Examining the Differential Item Functioning of the Rosenberg Self-Esteem Scale Across Eight Countries[1]

Lisa E. Baranik[2]
The University of Georgia

Adam W. Meade
North Carolina State University

Chad E. Lakey and
Charles E. Lance
The University of Georgia

Changya Hu
National Taiwan University of Science and Technology

Wei Hua
Singapore Management University

Alex Michalos
University of Northern British Columbia

We examined the differential item functioning (DIF) of Rosenberg's (1965) Self-Esteem Scale (RSES) and compared scores from U.S. participants with those from 7 other countries: Canada, Germany, New Zealand, Kenya, South Africa, Singapore, and Taiwan. Results indicate that DIF was present in all comparisons. Moreover, controlling for latent self-esteem, participants from individualistic countries had an easier time reporting high self-esteem on self-competence-related items, whereas participants from communal countries had an easier time reporting high self-esteem on self-liking items (Tafarodi & Milne, 2002). After adjusting for DIF, we found larger mean self-esteem differences between the countries than observed scores initially indicated. The suitability of the RSES, and the importance of examining DIF, for cross-cultural research are discussed.

Self-esteem refers to the extent to which one likes, values, accepts, and respects oneself at a general level (Brown, 1993; Rosenberg, 1979). Traditionally, researchers have viewed the possession of low versus high self-esteem as indicating the extent to which individuals take a negative or positive orientation toward themselves, and subsequently the extent to which individuals are (or are not) functioning well psychologically. For example, researchers have demonstrated that, when compared to those with low self-esteem, individuals with high self-esteem reveal greater happiness and fewer suicidal thoughts and depression (Harter, 1993); more clearly defined self-concepts (Campbell, 1990); greater optimism toward meeting goals (Scheier

[1]The authors thank Christina Collepardi for her helpful comments on the paper. Charles E. Lance was supported in part by National Institute on Drug Abuse Grant No. R01 DA019460-01A1; National Institute on Aging Grant No. AG15321; and National Cancer Institute Grant No. 5R03CA117470-02.

[2]Correspondence concerning this article should be addressed to Lisa Baranik, Department of Psychology, The University of Georgia, Athens, GA 30602-3013. E-mail: baranik@uga.edu

Journal of Applied Social Psychology, 2008, **38**, 7, pp. 1867–1904.

& Carver, 1985); and greater feelings concerning the capability to handle the ramifications of goals going unmet (Larrick, 1993). Individuals with high self-esteem experience less negative affect in response to failure than do those with low self-esteem (Kernis, Brockner, & Frankel, 1989) and, more specifically, less negative self-relevant emotions (e.g., humiliation; Brown & Dutton, 1995). Moreover, when compared to those with low self-esteem, individuals with high self-esteem evidence greater psychological well-being and other adaptive attitudes and behaviors that facilitate positive outcomes (Brown, Collins, & Schmidt, 1988; Rosenberg, Schoenbach, Schooler, & Rosenberg, 1995).

Perhaps the most often used measure of self-esteem is Rosenberg's (1965) Self-Esteem Scale (RSES). Generally, participants respond to 10 self-esteem-relevant items and are instructed to base their responses on how they typically, or generally, feel about themselves. One strength of this measure is its *face validity*; that is, the questions used to assess self-esteem all seem to tap into overall evaluation of self-liking. Another strength of this measure is the brevity of the assessment. Moreover, research supports that this scale is, at least in Western cultures, a reliable and valid measure of one's overall global feelings of self-worth (Blascovich & Tomaka, 1991).

However, the assessment of self-esteem—and the RSES (Rosenberg, 1965) as a measure of self-esteem—is not without criticism. For example, Tafarodi and colleagues (e.g., Tafarodi, Lang, & Smith, 1999; Tafarodi & Milne, 2002; Tafarodi & Swann, 1995, 1996; Tafarodi & Walters, 1999) have argued that the parsimony of a unidimensional self-esteem construct (as is commonly measured by the RSES) does not adequately reflect the extent to which feelings of self-worth can vary along the related, but possibly distinct, dimensions of self-competence and self-liking. That is, in their view, an individual's feelings of self-worth can manifest from feeling that one is an effective agent and is capable of exerting influence over environmental demands (i.e., self-competence) to the extent to which one recognizes and experiences inherent value in oneself (i.e., self-liking).

