
20. I could not get "going".	0	1	2	3

Note: Items are summed after reverse scoring of items 4, 8, 12, and 16. Total CES-D scores range from 0-60, with higher scores indicating higher levels of general depression.

Figure 2. Scoring the CES-D.

Scoring the CES-D

Each item has four options:

Option 0, *rarely or none of the time / less than 1 day*

Option 1, *some or a little of the time / 1-2 days*

Option 2, *occasionally or a moderate amount of the time / 3-4 days*

Option 3, *most of the time / 5-7 days*

ORDINAL scoring method

- All four options are scored 0, 1, 2, or 3, respectively
- The total score ranges from 0 - 60.

PRESENCE scoring method

- The respondent's report of having experienced the symptom at least *some of the time* during the preceding week (i.e. for 1 to 7 days).

Option 0 —is assigned→ 0 (indicating no depression)

Option 1, Option 2, Option 3 — assigned → 1 (indicating depression)

- The total score ranges from 0 – 20.

PERSISTENCE scoring method

- The respondent's report of having experienced the symptom for *3-7 days* during the preceding week.

Option 0, Option 1 → 0 (indicating no depression)

Option 2, Option 3 → 1 (indicating depression)

- The total score ranges from 0 – 20.

3.3 Analysis

Based on the observation that reliability coefficients are often used as a standard of comparison to show equivalence of measures, they will be reported in this thesis. Reporting reliability coefficients is such common practise that not reporting them gives the impression of an incomplete analysis. Specifically, for all three scoring methods of the CES-D, coefficient alpha reliabilities will be reported for males and females separately, and for the overall scale (males and females combined). However, although this classical test statistic is computed as an estimate of the test reliability, it should be interpreted cautiously because item-level DIF does not manifest itself in the reliability coefficient (Zumbo, in press). That is, the reliability estimates may be the same for males and females, however, some items may still be DIF.

Next, gender DIF will be investigated for each item of the CES-D using Zumbo's (1999) ordinal method of logistic regression and corresponding effect size measure for each scoring method. In all three gender DIF analyses, gender (coded as 0=female, 1=male) was the grouping variable. This can be expressed as

$$Y^*=b_0+b_1TOTAL+b_2GENDER+b_3TOTAL*GENDER + \varepsilon_i.$$

As this equation demonstrates, female and male respondents initially are matched according to their total test score (characterized as variable TOTAL) on the CES-D. As discussed previously, this total test score depends on the scoring format. Appendix A provides the SPSS syntax file used to calculate the ordinal logistic regression and corresponding effect size estimator. It should be noted that this syntax file calls for a public domain SPSS macro (filename: ologit.inc) written by Prof. Dr. Steffen Kuhnel, and modified by John Hendricks, University of Nijmegen, The Netherlands (see Appendix A).

As noted earlier in this thesis, the criterion for a DIF item is that (a) the $\chi^2(2)$ has a p-value less than .01, and (b) the R^2 for this $2df$ test be greater than or equal to 0.035. Jodoin and Gierl (in press) showed that this is a statistically powerful criterion. In addition, Jodoin and Gierl's effect size criteria will be used to quantify the magnitude of DIF: R^2 values below 0.035 for negligible DIF, between 0.035 and 0.070 for moderate DIF, and above 0.070 for large DIF.

In addition, the proportional odds ratio for each item will be presented. The odds ratio will be used to help determine the direction of responding. That is, it will identify whether men or women are more likely to endorse the item. Moreover, it can be used to determine the odds of one group responding higher to an individual item than those in the corresponding group, after matching on overall depressive symptomology. For example, a proportional odds ratio of 2.0 can be translated to mean that those in group one (e.g., females) are twice as likely to endorse the item than those in the comparison group (e.g., males coded as zero).

Chapter Four

Results

4.1 Introduction

The results of the analyses are reported for each scoring method separately, starting with the ordinal method, and followed by the presence and persistence method. For each scoring method, the analyses were calculated using SPSS 10.0 for Windows. Moreover, for each scoring method, results from Zumbo's ordinal logistic regression method and corresponding effect size measure, R-squared, are presented in a tabular format. Each table lists the CES-D item number, 1 through 20, the Chi-squared test statistic and the corresponding R-squared effect size measure for each step in the model. The final column reports the DIF computation, which includes the two degrees of freedom Chi-squared test statistic value with its p-value, as well as the corresponding R-squared effect size value.

4.2 Assumptions

The key assumption in using ordinal logistic regression for DIF is essential unidimensionality, which presumes that the items on the CES-D only measure one dominant factor. In the present case, results from a confirmatory factor analyses study support the unidimensionality of the scale (Zumbo, Gelin, & Hubley, in press). These authors found that a unidimensional model with method effects modelled for the four positively worded items was the best fit.

