A PSYCHOMETRIC STUDY OF THE ROSENBERG SELF-ESTEEM SCALE:
AN INVESTIGATION OF GENDER DIF

by

PATRICIA LOUISE SWENSON
B.A., Simon Fraser University, 1980
M.A., Simon Fraser University, 1986

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 COUNSELLING 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

August 2003

© Patricia Louise Swenson, 2003
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 Counselling Psychology and Special Education

The University of British Columbia
Vancouver, Canada

Date August 29, 2003
Abstract

The present study investigated the psychometric properties of the Rosenberg Self-Esteem Scale (RSES). The RSES is a ten-item self-administered instrument intended to measure global self-esteem. The RSES was administered to a group of 1,443 Canadian university students, including 569 males and 874 females. Professor Alex Michalos, as part of his International Study of Quality of Life, collected the data in 1980. Ages of participants ranged from 17 to 65 years. Mean ages were 21.94 years for males and 22.59 years for females. Coefficient alpha for the sample was 0.86. A two tailed t-test found no significant gender group differences for RSES total test scores. Factor analysis revealed one factor, accounting for 44.8% of variance in test scores. This result is indicative that the RSES is measuring one main variable, and that the RSES is a homogenous uni-dimensional measure. A gender group differential item functioning analysis was conducted using Zumbo's (1999) Logistic Regression method. DIF analyses found that three test items displayed statistically significant DIF (mainly uniform) at a probability level of 0.005 (which is the Bonferroni corrected Type I error rate). Effect sizes were less than the minimum R-squared of 0.130 required for significance according to the Zumbo-Thomas effect sizes for Likert data. Due to the extremely small effect sizes, it was concluded that the DIF findings were not significant.
# Table of Contents

Abstract ........................................................................................................ ii
List of Tables ................................................................................................ iv
List of Figures .............................................................................................. v
Acknowledgements ....................................................................................... vi

Chapter I:  Introduction
  Historical Use of the Rosenberg Self-Esteem Scale ......................... 1
  Development of the RSES ................................................................. 2

Chapter II:  Review of the Literature
  Target Ages for Use of the RSES ....................................................... 4
  Reliability of the RSES ................................................................. 4
  Factorial Validity Structure of the RSES with Different Age Groups ......................................................... 7
  RSES Structure: Global Uni-dimensional versus Two or More Dimensions ........................................ 7
  Evidence of Convergent Validity .................................................... 10
  Gender Differences on RSES Scores ........................................ 15
    No Gender Differences ............................................................ 15
    Group Differences Indicating Males Scored Higher than Females ......................................................... 16
    Variables That May Relate to Gender Differences ............... 17
    Test Item Bias ................................................................. 18

Chapter III:  Methodology
  Participants ....................................................................................... 23
  Instrument ....................................................................................... 23
  Data analyses .................................................................................... 24

Chapter IV:  Results
  Descriptive Statistics ........................................................................ 26
  Reliability ......................................................................................... 26
  Dimensionality of the RSES ......................................................... 27
  Differential Item Functioning (DIF) Analyses ................................ 28

Chapter V:  Discussion
  Psychometric Properties of the RSES for This Sample .................. 31
  Item Level Psychometric Analysis .................................................. 31
  Further RSES Psychometric Properties ....................................... 33
  Areas for Further Research ............................................................. 33

References ................................................................................................. 35
List of Tables

Table 1: Reliability for Studies Involving Various Participant Age Groups..........................5
Table 2: Convergent Validity with Other Global Self-Esteem Scales.................................10
Table 3: Variables that Positively Correlate with Self-Esteem......................12
Table 4: Variables that Negatively Correlate with Self-Esteem.....................14
Table 5: Ages of Participants..............................................23
Table 6: Descriptive Statistics of RSES Scores.................................26
Table 7: Item Analysis and Factor Loadings for the RSES.........................27
Table 8: Uniform and Non-Uniform DIF Analyses...............................30
List of Figures

Figure 1: Rosenberg Self-Esteem Scale........................................3

Figure 2: Scree Plot of Factorial Analysis of RSES Scores............. 28
Acknowledgements

I would like to sincerely thank Dr. Bruno D. Zumbo for his guidance, encouragement, and statistical expertise throughout all stages of this study. It is due to his mentorship and unwavering commitment to teaching excellence that I have been able to succeed. I would also like to thank my committee member, Dr. Anita Hubley, particularly for her detailed feedback on all aspects of this thesis. I would also like to acknowledge Dr. Susan James for being the external examiner, and providing valuable insights.

I am fortunate to have had the ongoing encouragement of family and many friends throughout my academic pursuits, and particularly thank my daughter Missy Stone and friend Lionel Webb for their unconditional support.

Finally, I would like to give thanks to Professor Alex Michalos for making available the data from his International Study of Quality of Life.
Chapter I
Introduction

The purpose of this thesis was to review the psychometric properties of the Rosenberg Self-Esteem Scale (RSES). This study examined reliability based on internal consistency of the total test and item scores, homogeneity and validity of the test based on factor analyses, gender group differences based on t-tests, and whether there is evidence of differential item functioning related to gender.

Historical Use of the Rosenberg Self-Esteem Scale

The RSES, developed in 1965, is the most commonly used instrument for measuring self-esteem. In a 1991 review of the literature, Blascovich and Tomaka found that out of 5,117 research studies on self-esteem, 19 different scales had been used, and 1,285 of the studies had used the RSES. A recent review of research studies reported in Academic Search Premier, ERIC and PsychINFO revealed that over the past twenty years the RSES has been adapted for use with a wide range of groups from all over the world. The RSES has been administered to ethnic groups in North America, Europe, Asia, Australia, Africa and the mid East, involving translations into many different languages including: Finnish, French, German, Italian, Spanish, Estonian, Persian, Chinese and Japanese. The RSES has also been translated into American Sign Language and a computerized version of the original paper and pencil test has been developed. The RSES was originally developed for use with adolescents, but has since been administered to a wide range of age groups, including intermediate elementary school students, adolescents, university students, adults of all ages, and elderly persons to over 100 years of age. Research has examined application of the RSES with a diversity of groups,
including persons with disabilities, physical conditions such as arthritis or tuberculosis, eating disorders, diagnosed mental disorders including depression and anxiety, drug addictions, as well as groups of elite athletes and high achievers. The research literature has continued to examine how psychometric properties of the RSES, including reliability and validity, have been maintained throughout the expansion of its application.