Moreover, although self-esteem has been studied extensively and linked to a wide range of outcome variables, much of this research has been conducted in North America and has been criticized for having low generalizability (Heine, Lehman, Markus, & Kitayama, 1999). Markus and Kitayama (1991) noted that conceptualizations of the self (e.g., self-esteem) are grounded in Western culture and reflect Western values (e.g., independence, individual uniqueness). Thus, the self-concept of individuals in Western cultures is rooted in traits that emphasize self–other distinctions and the self as an independent entity (Geertz, 1975). Conversely, the self-concept of those in Eastern or other cultures is rooted in interactions with others and is shaped inherently through social and cultural influences that emphasize relational or

interdependent processes. As such, in non-Western, collectivistic cultures, the self is more of a communal entity. In this vein, Taforodi and colleagues' (e.g., Tafarodi et al., 1999; Tafarodi & Swann, 1996) reasoning concerning distinctions of self-competence and self-liking has intuitive appeal for cross-cultural research.

Although there may be differences among individuals within certain cultures, the argument is that there are still distinct cultural uniformities that mark individuals between cultures (Markus, Mullally, & Kitayama, 1997; Tafarodi & Walters, 1999), such as the focus on uniqueness in countries like the United States, as compared to the focus on communality in non-Western countries (e.g., Japan). Heine et al. (1999) argued that "positive self-esteem, as it is currently conceptualized, operationalized, and measured, is not as prevalent, significant, sought after, discussed, functional, elaborated on, or desired in Japan as it is in North America" (p. 785). These considerations suggest that researchers in North America may view positive self-esteem as a desirable trait in part because it reflects their own values. In countries that do not share Western values, it may be erroneous to assume that positive self-esteem functions in an equivalent manner and, for that matter, is even viewed as a beneficial trait.

Other researchers disagree, and counter that esteeming oneself (i.e., pursuit of positive self-esteem) is a fundamental human drive. Classic theories of motivation, such as Maslow's (1943) hierarchy of needs, have depicted self-esteem in this manner, as do other, more contemporary theories (e.g., Brown, 1986; Crocker & Wolfe, 2001; Sedikides & Strube, 1997; Tesser, 1988). Some research offers support for these claims. For example, Sedikides, Gaertner, and Toguchi (2003) demonstrated that individuals from both individualistic and collectivistic cultures self-enhance on culturally ordained self-relevant dimensions in an effort to maintain or bolster their positive self-regard. In other research, Schmitt and Allik (2005) recently used the RSES (Rosenberg, 1965) to gather data in 53 countries and made a number of conclusions, including that the RSES had an invariant factor structure across all of the countries assessed, and that most respondents in countries do, in fact, report having positive mean self-esteem levels (i.e., above the scale midpoint). Thus, Schmitt and Allik asserted that their data provide empirical evidence to support that positive self-esteem is not solely a Western phenomenon.

However, the methods by which Schmitt and Allik (2005) came to their conclusions bear closer examination. Specifically, although Schmitt and Allik employed multiple statistical methods for investigating the psychometric properties of the RSES (Rosenberg, 1965) across countries, their criteria were less stringent than other recommended approaches (e.g., Vandenberg & Lance, 2000). Namely, before different countries can be compared using the

RSES, measurement invariance (MI) must be established. Verifying MI entails ensuring that (a) the same construct is being measured under different conditions (e.g., participants' nationality); and (b) the relationship between the items and the construct is identical across those conditions. Given that MI must be established before making meaningful comparisons between groups (Byrne & Campbell, 1999; Steenkamp & Baumgartner, 1998), Schmitt and Allik's findings may represent statistical artifacts, rather than substantive findings (van der Linden & Hambleton, 1997). In an effort to contribute to the cross-cultural self-esteem literature, we assessed the MI of the RSES across eight nations, using item response theory to evaluate the suitability of the scale for cross-cultural research.