4.3 Ordinal scored

4.3.1 Classical analyses

Using coefficient alpha, the reliabilities for the ordinal scored CES-D scale were 0.91 overall, 0.91 for males and 0.90 for females.

4.3.2 Gender DIF analyses

As displayed in Table 1, the results from the ordinal scoring method show that item 17 (crying) displays large gender DIF (DIF $R^2 = .218$). Moreover, comparing the R-squared values at steps #2 and #3, the data suggest that the "crying" item shows predominantly uniform DIF. Uniform DIF means that there is no interaction between the probability of endorsing item 17 and the group membership (e.g., gender) being examined. That is, for the "crying" item, DIF functions in a uniform fashion across the latent variable continuum. Moreover, the proportional odds of women responding higher on the item "I had crying spells" were 9.31 times that of men matched on the total score. That is, women were over nine times more likely to score higher on this item. It should be noted that items 2 (eating) and 18 (sad) showed Chi-squared p-values less than 0.01, but their effect sizes were small according to Jodoin and Gierl's (in press) criterion.

Table 1.

Results from Zumbo's (1999) ordinal logistic regression method and corresponding effect size measure for *ordinal* scoring method – CES-D Items.

CES-D ITEM NUMBER	STEP #1 Total Score in the Model			STEP #2 Total score, and Uniform DIF variable in the model			STEP #3 Total score, Uniform, and Non-uniform DIF variables in the model			DIF COMPUTATION Step #3 – Step #1		
	χ^2 with 1 *df*	p-value	R^2	χ^2 with 2 *df*	p-value	R^2	χ^2 with 3 *df*	p-value	R^2	χ^2 (2) test	p-value[a]	R^2
Item1	240.662	0.000	0.387	241.06	0.000	0.387	243.068	0.000	0.391	2.406	0.150	0.004
Item2	106.687	0.000	0.222	121.033	0.000	0.259	122.618	0.000	0.252	15.931	0.000	0.030
Item3	402.624	0.000	0.663	403.413	0.000	0.663	403.433	0.000	0.662	0.809	0.334	0.000
Item4	142.781	0.000	0.228	142.847	0.000	0.228	142.919	0.000	0.229	0.138	0.467	0.001
Item5	273.01	0.000	0.408	273.151	0.000	0.409	273.338	0.000	0.409	0.328	0.424	0.001
Item6	428.21	0.000	0.613	428.409	0.000	0.614	428.742	0.000	0.613	0.532	0.383	0.000
Item7	264.503	0.000	0.403	268.740	0.000	0.411	270.711	0.000	0.416	6.208	0.022	0.013
Item8	215.962	0.000	0.323	221.988	0.000	0.335	222.097	0.000	0.336	6.135	0.023	0.013
Item9	182.076	0.000	0.418	184.264	0.000	0.427	184.274	0.000	0.428	2.198	0.167	0.010
Item10	169.945	0.000	0.328	170.065	0.000	0.328	170.214	0.000	0.329	0.269	0.437	0.001
Item11	224.794	0.000	0.344	225.563	0.000	0.346	225.566	0.000	0.346	0.772	0.340	0.002
Item12	340.599	0.000	0.488	346.518	0.000	0.499	346.935	0.000	0.499	6.336	0.021	0.011
Item13	227.503	0.000	0.368	277.626	0.000	0.369	227.631	0.000	0.369	0.128	0.469	0.000
Item14	272.698	0.000	0.439	277.634	0.000	0.443	279.235	0.000	0.451	6.537	0.019	0.012
Item15	98.557	0.000	0.204	98.567	0.000	0.204	98.573	0.000	0.204	0.016	0.496	0.000
Item16	358.963	0.000	0.510	363.976	0.000	0.521	364.092	0.000	0.521	5.129	0.038	0.011
Item17	**160.926**	**0.000**	**0.374**	**215.41**	**0.000**	**0.539**	**217.483**	**0.000**	**0.592**	**56.557**	**0.000**	**0.218**
Item18	409.746	0.000	0.611	418.812	0.000	0.619	420.286	0.000	0.620	10.540	0.003	0.009
Item19	117.874	0.000	0.264	120.238	0.000	0.272	121.996	0.000	0.271	4.122	0.064	0.007
Item20	306.049	0.000	0.478	306.19	0.000	0.478	306.335	0.000	0.479	0.286	0.433	0.001

a. Probabilities calculated in Minitab for Windows version 12

Note: Items in bold clearly display gender DIF.