**Development of the RSES**

In developing the RSES, Rosenberg drew upon his prior experience with large-scale sample survey techniques, which had been used to assess opinions. He hypothesized the same methodology could be used to measure positive and negative attitudes toward the self. The RSES was initially developed to measure adolescents' global feelings of self-worth or self-acceptance. In doing so, Rosenberg (1989) ensured that key criteria were met. The instrument needed to be easy to administer to large samples, take no longer than a few minutes to complete, be a uni-dimensional measure so that respondents could be ranked along a single continuum based on total score, have face validity indicating measurement of one construct, and lead to reproducible results.

The scale includes ten items: five positive statements and five negative statements. Positive and negative items are mixed together to minimize respondent set. Each item has a number of options to choose from. Out of 40 recently reviewed research studies, 30 reported using four response choices. An example of four responses is: strongly agree, agree, disagree or strongly disagree. Respondents are instructed to choose the one response for each item that most closely resembles them. For data analyses, negatively worded items are reverse scored, so that high and low scores on both positive and negative items have the same meaning. With four possible choices per item, total test
scores range from 10 to 40, with higher scores indicating higher self-esteem. The RSES requires no more than a grade 5 reading level and can be administered in a few minutes (Gray-Little, Williams & Hancock, 1997). Figure 1 lists the RSES test items.

Figure 1

*Rosenberg Self-Esteem Scale*

1. I feel that I am a person of worth, at least on an equal plane with others.
2. I feel that I have a number of good qualities
3. All in all, I am inclined to feel that I am a failure.
4. I am able to do things as well as most other people.
5. I feel I do not have much to be proud of.
6. I take a positive attitude toward myself.
7. On the whole, I am satisfied with myself.
8. I wish I could have more respect for myself.
9. I certainly feel useless at times.
10. At times I think I am no good at all.
Chapter II

Review of the Literature

The present study on the psychometric properties of the RSES was developed after consideration of research findings previously published on this scale. In the following section, some of the previous research findings are summarized where the focus has been on the measurement of psychometric properties such as reliability, validity, dimensionality of the scale, and gender group differences in relation to possible test bias.

Target Ages for Use of the RSES

Rosenberg developed the RSES for use with adolescents. The scale was developed with a standardized sample of 5,024 adolescents from high schools in New York State (Bracken & Mills, 1994). Bracken and Mills noted that the RSES manual does not indicate any intended uses or applications for the instrument, including no intended age group. The scale has been used with all age groups.

Reliability of the RSES

Research studies have historically reported high levels of reliability for the RSES, internal and test-retest, since its development in 1965. Reliability has been consistently high for diverse groups including adolescents, students, or adults with physical illnesses, mental illnesses, drug addictions, low literacy, and for high achievers or with various translations of the test into other languages. Reliability has also continued to be consistently high throughout administration of the RSES to a wide range of age groups. Reliability, as reported in 23 recent research studies involving a wide range of subject age groups, is summarized in Table 1.
### Table 1

**Reliability for Studies Involving Various Participant Age Groups**

<table>
<thead>
<tr>
<th>Source</th>
<th>N</th>
<th>Ages</th>
<th>Internal consistency</th>
<th>Test-retest</th>
</tr>
</thead>
<tbody>
<tr>
<td>Verkuyten (2003)</td>
<td>1070</td>
<td>10-13 year olds</td>
<td>.76*</td>
<td></td>
</tr>
<tr>
<td>Hagborg (1996)</td>
<td>120</td>
<td>grade 5-8 students</td>
<td>.84</td>
<td></td>
</tr>
<tr>
<td>Marcotte et al. (2002)</td>
<td>547</td>
<td>11-18 year olds</td>
<td>.88</td>
<td></td>
</tr>
<tr>
<td>Lane et al. (2002)</td>
<td>91</td>
<td>11-21 year olds</td>
<td>.82</td>
<td></td>
</tr>
<tr>
<td>Yarcheski et al. (2003)</td>
<td>148</td>
<td>12-14 year olds</td>
<td>.87</td>
<td></td>
</tr>
<tr>
<td>Bagley et al. (1997)</td>
<td>2108</td>
<td>12-19 year olds</td>
<td>.85-.90</td>
<td></td>
</tr>
<tr>
<td>Francis &amp; Wilcox (1995)</td>
<td>802</td>
<td>16 year olds</td>
<td>.59</td>
<td></td>
</tr>
<tr>
<td>Feather (1991)</td>
<td>953</td>
<td>grade 11 students</td>
<td>.83</td>
<td></td>
</tr>
<tr>
<td></td>
<td>361</td>
<td>undergraduates</td>
<td>.86</td>
<td></td>
</tr>
<tr>
<td>Brems &amp; Lloyd (1995)</td>
<td>216</td>
<td>undergraduates</td>
<td>.91</td>
<td></td>
</tr>
<tr>
<td>Feather (1998)</td>
<td>186</td>
<td>undergraduates</td>
<td>.87</td>
<td></td>
</tr>
<tr>
<td></td>
<td>114</td>
<td>undergraduates</td>
<td>.81</td>
<td></td>
</tr>
<tr>
<td></td>
<td>310</td>
<td>undergraduates</td>
<td>.87</td>
<td></td>
</tr>
<tr>
<td>Gray-Little et al. (1997)</td>
<td>1234</td>
<td>undergraduates</td>
<td>.88</td>
<td></td>
</tr>
<tr>
<td>Salmela-Aro &amp; Nurmi (1997)</td>
<td>256</td>
<td>undergraduates</td>
<td>.85</td>
<td></td>
</tr>
<tr>
<td>Vispoel et al. (2001)</td>
<td>224</td>
<td>university students</td>
<td>.96**</td>
<td></td>
</tr>
<tr>
<td>Source</td>
<td>N</td>
<td>Ages</td>
<td>Internal consistency</td>
<td>Test-retest</td>
</tr>
<tr>
<td>---------------------------------------------</td>
<td>-------</td>
<td>---------------------</td>
<td>----------------------</td>
<td>--------------</td>
</tr>
<tr>
<td>Brumfitt &amp; Sheeran (1999)</td>
<td>243</td>
<td>university students</td>
<td>.87</td>
<td></td>
</tr>
<tr>
<td></td>
<td>14</td>
<td>adults</td>
<td>.91</td>
<td></td>
</tr>
<tr>
<td>Banos &amp; Guillen (1999)</td>
<td>266</td>
<td>adults</td>
<td>.88</td>
<td></td>
</tr>
<tr>
<td>Kingree (1995)</td>
<td>196</td>
<td>adults</td>
<td>.78</td>
<td></td>
</tr>
<tr>
<td>Westaway &amp; Wolmarans (1992)</td>
<td>100</td>
<td>adults</td>
<td>.78</td>
<td></td>
</tr>
<tr>
<td>White &amp; Schweitzer (2000)</td>
<td>88</td>
<td>adults</td>
<td>.87-.89</td>
<td></td>
</tr>
<tr>
<td>Hojat &amp; Lyons (1998)</td>
<td>212</td>
<td>university students</td>
<td>.78</td>
<td>.72</td>
</tr>
<tr>
<td></td>
<td>12</td>
<td>sub-sample (2 week retest)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Robins et al. (2001)</td>
<td>208</td>
<td>undergraduates (5/7 responses)</td>
<td>.89/.93</td>
<td></td>
</tr>
<tr>
<td></td>
<td>66</td>
<td>adults</td>
<td>.87</td>
<td></td>
</tr>
<tr>
<td></td>
<td>508</td>
<td>undergraduates (5 retests at ends of semester 1 &amp; years 1,2,3&amp;4)</td>
<td>.88-.90</td>
<td></td>
</tr>
<tr>
<td>Salyers et al. (2001)</td>
<td>202</td>
<td>adults (2 week retest)</td>
<td>.84</td>
<td>.80</td>
</tr>
<tr>
<td>Henriques &amp; Calhoun (1999)</td>
<td>395</td>
<td>university students</td>
<td></td>
<td>.88/.69</td>
</tr>
<tr>
<td></td>
<td></td>
<td>(Whites/Blacks-1 week retest)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Pullman &amp; Allik (1999)</td>
<td>616</td>
<td>197 undergraduates (2 week retest)</td>
<td>.84</td>
<td>.67/.62</td>
</tr>
<tr>
<td></td>
<td>419</td>
<td>adults (5 month/1 year retests)</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Note. * Positive RSES items only  ** Paper/pencil & computerized versions of RSES
Factorial Validity Structure of the RSES with Different Age Groups