Item Response Theory and the Rosenberg Self-Esteem Scale

The majority of studies examining the psychometric properties of the RSES (Rosenberg, 1965) have used exploratory or confirmatory factor analysis (CFA) to examine the factor structure of the scale (e.g., Carmines & Zeller, 1979; Farruggia, Chen, Greenberger, Dmitrieva, & Macek, 2004; Horan, DiStefano, & Motl, 2003; Marsh, 1996; Quilty, Oakman, & Risko, 2006; Schmitt & Allik, 2005; Tafarodi & Milne, 2002; Wang, Siegal, Falck, & Carlson, 2001). Although these studies certainly have helped to explain the overall factor-analytic structure of the RSES, they have not provided a detailed analysis of the psychometric properties of the RSES at the item level.

Item response theory (IRT) is a powerful psychometric method for examining the functioning of individual items in a scale. Specifically, IRT methods (Lord, 1980) link the probability of individuals' responses to their level of the latent construct being measured. Initially, researchers developed IRT to evaluate dichotomously scored ability tests. However, recent years have seen an increasing trend of using IRT for data with multiple response options (i.e., polytomous data). The marked rise in the number of researchers using IRT stems from the many advantages IRT affords for examining psychometric properties over traditional classical test theory (CTT) methods, such as exploratory and confirmatory factor analysis and internal consistency estimates. For instance, while CTT presumes an equal standard error of measurement for an entire test across all levels of the latent construct, IRT acknowledges differences in measurement precision for different scale items. Moreover, IRT allows for different standard errors for different levels of the latent trait. Other advantages of IRT measurement include item statistics that are not sample-dependent, and estimation of individuals' scores that are not dependent on the item properties (for a full discussion of the many advantages of IRT, see Embretson & Reise, 2000).

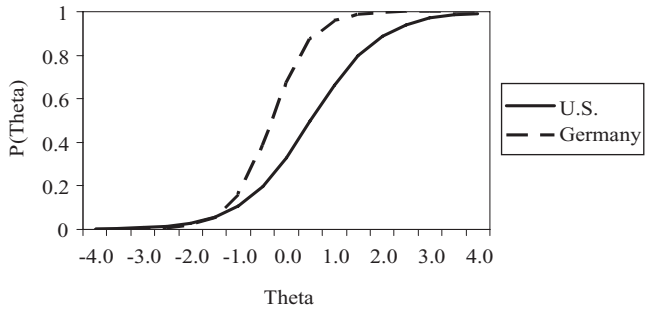

Figure 1. Item characteristic curves for Item 2: United States and Germany.

The merits of IRT can best be understood by examining the relationship between the probability of a specific response and the latent trait in the IRT model. For example, in the two-parameter logistic model, which models dichotomous data, the relationship between an individual's response and his or her level of the latent trait (e.g., self-esteem for the current paper), theta (θ) is represented by item characteristic curves (ICCs), which are nonlinear and S-shaped (see Figure 1 for two hypothetical ICCs). The slope of the ICC is referred to as the alpha (α) parameter, and it represents how well an item distinguishes between individuals at different levels of theta. High alpha values, which translate into steep ICC curves, indicate that the item is highly discriminating around theta levels where the slope is the steepest. High alpha parameters also signify that an individual's response to the item conveys a great deal of information about the individual's level of self-esteem when that person's theta level is similar to the difficulty level of the item. Items' alpha parameters are related mathematically to item-total correlations and are roughly analogous to factor loadings in CTT (McDonald, 1999). The placement of the ICC on the x-axis (i.e., level of the trait) is referred to as the beta (β) parameter and represents the amount of theta an individual needs to endorse an item with a .50 probability in the two-parameter IRT model.

High beta values, which translate into ICC curves located toward the right side of the x-axis, indicate that the item tends to be endorsed only by individuals who have high trait levels. Low beta values, on the other hand, indicate that the item will tend to be endorsed both by individuals who have low levels of the trait, as well as those who have high levels of the trait. In other words, it does not take much of the trait (e.g., self-esteem) to endorse items with low beta values. Thus, alpha and beta parameters provide valuable information for examining how well the item functions for individuals at all levels of the trait continuum.

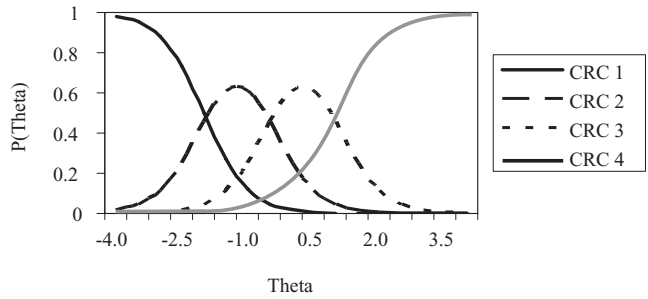

Figure 2. Category response curves for a hypothetical item with four response options.