4.4 Presence scored

4.4.1 Classical analyses

Using coefficient alpha, the reliabilities for the presence scored CES-D scale were 0.86 overall, 0.85 for males and 0.87 for females.

4.4.2 Gender DIF analyses

The results for the presence scoring method, displayed in Table 2, show that only item 17 (crying) shows large gender DIF. Furthermore, the difference in R-squared values from Step #2 to Step #3 was relatively small suggesting that the DIF is predominantly uniform. The proportional odds of women responding higher on the item "I had crying spells" were 7.51 times that of men matched on overall depressive symptoms. That is, women were 7.51 times more likely to endorse this item in a presence format than men. Unlike the ordinal scoring method, none of the other items had a Chi-squared p-value less than 0.01 (the threshold for DIF).

Table 2.

Results from Zumbo's (1999) ordinal logistic regression method and corresponding effect size measure for *presence* scoring method – CES-D Items.

CES-D ITEM NUMBER	**STEP #1** Total Score in the Model			**STEP #2** Total score, and Uniform DIF variable in the model			**STEP #3** Total score, Uniform, and Non-uniform DIF variables in the model			**DIF COMPUTATION** Step #3 – Step #1		
	χ^2 with 1 df	p-value	R^2	χ^2 with 2 df	p-value	R^2	χ^2 with 3 df	p-value	R^2	χ^2 (2) test	p-value[a]	R^2
Item1	236.924	0.000	0.448	236.926	0.000	0.449	237.396	0.000	0.452	0.472	0.395	0.003
Item2	109.651	0.000	0.254	116.730	0.000	0.270	116.734	0.000	0.271	7.083	0.014	0.017
Item3	375.926	0.000	0.733	376.278	0.000	0.736	377.825	0.000	0.749	1.899	0.193	0.016
Item4	6.781	0.000	0.027	6.855	0.000	0.027	7.077	0.000	0.027	0.296	0.431	0.001
Item5	277.150	0.000	0.522	278.532	0.000	0.525	281.566	0.000	0.537	4.416	0.055	0.015
Item6	411.113	0.000	0.712	416.786	0.000	0.722	417.222	0.000	0.722	6.109	0.024	0.011
Item7	285.175	0.000	0.529	290.452	0.000	0.540	292.304	0.000	0.548	7.129	0.014	0.019
Item8	24.040	0.000	0.107	27.364	0.000	0.124	27.365	0.000	0.123	3.325	0.095	0.017
Item9	185.319	0.000	0.492	190.886	0.000	0.518	190.911	0.000	0.516	5.592	0.031	0.025
Item10	179.463	0.000	0.387	180.157	0.000	0.391	181.041	0.000	0.396	1.578	0.227	0.009
Item11	155.945	0.000	0.387	158.508	0.000	0.390	158.51	0.000	0.390	2.565	0.139	0.003
Item12	27.883	0.000	0.164	28.603	0.000	0.168	28.767	0.000	0.165	0.884	0.321	0.001
Item13	246.068	0.000	0.463	249.217	0.000	0.472	250.872	0.000	0.479	4.804	0.045	0.017
Item14	206.855	0.000	0.396	213.297	0.000	0.406	213.300	0.000	0.406	6.445	0.020	0.010
Item15	127.910	0.000	0.300	128.620	0.000	0.304	128.632	0.000	0.304	0.722	0.348	0.004
Item16	47.661	0.000	0.228	48.773	0.000	0.234	49.495	0.000	0.229	1.834	0.200	0.001
Item17	**170.730**	**0.000**	**0.455**	**211.890**	**0.000**	**0.569**	**214.467**	**0.000**	**0.647**	**43.737**	**0.000**	**0.191**
Item18	373.058	0.000	0.659	376.296	0.000	0.663	377.145	0.000	0.667	4.087	0.065	0.008
Item19	155.434	0.000	0.393	155.524	0.000	0.393	155.948	0.000	0.393	0.514	0.387	0.000
Item20	274.694	0.000	0.514	276.050	0.000	0.517	277.246	0.000	0.522	2.552	0.140	0.008

a. Probabilities calculated in Minitab for Windows version 12.

Note: Items in bold clearly display gender DIF.

4.5 Persistence scored

4.5.1 Classical analyses

Using coefficient alpha, the reliabilities for the persistence scored CES-D scale were 0.87 overall, and 0.89 for males and 0.85 for females.

4.5.2 Gender DIF analyses

For the persistence scoring method, the results from Table 3 show that items 7 (effort) and 8 (hopeful) show moderate gender DIF. Furthermore, comparing the R-squared values at Step #2 and #3, the data suggest that the DIF for these items is predominantly uniform. The proportional odds of men responding higher on item 7 were 2.03 times that of women matched on the total score, and the proportional odds of men responding higher on item 8 were 1.90 times that of women matched on the total score.