Whiteside-Mansell and Corwyn (2003) questioned the validity of using the RSES with such a variety of differing age groups. They compared RSES scores for a group of 414 adolescents (aged 12-17 years) to a group of 900 adults (aged 18-82) and concluded, “Strong tests of factorial invariance indicted that the use of the scale was supported and comparable for adolescents and adults. Follow up analyses indicated that there were not mean differences on the latent construct between the two groups.” (p. 169). Although still widely used with adolescents, information as summarized in Table 1 reveals the scale has also become commonly used with groups of adults.

RSES Structure: Global Uni-dimensional versus Two or More Dimensions

Rosenberg (1989) intended the RSES to be a global measure of general self-esteem. Rosenberg suggested high self-esteem indicates a person respects and considers him/herself worthy, does not consider him/herself superior to others, recognizes self-limitations, and expects to grow and improve. Low self-esteem implies self-rejection or self-contempt, feeling disagreeable about one self and wishing it were otherwise.

In proposing that the RSES be a global rather than a multi-dimensional measure, Rosenberg (1979) explained, “…self-concept is not a collection but an organization of parts, pieces, and components and that these are hierarchically organized and interrelated in complex ways.” (p.73). Rosenberg, Schooler, Schoenbach and Rosenberg (1995) described global self-esteem to be an individual’s positive or negative attitude toward the self as a totality, which is strongly related to overall psychological well being. They differentiated this from specific self-esteem, which they said relates more closely to, and is a better predictor of, some related specific behavior. Rosenberg et al. argued global
and specific self-esteem have much in common, but the concepts should not be interchanged, noting they are often interchanged in the literature.

Based on exploratory and confirmatory factor analyses, researchers have frequently reported statistical verification that a single factor accounts for significant variance in RSES scores, supporting a uni-dimensional global model of self-esteem (Bagley, Bolitho & Bertrand, 1997; Pullman & Allik, 1999; Shevlin, Bunting & Lewis, 1995; Tomas & Oliver, 1999).

Researchers have also argued in support of a uni-dimensional global structure of the RSES based on high correlations between RSES scores and total test scores of other instruments that are intended to measure global self-esteem (Brems & Lloyd, 1995; Brumfitt & Sheeran, 1999; Francis & Wilcox, 1995; Griffiths, et al. 1999; McCurdy & Kelly, 1997; Silber & Tippett, 1965).

Crocker and Algina (1986) noted that high correlations among test item scores can be interpreted as evidence of uni-dimensionality, indicating that the test measures a single trait; however, such statistical dependence can just as easily be indicative of two different traits which are highly correlated with each other. In making inferences, therefore, assumptions underlying dimensionality and validity must be scrutinized.

Ongoing debate continues in the research literature as to whether the RESE actually measures one variable or more. A common alternate hypothesis is that two variables are represented: self-confidence represented by the positively stated items of the instrument, and self-deprecation represented by the negatively stated items. Research has supported this hypothesis, using exploratory and confirmatory factor analyses.
(Owens, 1993, 1994; Shahini, Dipboye & Phillips, 1990) and structural equation modeling (Sheasby, Barlow, Cullen & Wright, 2000).

Other researchers have argued that, although statistical evidence has indicated that variance in RSES scores could be accounted for by two relatively independent factors, these findings were most likely due to method effects rather than the existence of two different dimensions of self-esteem in the RSES. Method effects involve patterns of response bias or response set found for tests that contain positively and negatively worded test items. Thomas and Oliver (1999) found that method effects were associated with responses to item wording, especially for negatively worded items, on a Spanish translation of the RSES. Wang, Siegal, Falck and Carlson (2001) identified method effects in an administration of the RSES to drug addicts, associated primarily with positively rather than negatively worded items. This, they explain, was likely due to interactions between characteristics of this particular group of respondents and the positive statements regarding the self.

Regardless of whether method effects relate to negatively or positively stated items, these researchers have pointed out that, although statistical evidence is the necessary first step in analyzing factor outcomes, a comprehensive analysis of results is necessary prior to making valid conclusions regarding factorial structure of the RSES.

Researchers have also concluded that, based on factor analyses of RSES data, self-esteem or self-concept scores as measured by the RSES, appear to be multidimensional (Byrne & Shavelson, 1987).
Evidence of Convergent Validity

Researchers have compared the scores of groups tested with both the RSES and with other scales that were intended to measure global self-esteem. Correlations among subjects’ scores on the tests have been calculated to examine whether there is evidence of meaningful relationships. Strong positive correlations indicate homogeneity, or that the scales appear to be measuring the same construct. Research results have indicated significant positive relationships between measures of global self-esteem in all the cases reviewed. Research results on these studies are summarized in Table 2, including names of the various global self-esteem scales, correlations between scores on those scales and scores on the RSES, and the sources of each of the research findings.