An extension of the basic two-parameter logistic IRT model is Samejima's (1969) graded response model (GRM), which has been popular for use with personality assessments (e.g., Robie, Zickar, & Schmit, 2001; Zickar, 2002; Zickar & Robie, 1999) because it models ordinal, polytomous response scales (e.g., Likert scales). The GRM consists of a two-step process for estimating the probability that a participant responded to a specific response option. First, boundary response functions (BRFs), which are very similar to ICCs, are calculated. BRFs represent separations between response categories. Each item has one alpha (discrimination) parameter, but a number of beta parameters, which represent the amount of self-esteem needed to respond above their respective threshold with a .50 probability. For example, a 4-point Likert scale will have one alpha and three beta parameters. Second, once researchers estimate the alpha and beta values for the BRFs (note that *a* and *b* parameters are the sample estimated counterparts of the α and β parameters), category response curves (CRCs) can be computed in the second step. CRCs indicate the relationship between the respondent's level of self-esteem and the probability that the participant will respond to a particular response category (see Figure 2 for a sample CRC for an item with four response options). BRFs and CRCs relate such that high alpha values calculated in the first step lead to CRCs with more peaked curves, whereas beta values calculated in the first step determine where on the self-esteem continuum the boundaries between CRCs are located.

IRT item-level information, which shows that an instrument may provide uneven information across the trait continuum, is especially valuable when examining the RSES (Rosenberg, 1965). In light of evidence that most individuals score above the scale midpoint on the RSES (i.e., they have high self-esteem; Schmitt & Allik, 2005), it is imperative to establish that the RSES is accurately measuring self-esteem at high levels of the trait continuum. Recognizing this issue, Gray-Little, Williams, and Hancock (1997) employed

the GRM to examine the functioning of the RSES across different levels of self-esteem. The results show that Items 3, 5, and 6 (see Table 1 for the items) provided the most information about individuals across the self-esteem continuum. Items 4, 8, 9, and 10, on the other hand, distinguished among individuals with high self-esteem, yet they did not distinguish across lower levels of self-esteem. Finally, Items 1, 2, and 7 distinguished among individuals with low levels of self-esteem, but provided little information about individuals with high self-esteem.

Methods for Examining Measurement Invariance

In addition to examining the psychometric properties of tests with dichotomous and polytomous response options, one common use of IRT is to investigate measurement invariance (MI) across different groups of respondents. When a measure is invariant, it consistently measures a construct in the same way across conditions under which the construct is being observed (Horn & McArdle, 1992). Different conditions of measurement include factors such as gender, age, time, treatment conditions, and cultures. For example, Riordan and Vandenberg (1994) examined if the construct of organization-based self-esteem was measured consistently across cultures. The results indicate that citizens from the United States and citizens from Korea did not respond to items measuring organization-based self-esteem similarly. As a function of the failure to establish MI, meaningful comparisons could not be made between citizens from these two countries. As Vandenberg and Lance (2000) stated, "If one set of measures means one thing to one group and something different to another group, a group mean comparison may be tantamount to comparing apples and spark plugs" (p. 9). In sum, MI is not met when the scores of a test reflect both the trait that it was intended to measure and other unintended factors (e.g., cultural influences).

IRT researchers typically refer to a lack of MI as *differential item functioning* (DIF; Holland & Wainer, 1993). DIF is said to occur when an item functions differentially under different conditions, such as when individuals from two cultures with equal self-esteem levels respond differently to items that are nominally identical. When an item functions differentially, its ICCs (or BRFs) change under the different conditions as a result of different alpha parameters, beta parameters, or both. *Uniform DIF* refers to DIF that consistently "favors" one group over another; that is, it consistently takes less of the trait to have the same probability of endorsing an item for one group, as compared to another group. Typically, uniform DIF occurs where there is only DIF on the beta parameter. Conversely, *non-uniform DIF* refers to items that favor one group at some levels of the latent trait and the other group at

Table 1

Summary of DIF Analyses: Countries Having an Easier Time Indicating High Self-Esteem When Responding to RSES