It should also be noted that the calculation for items 2, 3, 9, 10, 15, 17, and 19 could not be computed because there was a low probability of endorsing the items. In other words, there was not enough variability in the item responses for the ordinal logistic regression to be computed. It is very important to note that item 17 (crying) is among these items that have very low variability. Table 4 shows the endorsement proportions for each item on the CES-D using the persistence scoring method. In classical test theory, these generally are defined as the proportion of examinees who answer an item correctly and, thus, are a measure of item difficulty. In this context, however, endorsement proportions are more accurately indicative of how much latent variable is required before an individual endorses a particular item. For example, using the persistence scoring method, an item is only endorsed by an individual if they experience the symptom for 3-7 days. In other words, this scoring method requires that an individual must have a lot of the latent variable, depression, before they endorse an item.

Thus, with such a strict criteria for endorsing an item, it is no surprise that all 20 CES-D items have very low endorsement proportions, ranging from 3.8% to 27%. Table 4 also shows that items 2, 3, 9, 10, 15, 17, and 19 have low endorsement proportions ranging from 3.8 % to 6.7%. Consequently, these items exhibit very little variability. Endorsement proportions by gender were also calculated. As shown in Table 4, females had higher endorsement proportions than males for 16 of the 20 CES-D items. However, according to Cohen's (1992) effect size criteria, the difference in endorsement proportions between males and females was trivial. Moreover, as noted in the introduction of this thesis, one should interpret these gender differences cautiously because of the absence of matching males and females on the latent trait– which is why DIF analyses are needed. The implications of these very low endorsement proportions will be presented in the discussion section of this thesis, Chapter V.

Table 3.

Results from Zumbo's (1999) ordinal logistic regression method and corresponding effect size measure for *persistence* scoring method – CES-D Items.

CES-D ITEM NUMBER	**STEP #1** Total Score in the Model			**STEP #2** Total score, and Uniform DIF variable in the model			**STEP #3** Total score, Uniform, and Non-uniform DIF variables in the model			**DIF COMPUTATION** Step #3 – Step #1		
	χ^2 with 1 df	p-value	R^2	χ^2 with 2 df	p-value	R^2	χ^2 with 3 df	p-value	R^2	χ^2 (2) test	p-value[a]	R^2
Item1	125.38	0.000	0.340	125.406	0.000	0.340	127.678	0.000	0.361	2.298	0.158	0.021
Item2												
Item3												
Item4	97.684	0.000	0.217	97.817	0.000	0.217	97.863	0.000	0.216	0.179	0.457	-0.001
Item5	142.074	0.000	0.344	142.581	0.000	0.348	142.733	0.000	0.350	0.659	0.360	0.006
Item6	215.327	0.000	0.526	218.023	0.000	0.532	219.102	0.000	0.559	3.775	0.076	0.033
Item7	**142.206**	**0.000**	**0.350**	**147.275**	**0.000**	**0.379**	**150.656**	**0.000**	**0.395**	**8.45**	**0.007**	**0.045**
Item8	**160.764**	**0.000**	**0.351**	**167.954**	**0.000**	**0.382**	**168.57**	**0.000**	**0.391**	**7.806**	**0.010**	**0.040**
Item9												
Item10												
Item11	170.176	0.000	0.375	170.543	0.000	0.374	173.604	0.000	0.406	3.428	0.090	0.031
Item12	231.299	0.000	0.507	231.989	0.000	0.512	232.161	0.000	0.517	0.862	0.325	0.010
Item13	102.03	0.000	0.282	102.872	0.000	0.287	103.415	0.000	0.285	1.385	0.250	0.003
Item14	172.153	0.000	0.424	172.212	0.000	0.424	172.258	0.000	0.426	0.105	0.474	0.002
Item15												
Item16	253.665	0.000	0.553	253.853	0.000	0.555	254.304	0.000	0.564	0.639	0.363	0.011
Item17												
Item18	211.519	0.000	0.565	215.059	0.000	0.580	217.085	0.000	0.569	5.566	0.031	0.004
Item19												
Item20	178.343	0.000	0.449	178.401	0.000	0.450	179.457	0.000	0.465	1.114	0.286	0.016

a. Probabilities calculated in Minitab for Windows version 12

Note: Items in bold clearly display gender DIF.

Table 4. Item endorsement proportions by gender for the CES-D scale (310 males; 290 females) using the *persistence* scoring method.