Table 2

Convergent Validity with Other Global Self-Esteem Scales

<table>
<thead>
<tr>
<th>Self-Esteem Scale</th>
<th>Source</th>
</tr>
</thead>
<tbody>
<tr>
<td>Coopersmith Self-Esteem Inventory</td>
<td>Francis &amp; Wilcox (1995)</td>
</tr>
<tr>
<td></td>
<td>.52</td>
</tr>
<tr>
<td></td>
<td>.60</td>
</tr>
<tr>
<td></td>
<td>.68</td>
</tr>
<tr>
<td>MMPI - 2 Low Self-Esteem Subtest</td>
<td>McCurdy &amp; Kelly (1997)</td>
</tr>
<tr>
<td></td>
<td>-.61</td>
</tr>
<tr>
<td>Single - Item Self-Esteem Scale</td>
<td>Robins et al. (2001)</td>
</tr>
<tr>
<td></td>
<td>.69</td>
</tr>
<tr>
<td>Visual Analogue Self-Esteem Scale</td>
<td>Brumfitt &amp; Sheeran (1999)</td>
</tr>
<tr>
<td></td>
<td>.60</td>
</tr>
</tbody>
</table>
Table 2 (continued)

<table>
<thead>
<tr>
<th>Self-Esteem Scale</th>
<th>Source</th>
</tr>
</thead>
<tbody>
<tr>
<td>RSES (neg items)</td>
<td></td>
</tr>
<tr>
<td>MMPI - 2 Low Self-Esteem Subtest</td>
<td>.69</td>
</tr>
<tr>
<td></td>
<td>Brems &amp; Lloyd (1995)</td>
</tr>
</tbody>
</table>

Research investigations have examined whether the RSES is valid and is actually measuring what it is intended to measure by comparing scores on the RSES to scores on other different tests, which are intended to measure constructs other than global self-esteem. Significant positive correlations between RSES scores and measures of other constructs considered similar to, or congruent with, global self-esteem, provide evidence of convergent validity, and suggest that the scales are measuring similar variables. Examples are studies identifying positive correlations between RSES global self-esteem scores and scores on tests measuring happiness, positive affect, self-worth, positive self-attitude, confidence, high self-efficacy, life satisfaction and pleasant affect. Results of findings from a number of recent studies comparing RSES scores to scores on scales intended to measure constructs which are understood to be positively related to global self-esteem are summarized in Table 3. Information displayed in Table 3 includes the sources of the research studies, the various traits that were being measured by the other scales, and the significant correlations between scores on those scales and scores on the RSES. All research findings reported evidence of statistically significant positive relationships and convergent validity.
### Table 3
*Variables that Positively Correlate with Self-Esteem*

<table>
<thead>
<tr>
<th>Source</th>
<th>Variable</th>
<th>r</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Positive affect</td>
<td>.33</td>
</tr>
<tr>
<td></td>
<td>Balanced affect</td>
<td>.38</td>
</tr>
<tr>
<td></td>
<td>Extraversion</td>
<td>.39</td>
</tr>
<tr>
<td>Hale et al. (1992)</td>
<td>Generalized expectancy for success</td>
<td>.46</td>
</tr>
<tr>
<td></td>
<td>Life orientation</td>
<td>.58</td>
</tr>
<tr>
<td></td>
<td>Internal locus of control</td>
<td>.24</td>
</tr>
<tr>
<td>Hojat &amp; Lyons (1998)</td>
<td>Think worthy of self</td>
<td>.73</td>
</tr>
<tr>
<td></td>
<td>Positive self attitudes</td>
<td>.61</td>
</tr>
<tr>
<td>Lorr &amp; Wunderlich (1986)</td>
<td>Confidence</td>
<td>.65-.69</td>
</tr>
<tr>
<td></td>
<td>Popularity</td>
<td>.39-.43</td>
</tr>
<tr>
<td></td>
<td>Social assertiveness</td>
<td>.36</td>
</tr>
<tr>
<td></td>
<td>Pleasant affect (self/family-friends ratings)</td>
<td>.45/.39</td>
</tr>
<tr>
<td>Yaniko &amp; Lu (2000)</td>
<td>Likeability</td>
<td>.36</td>
</tr>
<tr>
<td></td>
<td>Task accomplishment</td>
<td>.30</td>
</tr>
<tr>
<td></td>
<td>Giftedness</td>
<td>.42</td>
</tr>
<tr>
<td></td>
<td>Power</td>
<td>.37</td>
</tr>
</tbody>
</table>
Negative correlations between RSES scores and measures of other constructs considered dissimilar to, or at variance with, global self-esteem, also provide evidence of convergent validity, further indicating that RSES is measuring what it intended to measure and not what the other tests are measuring. Examples are studies that compare RSES global self-esteem scores to scores on tests measuring variables such as depression, anxiety, ineffectiveness, lack of satisfaction of self, feeling useless, loneliness, high sensitivity to teasing and feelings of vulnerability. Results of findings from a number of recent studies comparing RSES scores to measures of constructs understood to be dissimilar to global self-esteem are summarized in Table 4. Information displayed in Table 4 includes the sources of the research studies, the various traits, which were being measured by the other scales, and the significant correlations between scores on those scales and scores on the RSES. In all cases there was evidence of statistically significant negative correlations.
Table 4

*Variables that Negatively Correlate with Self-Esteem*

<table>
<thead>
<tr>
<th>Source</th>
<th>Trait</th>
<th>r</th>
</tr>
</thead>
<tbody>
<tr>
<td>Brumfitt &amp; Sheeran (1999)</td>
<td>Depression</td>
<td>-.53 to -.47</td>
</tr>
<tr>
<td></td>
<td>Anxiety</td>
<td>-.42 to -.50</td>
</tr>
<tr>
<td>Griffiths et al. (1999)</td>
<td>Depression</td>
<td>-.73</td>
</tr>
<tr>
<td></td>
<td>Ineffectiveness</td>
<td>-.66</td>
</tr>
<tr>
<td>Hojat &amp; Lyons (1998)</td>
<td>Not satisfied with self</td>
<td>-.68</td>
</tr>
<tr>
<td></td>
<td>Think not well of self</td>
<td>-.60</td>
</tr>
<tr>
<td>Hojat &amp; Lyons (1998)</td>
<td>Feel useless</td>
<td>-.53</td>
</tr>
<tr>
<td></td>
<td>Loneliness</td>
<td>-.49</td>
</tr>
<tr>
<td></td>
<td>Test anxiety</td>
<td>-.22</td>
</tr>
<tr>
<td></td>
<td>General anxiety</td>
<td>-.44</td>
</tr>
<tr>
<td>Schimmack &amp; Diener (2003)</td>
<td>Unpleasant affect (self/family &amp; friends rating)</td>
<td>-.40/- .36</td>
</tr>
<tr>
<td>Thompson et al. (1995)</td>
<td>Perception of weight teasing effect</td>
<td>-.27</td>
</tr>
<tr>
<td></td>
<td>Perception of competency teasing effect</td>
<td>-.20</td>
</tr>
<tr>
<td>Westaway &amp; Walmarans (1992)</td>
<td>Depression</td>
<td>-.54 to -.56</td>
</tr>
<tr>
<td>Wilson &amp; Lavelle (1989)</td>
<td>Loneliness</td>
<td>-.32 to -.58</td>
</tr>
<tr>
<td></td>
<td>Depression</td>
<td>-.32 to -.58</td>
</tr>
<tr>
<td></td>
<td>Anxiety</td>
<td>-.32 to -.58</td>
</tr>
<tr>
<td>Yaniko &amp; Lu (2000)</td>
<td>Vulnerability</td>
<td>-.49</td>
</tr>
</tbody>
</table>
Studies have reported significant statistical relationships between RSES scores and other important life events. Yarcheski, Mahon and Yarcheski (2003) found self-esteem was positively related to positive health practices, noting, however, self-esteem was not as powerful as social support in relating to health practices of early adolescents. Shahini et al. (1990) found global self-esteem was significantly related to 12 variables associated with work-related attitudes, including: regard of client, job pressure, recognition of results, clarity of structure, job tension, work satisfaction, pay satisfaction, co-worker satisfaction, organizational commitment, how long workers intended to stay with the organization, how long they had intended to stay when they accepted the job, and whether they were looking for other jobs. Salmela-Aro and Nurmi (1997) found low self-esteem was a predictor that young adults were likely to develop self-related goals whereas high self-esteem predicted a likelihood of becoming interested in family-related goals.