	Country being compared to U.S.							
	Independent			Asian			African	
RSES item	Canada	Germany	New Zealand	Taiwan (E)	Taiwan (C)	Singapore	South Africa	Kenya
1. I feel that I'm a person of worth, at least on an equal basis with others.							U.S.	U.S.
2. I feel that I have a number of good qualities.	U.S.	U.S.		U.S.	U.S.		U.S.	U.S.
3. All in all, I am inclined to feel that I am a failure.[c]		Germany		U.S.		U.S.[a]		Kenya
4. I am able to do things as well as most people.	U.S.	U.S.	U.S.	Taiwan (E)	Taiwan (C)			
5. I feel I do not have much to be proud of.[c]		U.S.		U.S.	U.S.[a]	U.S.	U.S.	
6. I take a positive attitude toward myself.				Taiwan (E)	Taiwan (C)[a]		South Africa	Kenya
7. On the whole, I am satisfied with myself.		Germany		Taiwan (E)	Taiwan (C)	Singapore[b]	South Africa[a]	
8. I wish I could have more respect for myself.[c]								
9. I certainly feel useless at times.[c]				Taiwan (E)	Taiwan (C)[b]		South Africa	Kenya
10. At times I think I am no good at all.[c]	Canada	Germany		Taiwan (E)[b]				

Note. DIF = differential item functioning. RSES = Rosenberg (1965) Self-Esteem Scale. Item 8 was not included in DIF analyses. Taiwan (E) = English-speaking; Taiwan (C) = Chinese-speaking.
[a]Item discriminated better among U.S. participants. [b]Item discriminated better among participants from the country being compared to the U.S. [c]Item was reverse-scored.

other levels of the latent trait. Non-uniform DIF is associated typically with DIF on the alpha parameter. The implication of DIF is that individuals in different groups may have the same level of a trait, but because the item is functioning differentially, the two groups may have different observed scores on the item. For example, citizens in Canada and Kenya may both have equal levels of self-esteem, but if the measure functions differentially, observed mean scores may incorrectly show that one culture has higher self-esteem than the other.

Measurement Invariance and Cross-Cultural Research

MI is especially important in cross-cultural research. Cross-cultural research necessarily involves the comparison of multiple cultures on constructs of interest using measures of those constructs. In order to verify data patterns, such as the prevalence of self-esteem, the measure of that construct must function the same way in the cultures under investigation.

Researchers face a number of challenges when attempting to provide evidence for the generalizability of a theory to other cultures. In addition to logistic and cost impediments, researchers must also consider linguistic challenges. Drasgow and Probst (2004) outlined three specific linguistic challenges in cross-cultural research. First, if cultures use different languages, researchers must ensure that the translation from the original instrument into another language is accurate. Second, concepts described in the instrument must be understood in the same way by both cultures. Because the idea of self-esteem and the scales used to measure it (e.g., RSES; Rosenberg, 1965) have been developed in a manner congruent with Western ideals (i.e., valuing independence), it is possible that Asian and African countries, whose self-systems are rooted in more communal ideals, will interpret items on the instrument differently. For example, Item 4 ("I am able to do things as well as most people") might be easier to agree with for individuals who come from a more individualistic culture than for those who come from a more collectivistic culture. Participants from individualistic cultures may have more experience making other-referent comparisons than do participants from collectivistic cultures, so they may score higher on this item. This is not because they have higher self-esteem, but because of a systematic cultural difference that is related, for example, to how the item is read and understood. Finally, response scales must be utilized in the same way by both cultures. For example, even if two cultures can understand Item 4 equally well, a *strongly agree* response in one culture may not indicate the same level of self-esteem as a *strongly agree* response in another culture. In short, the measure used for cross-cultural comparisons must function equivalently across those cultures being examined.