CES-D item number	Overall endorsement proportions	Male endorsement proportions	Female endorsement proportions
1	0.098	0.094	0.103
2	**0.067**	**0.032**	**0.103**
3	**0.065**	**0.061**	**0.069**
4	0.200	0.184	0.217
5	0.132	0.129	0.134
6	0.118	0.090	0.148
7	0.125	0.139	0.110
8	0.230	0.248	0.210
9	**0.048**	**0.052**	**0.045**
10	**0.067**	**0.052**	**0.083**
11	0.272	0.242	0.303
12	0.193	0.181	0.207
13	0.098	0.084	0.114
14	0.120	0.106	0.134
15	**0.057**	**0.061**	**0.052**
16	0.193	0.174	0.214
17	**0.050**	**0.026**	**0.076**
18	0.095	0.071	0.121
19	**0.038**	**0.029**	**0.048**
20	0.112	0.103	0.121
Mean	0.12	0.11	0.13
SD	0.07	0.07	0.07

Note: The items in **bold** could not be computed with the ordinal logistic regression (see Table 3) because there was a low probability of endorsing the items. This table is only intended to show endorsement proportions and should not be used to investigate or interpret DIF.

Chapter V

Discussion

The primary objective of this thesis was to conduct gender DIF for the CES-D scale with each scoring method. In doing so, this thesis explored whether (a) gender DIF existed for the CES-D for the ordinal, presence, and persistence scoring formats, (b) any CES-D items were identified as DIF irrespective of the scoring method and (c) any CES-D items were found to display DIF for only some of the scoring methods.

5.1 Gender DIF for the CES-D Items

The results in Chapter IV showed that, depending on the scoring method used, at least three of the 20 CES-D items functioned differently among males and females. For the ordinal scoring method, the "crying" item (item 17) showed predominately uniform DIF. Similarly, the "crying" item also showed uniform DIF when the CES-D was scored using the presence method. These findings were consistent with those of Cole et al.'s (2000) study of seniors.

However, it is not known whether or not the "crying" item functions differently for males or females when the persistence scoring method is used because this data sample does not have enough variability in the responses of these items for the ordinal logistic regression to be computed. Although an item showing almost no variability may be considered a poor indicator of the construct being measured, this is unlikely in this context because the items hold up well using the other two scoring methods. It is more likely that these are good items, and that the low variability for these items is only because of the strict criteria used with the persistence method of scoring. Furthermore, given that this data came from the general population, it is unlikely that enough people in this sample would experience symptoms such

as "I had crying spells" for 3-7 days, creating very little variability. That is, individuals from the general population are more likely to experience such symptoms less than two days, an endorsement indicating an individual is unlikely to be depressed with the persistence scoring method.

The results found with the persistence scoring method showed that item 7 (everything was an effort) and item 8 (felt hopeful) were flagged as showing gender DIF. Although these items were flagged as showing gender DIF, this result should be interpreted with some caution because of the low variability found in the responses (item 7, p= 0.125; item 8, p= 0.230). In fact, a close look at all the responses with this scoring format reveals low variability for all the items, which supports the notion that using the persistence scoring method with data from a general population is not appropriate. One needs variability in the item responses for an item to have psychometric utility – i.e., the aim of psychometric methods is to differentiate among respondents and therefore an item without variability does not aid in this differentiation.

To this end, it can be concluded that the "crying" item displays gender DIF for both the ordinal and presence scoring methods. That is, an individual's gender influences how one endorses the "crying" item on the CES-D. In other words, re-scoring the CES-D from the ordinal method to the presence method does not remove the gender DIF of the "crying" item. On the other hand, items 7 and 8 display gender DIF only with the persistence method.

5.2 The Effect of Different Scoring Methods on DIF

The findings in this thesis not only demonstrate that DIF is a property of the item (e.g., the crying item), they also show that DIF is a property of the scoring method. As

mentioned in the introduction of this thesis, no previous study had compared DIF for various scoring formats with the same instrument. That is, prior to this thesis, it was not known whether DIF was dependent on various scoring formats. Given that the CES-D has various scoring formats, this instrument was used to ascertain whether DIF was dependent on the scoring method. As presented in the results section, Chapter IV, different items displayed gender DIF depending on the scoring format used. Thus, when one thinks about DIF one must think about the items, the scoring method, and the interaction of the items and scoring method used. Furthermore, because the scoring method is dependent on the purpose of the instrument, DIF is also a property of the purpose of the measure. With this in mind, researchers exploring why an item displays DIF must consider not only the item, but also the scoring format used and the purpose of the instrument. This is relevant in practice because there are several other instruments that have a Likert-type response format and a variety of suggested binary scoring methods, such as the widely used General Health Questionnaire (Goldberg, 1972; Goldberg & Williams, 2000).