**Gender Differences on RSES Scores**

For the research studies currently reviewed, investigations into whether there are significant gender differences among RSES total test scores has resulted in one of two conclusions. Either no significant gender differences have been detected or males have been found to score higher than females.

**No Gender Differences**

A number of studies involving respondent groups of different ages and with varying characteristics, have reported finding no significant gender differences. Hagborg (1996) studied 120 students, 15 females and 15 males at each of four grade levels, 5 through 8, and found no significant gender differences. Abu-Saad (1999) studied 1560
grade 11 and 12 Israeli-Arab adolescents, 45% female and 55% male, finding gender differences were not significant in accounting for variance in scores. In two studies involving university students, Swanson and Lease (1990) found RSES scores were almost identical for 59 females and 53 males based on results of a t-test, and in a study comparing 161 females and 36 males, Pullmann and Allik (2000) found no significant differences based on results of a one-way ANOVA grouped for gender. McCurdy and Kelly (1997) compared 82 female and 33 male adults, and found no significant gender differences. Banos and Gullen (2000) examined gender differences for groups of adults, including 34 females and 18 males diagnosed with social phobias, and 135 females and 79 males who comprised a control group with no phobias, and found no significant gender differences. White and Schweitzer (2000) found no significant gender differences for a group of 88 adults including 44 with chronic fatigue syndrome.

Group Differences Indicating Males Scored Higher than Females

In all studies where significant gender group differences on RSES total test scores were detected, males were found to score higher than females. These studies have included groups sampled from many different cultures and age groups.

The following studies detecting significant gender differences in favor of males indicate the consistency of findings across diverse cultural groups. Kawabata et al. (1999) compared 999 female and 1089 male grade 4-9 students in Japan, noting the differences increased with age. Marcotte et al. (2002) compared 279 female and 268 male Quebec French speaking adolescents aged 11-18 years old. Moran and Eckenrode (1991) compared 70 female and 48 male grade 7-11 students in New York. Bagley et al. (1997) compared 1024 female and 1084 male high school students in Alberta (also

Among the studies reviewed involving RSES total test scores of adults, Henriches and Calhoun (1999) found a group of 218 female university students scored lower than 177 males. Kendler, Gardner and Prescott (1998) compared 3793 twin pairs, including 856 female-female pairs, 1517 male-male pairs and 1420 female-male pairs, and found there was a small significant gender difference favoring males. Kingree (1995) studied 41 female and 155 male substance abusers in long term treatment, finding males had higher RSES scores.

**Variables That May Relate to Gender Differences**

Where significant gender differences have been detected, speculation has been made as to what issues may be interacting with global self-esteem in ways that put females at a disadvantage for scoring as high as males on the RSES.

Leonardi et al. (1998) noted that adolescent females tend to score lower than males, suggesting females', "...disbelief in their capabilities seems to be shaped by the family, the educational system, the mass media and the culture at large" (p. 158). Kingree (1995) found gender differences favoring males were statistically linked to female reports of less family support, more self-blame and more parental blame than
males. Moran and Eckenrode (1991) noted males appeared to gain greater social benefits and pay fewer costs from peer relationships than females, and suggested gender differences favoring males on RSES scores were a reflection of the impact of these differences. Verkuyten (2003) noted general developmental and societal factors are important factors for self-esteem scores. He suggested lower self-esteem among females might relate to different factors depending upon culture. For example, women from minority cultures may be coping with more traditional gender roles, and majority women from Europe and North America may experience more pressure regarding appearance. Bagley et al. (1997) suggested gender differences might relate to females answering RSES items differently than males, "...since the response style of females to self-esteem questions may be less egoistic, and more self-deprecating in ways which imply that relationship styles rather than self-aggrandizement are more important" (p.89). Kendler et al. (1998) suggested that biological differences between genders may affect self-esteem scores. For example, factors such as menstruation may have a negative effect on self-esteem.

Test Item Bias

Zumbo and Hubley (2003) pointed out that gender differences have historically been a focus of test bias studies. They noted that given tests are comprised of items, it is reasonable to consider test bias in terms of differential item functioning (DIF). Zumbo and Hubley explain that DIF allows differentiation between group differences that are due to item impact, indicating true group differences on the underlying ability being measured, and group differences due to item bias.
Zumbo (1999) explained, "...DIF is a necessary but not sufficient, condition for item bias. Thus, if DIF is not apparent for an item, then no item bias is present. However, if DIF is apparent, then its presence is not sufficient to declare item bias; rather, one would have to apply follow-up item bias analyses (e.g., content analysis, empirical evaluation) to determine the presence of item bias" (p.12).

While some group differences on test scores do reflect test bias, other group differences relate to actual differences in ability on the variable being measured. If a test item is an accurate representation of the construct that the test is intended to measure, then differential performance by groups on that item can reflect a true difference between the groups (Gierl, Bisanz, Bisanz, Boughton & Khaliq, 2001). Performance differences occur frequently in test and item data because individuals do differ with respect to abilities measured by items and tests. Dorans and Holland (1993) refer to these true differences as impact, which is distinct from DIF. In cases where true group differences in ability exist, researchers are often able to develop hypotheses, usually regarding antecedent circumstances, which explain reasons why one group may score differently than another group.