Although Schmitt and Allik (2005) addressed a number of psychometric concerns in their study of self-esteem, they did not conduct a rigorous test of the invariance of the measurement properties of the RSES (Rosenberg, 1965) across the different countries. Rather, Schmitt and Allik followed van de Vijver and Leung's (1997) guidelines for examining structural equivalence, which is demonstrated when the internal relationships and the instrument's relationships with other theoretically related variables are consistent when administered across different countries (see van de Vijver & Leung, 2001). Importantly, the term *structural equivalence* is not to be confused with MI, as the latter term implies much more stringent standards with respect to the way items relate to constructs, as well as parametric tests used to evaluate the degree to which a measure functions equivalently across conditions (for a review, see Vandenberg & Lance, 2000). Using this framework, Schmitt and Allik examined similarities among the different countries' internal consistencies (alpha coefficients), and found a single dominant principal component in most countries and similar meta-traitedness indexes (an estimate of the consistency with which individuals responded to items on the same scale; Baumeister & Tice, 1988) across countries. Furthermore, the general directionality (i.e., positive or negative) of the correlations between the RSES and neuroticism and extraversion was largely the same across cultures.

Although Schmitt and Allik (2005) argued convincingly that a single factor underlies the RSES (Rosenberg, 1965) in nearly all cultures, they specifically noted that the presence of a single factor does not imply more stringent levels of invariance. Moreover, in van de Vijver and Leung's (2001) own words,

> It should be noted that, even if a personality inventory shows structural equivalence in a (cross-)cultural study, the scores based on this instrument may not be comparable across cultures. In more technical terms, structural equivalence does not yet imply measurement unit equivalence. (p. 1018)

Stated differently, it is essential to establish MI before making comparisons between different groups. The presence of a common factor and similar (e.g., positive) correlations among the RSES and personality dimensions across cultures hardly ensures that respondents are interpreting the scale in the same way or that the meaning of response options is identical across cultures. Although a single factor may underlie the data in two cultures, the amount of the latent trait needed to choose a response option of 4 (*agree*), for example, may differ greatly across cultures. Thus, while Schmitt and Allik provided a much needed preliminary investigation of the functioning of the RSES across cultural groups, further work is needed to establish the MI of the

RSES in order to determine the extent to which the RSES can be used to make accurate cross-cultural comparisons.

There are several potential approaches used to examine the equivalence of a measure across groups. However, the two most rigorous and predominant approaches utilize IRT or CFA (Raju, Laffitte, & Byrne, 2002; van de Vijver & Leung, 2001; Vandenberg & Lance, 2000). These two approaches, when applied to single constructs, are conceptually similar in that they both investigate the equivalence of the relationships among items and the latent trait across different conditions (e.g., cultural groups). While a thorough review of the differences between these two methods is beyond the scope of this paper (for an excellent review, see Raju et al., 2002), one notable difference is the use of a nonlinear model (IRT), as compared to a linear model (CFA).

More importantly for tests of invariance, however, IRT models for Likert-type data provide more information about how items relate to latent traits (via the estimation of more item parameters) than do CFA models (Raju et al., 2002). Specifically, CFA models estimate a single factor loading and item intercept for each item, whereas IRT models (e.g., graded response model; Samejima, 1969) commonly estimate as many parameters per item as there are response options. These additional parameters provide information related to the amount of latent trait that is necessary for a respondent to be likely to endorse one response option (e.g., 5) as compared to another (e.g., 4). As such, these additional item parameters then allow for more accurate detection of DIF (Maurer, Raju, & Collins, 1998; Meade & Lautenschlager, 2004; Raju et al., 2002). Thus, IRT provides the most stringent test available of the equivalence of the functioning of a measure across cultures.

The Present Study

In the current investigation, we seek to examine the MI of the RSES (Rosenberg, 1965) using IRT DIF analyses between the United States and seven other countries. Specifically, we measured the self-esteem of individuals from three countries with traditionally independent self-concepts similar to those in the U.S. (i.e., Canada, Germany, and New Zealand), and four countries with traditionally communal or interdependent self-concepts (i.e., Kenya, Singapore, South Africa, and Taiwan). Drasgow and Probst (2004) explained that the standard procedure for adapting an instrument for cross-cultural use begins with a survey developed in a source language or culture, which is then adapted into a target language or culture. Because the RSES was originally developed in English and researched primarily in the U.S., the U.S. sample serves as the source sample, and the other reported samples serve as the target samples.