5.3 Implications

Exploring why an item displays DIF is pertinent for decisions regarding what to do with items that function differently for different groups. That is, an item displaying DIF should not be discarded from the instrument before experts in the area can clearly understand why the item is endorsed differently for different groups. An item displaying DIF may reflect item impact (i.e. the groups truly differ on the underlying factor being measured). For example, if the "crying" item is tapping depressive symptomology (i.e. a relevant characteristic of the measure), item impact may be present. For instance, women may be more depressed than men and therefore, they endorse the crying item more. However, if the

"crying" item is tapping a factor other than depression symptomology, item bias may be present. For example, the *crying* item would be biased if men endorsed it more than women because men are socialized not to express their emotions. In order to explore this issue, a talk-aloud protocol may be used wherein individuals orally describe what they are thinking as they respond to an item. Accordingly, this method will provide information as to the *process* of responding to an item. Moreover, this method lends itself nicely to answering the question "What is it, in the process of responding to items, that causes bias?" That is, are there differences in the reasoning processes people use in biased versus unbiased items? Although this talk-aloud protocol approach has not yet been used in the DIF literature, the information provided by this approach may help researchers decide what to do with an item that displays DIF and it may suggest why the item displays DIF.

In addition, one also needs to consider the characteristics of the sample with which the DIF item was found. One should make sure that the item displays DIF across different samples. For example, an item may show DIF between groups of seniors and young adults, but may not show DIF between groups of seniors and children (e.g., irritability). In terms of this thesis's findings, although DIF was found with a sample from the general population, comparisons and generalizations across other populations may be problematic. Specifically, this sample came from Northern British Columbia, an area that is comprised of numerous small, rural towns and one city (population ~80,000). In order to generalize these findings across populations, it is necessary that future studies replicate this study with different populations (such as those from larger urban areas).

Future studies could explore gender DIF for other depression inventories and compare whether similar items display gender DIF. Comparisons across different items from

different depression scales may provide a great deal of information regarding the types of items that are found to display gender DIF, and this information may be used as a basis for making informed judgements and decisions pertaining to the future use and role of such items.

References

American Psychiatric Association. (1994). Diagnostic and statistical manual of mental disorders (4th ed.). Washington, DC: Author.

Aneshensel, C.S., Frerichs, R.R., & Huba, G.J. (1984). Depression and physical illness: A multiwave, nonrecursive causal model. Journal of Health and Social Behavior, 25, 350-371.

Angst, J., & Dobler-Mikola, A. (1984). Do the diagnostic criteria determine the sex ratio in depression? Journal of Affective Disorders, 7, 189-198.

Beck, A.T., Ward, C.G., Mendelson, M., Mack, J., & Erbaugh, J. (1961). An inventory for measuring depression. Archives of General Psychiatry, 4, 561-571.

Caetano, R. (1987). Alcohol use and depression among U.S. Hispanics. British Journal of Addictions, 82, 1245-1251.

Callahan, C.M., & Wolinsky, F.D. (1994). The effect of gender and race on the measurement properties of the CES-D in older adults. Medical Care, 32, 341-356.

Clark, V.A., Aneshensel, C.S., Frerichs, R.R., & Morgan, T.M. (1981). Analysis of effects of sex and age in response to items on the CES-D scale. Psychiatry Research, 5, 171-181.

Cohen, J. (1992). A power primer. Psychological Bulletin, 112, 155-159.

Cole, S.R., Kawachi, I., Maller, S.J., & Berkman, L.F. (2000). Test of item-response bias in the CES-D scale: Experience from the New Haven EPESE study. Journal of Clinical Epidemiology, 53, 285-289.

Coryell, W., Endicott, J., & Keller, M. (1992). Major depression in a nonclinical sample: Demographic and clinical risk factors for first onset. Archives of General Psychiatry, 49, 117-125.

Craig, T.J., & Van Natta, P.A. (1976). Presence and persistence of depressive symptoms in patient and community populations. American Journal of Psychiatry, 133, 1426-1429.

Culbertson, F.M. (1997). Depression and gender: An international review. American Psychologist, 52, 25-31.

DeForge, B.R., & Sobal, J. (1988). Self-report depression scales in the elderly: The relationship between the CES-D and Zung. International Journal of Psychiatry in Medicine, 18, 325-338.

Devins, G.M., Orme, C.M., Costello, C.G. Binik, W.M., Frizzell, B., Stam, H.J., & Pullin, W.M. (1988). Measuring depressive symptoms in illness populations: Psychometric properties of the Center for Epidemiologic Studies Depression (CES-D) scale. Psychology and Health, 2, 139-156.