On the other hand, some group differences on test and test item scores have been attributed to test biases. Individual test items have been found to be biased in favor of, or against, particular groups of test takers, for example when non-essential characteristics of the item (such as vocabulary incidental to what the test is intended to measure) are less familiar to members of one group of test takers than they are to another group (Linn & Harnisch, 1981). The interaction of test item characteristics with characteristics of test takers can affect performance. If answering positively or correctly requires familiarity
with something irrelevant to what the test is intended to measure, test scores will vary depending on familiarity with that extraneous information.

In order for an assessment to be valid, the items included on the test must adequately represent the construct that the test is intended to measure. As well as including adequate and meaningful representation of what the test is intended to measure, the test should be homogeneous; that is, it should not include information unrelated to what is being assessed. If construct irrelevant information is also included, the test is measuring more than one thing.

It is assumed that individuals matched on total test scores should have equal probabilities of answering individual test items the same way. DIF studies typically compare a focal group of interest and a matched reference group, based on variables such as gender, language proficiency or ethnicity. A DIF statistical analyses, “represents the average factor by which the odds that a reference group (majority) member gets an item correct exceeds the corresponding odds for a comparable focal (minority) group member” (Clauser, Mazor & Hambleton, 1991, p. 354). It is widely accepted that, “an item is differentially functioning (DIF) if examinees of the same ability level, but belonging to different subgroups, have different probabilities of selecting the correct response” (Clauser et. al., 1991, p. 353).

It is important that the groups being compared are matched on the variable being assessed. The total test score is normally considered a valid measure of the variable for DIF analyses. DIF statistical analyses provide the degree and direction of DIF, but not an explanation of the cause.
There are non-statistical and statistical methods of detecting bias at the item level. A non-statistical method for detecting biases on a scale such as the RSES would involve examination of test items by content experts to analyze whether it may be easier for one group than another to give a positive rating on the item. Zumbo and Hubley (2003) outlined three statistical frameworks, which have been developed in the literature to detect test item biases: modelling item responses via contingency tables and/or regression models; item response theory, and multidimensional models.

Zumbo and Hubley (2003) noted each of the statistical methods to detect DIF could be applied in a confirmatory or exploratory manner. They explained that multidimensional approaches to DIF are traditionally exploratory, involving statistical detection of group item differences, which are subsequently studied by content specialists to identify whether or not there is evidence of impact or bias and propose possible reasons for this. The alternative confirmatory approach involves a theory-based approach in which an explanation of why DIF may be present is based on the previous research literature and is then tested using statistical methods.

Messick (1996) argued that test validity is not just a property of the assessment, but also of the meaning and interpretation of the test scores. He added that the extent to which meanings and outcomes of test scores hold across population groups and contexts is a continuing empirical question and this is the reason why validity is a continuing process, not a static test characteristic. Validity is an issue not only of test content representation, but also of data interpretation and resulting consequences. Fairness extends beyond valid interpretations and includes concerns about social policy and the outcomes of interpretations of test data (Bond, Moss & Carr, 1996). Since self-esteem
tests are used on a wide basis, validity and fairness are essential. If a test is invalid for one group and interpretation of the scores results in consequences that deny that group some benefit (such as treatment or access to a program), fairness has been denied.

The above noted review of research findings related to gender differences on the RSES has revealed that previous studies have focused only on gender differences for RSES total test scores, not for RSES test item scores. Clearly what is missing are gender DIF analyses. DIF analyses are the most powerful and appropriate way to examine gender differences because these procedures match participants on the variable of interest, before checking for gender differences.

The present study will examine psychometric properties of the RSES scores for a sample of 1443 Canadian university students. Gender group analyses will include comparisons of average total test scores using two tailed independent samples t-tests, as well as DIF analyses of test item scores using Zumbo’s (1999) Logistic Regression. Dimensionality of total test scores based on exploratory factorial analysis will be conducted to confirm homogeneity of the RSES for this sample, as homogeneity is an assumption of DIF analysis, where total test scores are used for matching the gender groups.
Chapter III
Methodology

Participants

Participants included in this study were a sample of English speaking Canadian university students. Data were collected in 1980 by Professor Alex Michalos as part of his International Study of Quality of Life. The RSES was administered to a total of 1443 participants, including 569 males and 874 females.

Age data was collected on 1415 of the participants, including 562 males and 853 females. Descriptive statistics on the ages of participants are displayed in Table 5.

Table 5

Ages of Participants

<table>
<thead>
<tr>
<th></th>
<th>Mean</th>
<th>Std. Dev.</th>
<th>N</th>
<th>Min</th>
<th>Max</th>
</tr>
</thead>
<tbody>
<tr>
<td>Males</td>
<td>21.94</td>
<td>5.727</td>
<td>562</td>
<td>17</td>
<td>64</td>
</tr>
<tr>
<td>Females</td>
<td>22.59</td>
<td>7.140</td>
<td>853</td>
<td>17</td>
<td>65</td>
</tr>
<tr>
<td>Overall</td>
<td>22.33</td>
<td>6.621</td>
<td>1415</td>
<td>17</td>
<td>65</td>
</tr>
</tbody>
</table>

Based on results of a two tailed independent samples t-test, no significant gender group differences were identified for age, $t(1413) = 1.83, p = 0.068$.

Instrument

As noted above, the Rosenberg Self-Esteem Scale is a ten item self-administered instrument intended to measure global self-esteem. The RSES includes five positively worded items, and five negatively worded items. Positive and negative items are mixed
together to minimize respondent set. For this study, items were administered in the order presented in Figure 1. Each item had four possible response choices: (1) strongly agree, (2) agree, (3) disagree and (4) strongly disagree. For each item, respondents were asked to choose the one response choice that most closely resembled themselves.

Data Analyses

For data analyses, RSES responses for negatively worded items were reversed scored so that high and low scores on positively and negatively worded items had the same meaning. With four possible choices for each of the ten items, total test scores ranged from 10 to 40. RSES total test raw scores and item raw scores were obtained.

All statistical calculations were conducted using the Statistical Package for the Social Sciences, Version 11. Statistical analyses were conducted for the total sample of 1443, including 569 males and 874 females, except for calculations involving age, in which case data was available on 1415 participants, including 562 males and 853 females.

Analyses included gender comparisons for age and total test scores based on two tailed independent samples t-tests, item discrimination based on corrected item-total test score correlations, dimensionality of total test scores based on exploratory factor analyses, and gender DIF using Zumbo’s (1999) Logistic Regression.

A detailed description of the Logistic Regression DIF method can be found in Zumbo (1999) or Gelin and Zumbo (2003). Zumbo (1999) explained that, for ordinal item scores, Logistic Regression analysis involves entering variables in three steps. Step #1 involves entering the conditioning variable (total test scores). Step #2 involves entering the grouping variable (in this case, gender). Step #3 involves entering the
interaction term (total test score*group). As Gelin and Zumbo (2003) noted, the model has one dependent variable (item response) and three independent variables (total test scores, gender groupings and the interaction of total and gender).