Fennig, S., Schwartz, J.E., & Bromet, E.J. (1994). Are diagnostic criteria, time of episode and occupational impairment important determinants of the female: Male ratio for major depression? Journal of Affective Disorders, 30, 147-154.

Frank, E., Carpenter, A.B., & Kupfer, D.J. (1988). Sex differences in recurrent depression: Are there any that are significant? American Journal of Psychiatry, 145, 41-45.

Gatz, M., & Hurwicz, M.L. (1990). Are old people more depressed? Cross-sectional data on Center for Epidemiological Studies Depression Scale factors. Psychology and Aging, 5, 284-290.

Goldberg, D. P. (1972). The Detection of Psychiatric Illness by Questionnaire. Maudsley Monograph No. 21. Oxford University Press : Oxford.

Goldberg, D., & Williams, P., (2000). A User's Guide To The General Health Questionnaire. NFER-NELSON.

Hamilton, M. (1960). A rating scale for depression. Journal of Neurology, Neurosurgery and Psychiatry, 23, 56-65.

Hertzog, C., Van Alstine, J., Usala, P.D., Hultsch, D.F., & Dixon, R. (1990). Measurement properties of the Center for Epidemiological Studies Depression scale (CES-D) in older populations. Psychological Assessment: A Journal of Consulting and Clinical Psychology, 2, 64-72.

Hudson, W. (1982). A measurement package for clinical workers. Journal of Applied Behavioral Science, 18, 229-238.

Jodoin, M.G., & Gierl, M.J. (in press). Evaluating Type I error and power rates using an effect size measure with the logistic regression procedure for DIF detection. Applied Measurement in Education.

Kessler, R.C., Foster, C., Webster, P.S., & House, J.S. (1992). The relationship between age and depressive symptoms in two national surveys. Psychology and Aging, 7, 119-126.

King, D.A., & Buchwald, A.M. (1982). Sex differences in subclinical depression: Administration of the Beck Depression Inventory in public and private disclosure situations. Journal of Personality and Social Psychology, 42, 963-969.

Krause, N. (1986). Stress and sex differences in depressive symptoms among older adults. Journal of Gerontology, 41, 727-731.

Leon, A.C., Klerman, G.L., & Wickramaratne, P. (1993). Continuing female predominance in depressive illness. American Journal of Public Health, 83, 754-757.

Liang, J., Tran, T.V., Krause, N., & Markides, K.S. (1989). Generational differences in the structure of the CES-D scale in Mexican Americans. Journal of Gerontology: Social Sciences, 44, S110-S120.

Lopez, S.R. (1989). Patient variable biases in clinical judgement: Conceptual overview and methodological considerations. Psychological Bulletin, 106, 184-203.

Loring, M. & Powell, B. (1988). Gender, race, and DSM-III: A study of the objectivity of psychiatric diagnostic behavior. Journal of Health and Social Behavior, 29, 1-22.

Myers, J.K., & Weissman, M.M. (1980). Use of a self-report symptom scale to detect depression in a community sample. American Journal of Psychiatry, 137, 1081-1084.

Nolen-Hoeksema, S. (1987). Sex differences in unipolar depression: Evidence and theory. Psychological Bulletin, 101, 259-282.

Nolen-Hoeksema, S. (1990). Sex differences in depression. Stanford, CA: Stanford University Press.

Potts, M.K., Burnam, M.A., & Wells, K.B. (1991). Gender differences in depression detection: A comparison of clinical diagnosis and standardized assessment. A Journal of Consulting and Clinical Psychology, 3, 609-615.

Radloff, L.S. (1977). The CES-D scale: A self-report depression scale for research in the general population. Applied Psychological Measurement, 3, 385-401.

Radloff, L.S., & Locke, B.Z. (1986). The community mental health assessment survey and the CES-D Scale. In A.E. Slaby (Series Ed.), Community surveys of psychiatric disorders (pp. 177-189). New Brunswick, NJ: Rutgers University Press.

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

Roberts, R.E., Andrews, J.A., Lewinsohn, P.M., & Hops, H. (1990). Assessment of depression in adolescents using the Center for Epidemiological Studies Depression scale. Psychological Assessment, 2, 122-128.

Roberts, R.E., & Vernon, S.W. (1983). The Center for Epidemiological Studies Depression scale: Its use in a community sample. American Journal of Psychiatry, 140, 41-46.

Santor, D.A. & Coyne, J.C. (1997). Shortening the CES-D to improve its ability to detect cases of depression. Psychological Assessment, 9, 233-243.

Santor, D.A., & Ramsay, J.O. (1998). Progress in the technology of measurement: Applications of item response models. Psychological Assessment, 10, 345-359.

Santor, D.A., Ramsay, J.O., & Zuroff, D.C. (1994). Nonparametric item analyses of the Beck Depression Inventory: Evaluating gender item bias and response option weights. Psychological Assessment, 6, 255-270.

Snyder, V.N.S., Cervantes, R.C., & Padilla, A.M. (1990). Gender and ethnic differences in psychosocial stress and generalized distress among Hispanics. Sex Roles, 22, 441-453.

Sonnenberg, C. M., Beekman, A. T. F., Deeg, D. J. H., & Van Tilburg, W. (2000). Sex differences in late-life depression. Acta Psychiatrica Scandinavica, 101, 286-292.

Stommel, M., Given, B.A., Given, C.W., Kalaian, H.A., Schulz, R., & McCorkle, R. (1993). Gender bias in the measurement properties of the Center for Epidemiologic Studies Depression Scale (CES-D). Psychiatry Research, 49, 239-250.

Tousignant, M., Brosseau, R., & Tremblay, L. (1987). Sex biases in mental health scales: Do women tend to report less serious symptoms and confide more then men? Psychological Medicine, 17, 203-215.

Vera, M., Alegria, M., Freeman, D., Robles, R.R., Rios, R., & Rios, C.F. (1991). Depressive symptoms among Puerto Ricans: Island poor compared with residents of the New York City area. American Journal of Epidemiology, 134, 502-510.

Weissman, M.M., Orvaschel, H., & Padian, N. (1980). Children's symptoms and social functioning self-report scales: Comparison of mother's and children's reports. Journal of Nervous and Mental Disease, 168, 736-740.

Weissman, M.M., Sholomskas, D., Pottenger, M., Prusoff, B.A., & Locke, B.Z. (1977). Assessing depressive symptoms in five psychiatric populations: A validation study. American Journal of Epidemiology, 106, 203-214.

Wilhelm, K., & Parker, G. (1994). Sex differences in lifetime depression rates: Fact or artefact? Psychological Medicine, 24, 97-111.

Winokur, G., & Clayton, P. (1967). Sex differences and alcoholism in primary affective illness. British Journal of Psychiatry, 113, 973-979.

Wrobel, N.H. (1993). Effect of patient age and gender on clinical decisions. Professional Psychology: Research and Practice, 24, 206-212.

Yesavage, J.A., Brink, T.L., Rose, T.L., Lum, O., Huang, V., Adey, M.B., & Leirer, V.O. (1983). Development and validation of a geriatric depression screening scale: A preliminary report. Journal of Psychiatric Research, 17, 37-49.

Zumbo, B.D. (1999). A handbook on the theory and methods of differential item functioning (DIF): Logistic regression modeling as a unitary framework for binary and likert-type (ordinal) item scores. Ottawa, ON: Directorate of Human Resources Research and Evaluation, Department of National Defense.

Zumbo, B.D. (in press). Does item-level DIF manifest itself in scale-level analyses?: Implications for translating language tests. Language Testing.

Zumbo, B. D., Gelin, M. N., & Hubley, A. M. (in press). The construction and use of psychological tests and measures. Encyclopedia of Life Support Systems. United Nations Educational, Scientific and Cultural Organization Publishing (UNESCO-EOLSS Publishing), France.

Zung, W.W.K. (1965). A self-rating depression scale. Archives of General Psychiatry, 12, 63-73.

Zunzunegui, M.V., Beland, F., Llacer, A., & Leon, V. (1998). Gender differences in depressive symptoms among Spanish elderly. Social Psychiatry and Psychiatric Epidemiology. 33, 195-205.

Appendix A

SPSS Syntax File for the Ordinal Logistic Regression and Corresponding Effect Size Estimator

```
* SPSS SYNTAX written by: Bruno D. Zumbo, PhD   .
* University of British Columbia   .
* e-mail: brunozumbo@ubc.ca   .

* Instructions .
* Copy this file and the file "ologit2.inc", and your SPSS data file into the same folder .
* Change the filename, currently 'ordinal.sav' to your file name .
* Change 'item', 'total', and 'grp', to the corresponding variables in your file.
* Run this entire syntax command file.

include file='ologit2.inc'.
execute.


GET
FILE='C:\ordinal.sav'.
 EXECUTE .

compute item= item1.
compute total= cesdtot.
compute grp= gender.

* Regression model with the conditioning variable, total score, in alone.
ologit var = item total
    /output=all.
execute.

* Regression model adding uniform DIF to model.
ologit var = item total grp
    /contrast grp=indicator
    /output=all.
execute.


* Regression model adding non-uniform DIF to the model.
ologit var = item total grp total*grp
    /contrast grp=indicator
    /output=all.
execute.
```