A Chi-square statistic for Step #3 is obtained and subtracted from the Chi-square value for Step #1. The resulting Chi-square statistic has two degrees of freedom and is a simultaneous measure of uniform and non-uniform DIF. The two degrees of freedom occur because they are the difference between the three degrees of freedom at Step #3 and the one degree of freedom at Step #1 (Gelin & Zumbo, 2003). Zumbo (1999) further explained that this three step sequential strategy for analysis enables comparison of variation in item scores due to the grouping plus interaction variables (DIF statistic) over and above variance due to the total test scores (uniform DIF).

Zumbo (1999) explained the model also allows effect size comparisons of each of the three sources of variance as, at each step, an R-squared is calculated. In order for a test item to be considered DIF, the DIF statistic must be at a statistically significant level and the effect size must be sufficient. Probability values for DIF statistics can be determined using a Chi-square probability table (Gelin & Zumbo, 2003). According to the Zumbo-Thomas effect size required for significance for Likert data (Zumbo, 1999), effect sizes are reasonable with an R-square value of at least 0.130.
Chapter IV

Results

Descriptive Statistics

Descriptive statistics for the total test scores for each gender and the overall sample are displayed in Table 6. Based on the results of a two-tailed independent samples t-test, gender group differences for total test scores were not significant, t(1441) = 1.29, p = .197.

Table 6

Descriptive Statistics of RSES Scores

<table>
<thead>
<tr>
<th></th>
<th>Mean</th>
<th>Std. Dev.</th>
<th>N</th>
<th>Min</th>
<th>Max</th>
</tr>
</thead>
<tbody>
<tr>
<td>Males</td>
<td>31.2</td>
<td>4.70</td>
<td>569</td>
<td>13</td>
<td>40</td>
</tr>
<tr>
<td>Females</td>
<td>30.9</td>
<td>4.89</td>
<td>874</td>
<td>14</td>
<td>40</td>
</tr>
<tr>
<td>Overall</td>
<td>31.0</td>
<td>4.82</td>
<td>1443</td>
<td>13</td>
<td>40</td>
</tr>
</tbody>
</table>

The correlation between age and total test score was r = 0.13 (p = 0.01). This indicates that total test scores tended to increase with age. Although this correlation is statistically significant, the result does not provide any particularly valuable insight.

Reliability

Reliability of RSES scores for the present sample (N = 1443) was high, with a coefficient alpha of 0.86 (for males the reliability was 0.84 and for females it was 0.87). Item-total test score correlations are displayed in Table 7.
Table 7

*Item Analysis and Factor Loadings for the RSES*

<table>
<thead>
<tr>
<th>Item #</th>
<th>Item</th>
<th>Corrected Item-Total Correlation</th>
<th>Factor Loading</th>
</tr>
</thead>
<tbody>
<tr>
<td>1P</td>
<td>I feel that I am a person of worth, at least on an equal basis with others.</td>
<td>.497</td>
<td>.614</td>
</tr>
<tr>
<td>2P</td>
<td>I feel that I have a number of good qualities.</td>
<td>.460</td>
<td>.576</td>
</tr>
<tr>
<td>4P</td>
<td>I am able to do things as well as most people.</td>
<td>.495</td>
<td>.606</td>
</tr>
<tr>
<td>6P</td>
<td>I take a positive attitude toward myself.</td>
<td>.690</td>
<td>.782</td>
</tr>
<tr>
<td>7P</td>
<td>On the whole, I am satisfied with myself.</td>
<td>.636</td>
<td>.734</td>
</tr>
<tr>
<td>3N</td>
<td>All in all, I am inclined to feel that I am a failure.</td>
<td>.599</td>
<td>.688</td>
</tr>
<tr>
<td>5N</td>
<td>I feel I do not have much to be proud of.</td>
<td>.547</td>
<td>.646</td>
</tr>
<tr>
<td>8N</td>
<td>I wish I could have more respect for myself.</td>
<td>.601</td>
<td>.687</td>
</tr>
<tr>
<td>9N</td>
<td>I certainly feel useless at times.</td>
<td>.553</td>
<td>.632</td>
</tr>
<tr>
<td>10N</td>
<td>At times I think I am no good at all.</td>
<td>.621</td>
<td>.698</td>
</tr>
</tbody>
</table>

Note. P=positively worded item; N=negatively worded item

**Dimensionality of the RSES**

Principle component analysis was conducted on RSES scores for this sample. A factor analysis of the RSES scores resulted in identification of one large Eigenvalue, accounting for 44.8% of the variance. These results are indicative that the RSES is essentially measuring one main variable, which can support an inference the test is unidimensional. This is important, as unidimensionality of RSES total test scores is an
assumption for gender DIF analyses, where total test scores are using as the matching variable for the groups. Factor loadings for the ten items are displayed in Table 7. The data is displayed in Figure 2 using a Scree Plot.

Figure 2

Scree Plot of Factorial Analysis of RSES Scores

Differential Item Functioning (DIF) Analyses

Logistic regression DIF statistical analyses were conducted on data for each of the ten items of the RSES for the total sample (N=1443).

For this study, in order for an item to be considered displaying significant DIF, the DIF statistic had to be significant at a probability level of 0.005 or less, which is the Bonferroni corrected Type I error rate. All DIF tests were conducted at this corrected
Type I error rate. Additionally, in order for an item to be considered displaying meaningful DIF, the effect size required was an R-squared of at least 0.130. This is the Zumbo-Thomas effect size required for significance for Likert data, as described by Zumbo (1999). All DIF results have been displayed in Table 8.

Three items displayed DIF (uniform and non-uniform) at a significance level of 0.005. The same three items were examined to see if the DIF was uniform, non-uniform, or both. For all three items, the non-uniform DIF tests were not statistically significant, and the uniform DIF tests were statistically significant. This means that significant gender group differences were similar at all total test scores levels, and no interactions existed between gender groups and total test scores.

DIF analysis results have been displayed in Table 8 for the three test items, which were detected to have statistically significant DIF. Table 8 displays information for each of the three test items, including: the test item numbers and whether each is a P (positive) or N (negative) statement, the wording for each item as stated on the test, and statistical results including chi-square values, degrees of freedom, probability values, and R-squared values for each of the three DIF analyses (Uniform DIF, Non-Uniform DIF, and Uniform plus Non-Uniform DIF).

Statistically, the DIF results appear to be predominantly uniform DIF; however, none of the DIF results were practically significant due to extremely small effect sizes. The conclusion is, therefore, that the RSES does not display gender DIF for this sample.
Table 8

*Uniform and Non-Uniform DIF Analyses*

<table>
<thead>
<tr>
<th>Item# - Item</th>
<th>Chi-Square</th>
<th>DF</th>
<th>p- value</th>
<th>R- squared</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Uniform and Non-Uniform DIF</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>4P - I am able to do things as well as most people.</td>
<td>13.638</td>
<td>2</td>
<td>0.001</td>
<td>0.008</td>
</tr>
<tr>
<td>6P - I take a positive attitude toward myself.</td>
<td>11.185</td>
<td>2</td>
<td>0.004</td>
<td>0.004</td>
</tr>
<tr>
<td>5N - I feel I do not have much to be proud of.</td>
<td>28.442</td>
<td>2</td>
<td>0.001</td>
<td>0.012</td>
</tr>
<tr>
<td><strong>Non-Uniform DIF</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>4P - I am able to do things as well as most people.</td>
<td>0.196</td>
<td>1</td>
<td>0.658</td>
<td>0.001</td>
</tr>
<tr>
<td>6P - I take a positive attitude toward myself.</td>
<td>0.073</td>
<td>1</td>
<td>0.787</td>
<td>0.001</td>
</tr>
<tr>
<td>5N - I feel I do not have much to be proud of.</td>
<td>0.001</td>
<td>1</td>
<td>1.000</td>
<td>0.001</td>
</tr>
<tr>
<td><strong>Uniform DIF</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>4P - I am able to do things as well as most people.</td>
<td>13.442</td>
<td>1</td>
<td>0.001</td>
<td>0.008</td>
</tr>
<tr>
<td>6P - I take a positive attitude toward myself.</td>
<td>11.112</td>
<td>1</td>
<td>0.001</td>
<td>0.004</td>
</tr>
<tr>
<td>5N - I feel I do not have much to be proud of.</td>
<td>28.442</td>
<td>1</td>
<td>0.001</td>
<td>0.012</td>
</tr>
</tbody>
</table>

*Note.* P=positively worded item; N=negatively worded item; DF= degrees freedom
Psychometric Properties of the RSES for This Sample

The present study involves psychometric analyses of RSES data, including reliability, gender group comparisons for total test scores, factor analysis and gender DIF. The high degree of reliability found for the RSES scores for the present sample is consistent with the findings of previous research studies on reliability, as displayed in Table 1. Comparisons of average total test scores for gender groups reveals no significant group differences for the present sample. This finding is consistent with previous research, which either found no significant gender group differences, or males scoring higher than females.

Factor analysis of the current data has resulted in one dominant factor. It would be reasonable to assume from these findings that the RSES is a homogenous measure of global self-esteem, rather than a measure of two or more different variables. This is consistent with the original intent of Rosenberg in developing the measure.

With the detection of one dominant factor in the RSES data, the assumption of homogeneity of the test has been met. This assumption is necessary for DIF analysis, since the total test scores are used as the matching variable for subjects. In DIF analyses, individual test item scores are compared for gender groups, after subjects have been matched based on their total test scores.

Item Level Psychometric Analyses

The unique contribution of this current study is the DIF analysis comparing gender groups at the test item level. Previous research on the RSES has involved only
analysis of psychometric properties of test scores at the scale level. None of the previous research studies that were identified in the review of the literature involved analyses of RSES scores at the item level. Zumbo (2003) noted that, while variation in test scores at the scale level may indicate no group differences, analysis at the item level may result in detection of group differences.

In psychometric analyses involving gender group differences of test scores, there is a need for DIF studies because this methodology enables researchers to rule out the possibility that group differences are an artifact of the measurement process. DIF analyses of RSES data are important because when making gender group comparisons you want to ensure that any statistically significant group differences that are reported in the literature reflect actual group differences.

Some studies on the RSES previously reported in the literature have reported statistically significant gender group differences based on total test score mean differences. Many of these studies have speculated reasons why males and females may have responded differently on the RSES test.

Results of the present study do not support a hypothesis that males and females respond differently on the RSES. Our study shows that males and females matched at the same level of RSES total test score, respond the same way. Our study provides evidence that males and females with the same level of self-esteem tend to give the same responses.

The current DIF analyses did not result in detection of significant DIF related to gender. Although three test items displayed statistically significant uniform DIF, due to
the very small effect sizes, the conclusion is that there is no meaningful gender DIF for this sample.

**Further RSES Psychometric Properties**

Zumbo (1999) pointed out that the literature has shown that sample sizes of at least 200 people per group are adequate for DIF analyses, and the more the better. The sample of participants involved in the present study (569 males and 874 females) was well beyond this minimum required number, contributing to the power of the findings.

In conducting DIF analyses, Zumbo (1999) cautioned, “...it will aid the interpretation of the results if the sample does not have missing data....one should conduct the analysis only on test takers from which you have complete data on the scale at hand (i.e., no missing data on the items comprising the scale and grouping variable for your analysis). (p.27-28). The data set used for the analysis in this study had no missing data.

**Areas for Further Research**

Although Rosenberg developed this scale for use with adolescents, very little previous research has compared psychometric properties of RSES scores for different age groups. Only one study was identified in the current literature review comparing RSES scores of adolescents to scores of adults (Whiteside-Mansell & Corwyn, 2003). In that study, no factorial variations were detected between the groups.

This present study, like so many other research studies on the RSES, has involved adult aged participants. The participants sampled for this study were a group of university students. Although their ages ranged widely from 17 to 65 years of age, over half of the sample was aged between 17-20 years old. The mean age for males was 21.94
years and for females was 22.59 years. This group is a fairly representative sample of college students. This study has not focused on the psychometric properties of the RSES scores in relation to different age groups, as there were no obvious age groupings in the data to make a meaningful comparison of groups.

Many studies reported in the literature have involved administration of the RSES to groups of adolescents. These studies have included a wide variety of different cultural groups, and interpretation of the RSES into many different languages. Studies involving adolescent gender group comparisons reported either no statistically significant group differences for RSES total test scores, or reported that males scored significantly higher than females. Given the RSES was originally developed for use with adolescents (Rosenberg, 1989), future psychometric analyses that include DIF analyses should be focused on adolescents.

Future research on the RSES may also involve qualitative investigations to explore whether any variables are interacting with the test. For example, because this is a self-administered test, test takers themselves may be able to provide valuable insights into the RSES.
References


depending on how an item is scored: An illustration with the Center for Epidemiological Studies Depression Scale, *Educational and Psychological Measurement*, 63, 65-74.


357-368.


self-esteem, body image, stressful life events, and pubertal status. *Journal of Emotional & Behavioral Disorders, 10, 29-42.*


Social Science & Medicine, 35, 1311-1315.