Table 2

Demographic Characteristics of Countries

	N	*M* age	Female	Single	First-year
1. United States	496	22.8	65%	87%	69%
2. Canada	1573	21.8	62%	88%	45%
3. Germany	794	23.4	45%	88%	47%
4. Kenya	273	22.6	43%	98%	47%
5. New Zealand	322	21.5	64%	88%	39%
6. Taiwan, Chinese-speaking	334	20.3	58%	—	48%
7. Taiwan, English-speaking	255	21.1	65%	99%	4%
8. Singapore	173	21.2	53%	94%	29%
9. South Africa	293	23.1	56%	91%	41%

Note. First-year = students currently in their first year of college. The proportion of single students from the Taiwan, Chinese-speaking sample was not available. However, university statistics show that almost all of the students are single.

Method

Participants

College students in the U.S., Canada, Germany, New Zealand, South Africa, Kenya, Singapore, and Taiwan participated in the current study. Table 2 lists the demographic information of each of the samples, and Table 3 lists the universities and university characteristics from which each of the samples were gathered. Overall, participants across each of the countries were in their early 20s, and there was a representative mix of males and females. Most participants were not married, and were in their first year of college. The majority of universities from which data were gathered were public schools that were located in an urban area.

Procedure and Description of the RSES

Data for the current study were collected in two parts. Data from the United States, New Zealand, Germany, Kenya, and South Africa were collected during a larger investigation of student well-being, which is discussed in full detail in Michalos (1991). Briefly, Michalos solicited global

Table 3

Universities and Characteristics From Which Samples Were Drawn

Country	University	Public/ Private	Urban/ Rural
United States	Ohio State University, Newark	Public	Urban
	Edison Community College	Public	Urban
Canada	Dalhousie University	Private	Urban
	University of Guelph	Public	Urban
	Mount Saint Vincent	Public	Urban
	Saint Mary's University	Public	Urban
	Simon Fraser University	Public	Urban
Germany	Federal College of Public Administration	Public	Urban
	University of Frankfurt	Public	Urban
	University of Mannheim	Public	Urban
Kenya	University of Nairobi	Public	Urban
New Zealand	Massey University	Public	Urban
Taiwan, Chinese-speaking	National Chengchi University	Public	Urban
	National Taiwan University	Public	Urban
	National Taiwan University	Public	Urban
	Chaoyang University of Technology	Private	Urban
Taiwan, English-speaking	National Chengchi University	Public	Urban
	National Taiwan University	Public	Urban
	National Taiwan University of Science and Technology	Public	Urban
Singapore	Singapore Management University	Private	Urban

participation by contacting scholars in numerous countries around the world. Scholars were selected by convenience. Specifically, Michalos wrote to other scholars interested in well-being, who he knew through his position as editor and founder of the journal *Social Indicators Research* and through various organizations (e.g., United Nations Educational, Scientific and

Cultural Organization [UNESCO, Paris]; Organization for Economic Cooperation and Development [OECD, Paris]; International Sociological Association). Across 48 countries, 68 scholars volunteered to participate. Data were collected from Fall 1984 to Fall 1986 from 39 countries. A full listing of these countries, literacy rates, and per capita gross national products can be found in Michalos (1991). To summarize the samples, however, Michalos said that the sample was "biased toward relatively developed countries" (p. 69).

Data were sampled systematically in an effort to balance the different genders. Scholars gathered college student participants—oftentimes in large, introductory classrooms—to complete a survey containing demographic questions, a number of items concerning overall and life facet satisfaction, and the RSES (Rosenberg, 1965). Michalos (1991) noted that students' major course of study has almost no effect on well-being, meaning that the classes in which the data were collected should not have influenced the results.

Data from Taiwan and Singapore were collected by the fifth and sixth authors, respectively. There were two samples collected from Taiwan: one from English-speaking students, and one from Chinese-speaking students. Similar to the methods used by Michalos (1991), data were collected from large, introductory classrooms on college campuses. Details about the universities from which the data were collected are reported in Table 3. The surveys focused on demographic information and the RSES (Rosenberg, 1965).

For all countries, we used the full 10-item version of the RSES (Rosenberg, 1965) that is shown in Table 1. A 4-point Likert-type response scale was used, ranging from 1 (*strongly disagree*) to 4 (*strongly agree*). In addition, a fifth response option was provided indicating *no opinion*, which was located to the right of the other four options on the survey administered to participants. Alpha coefficients for each country were as follows: United States, $\alpha = .86$; Canada, $\alpha = .86$; Germany, $\alpha = .75$; New Zealand, $\alpha = .78$; South Africa, $\alpha = .68$; Kenya, $\alpha = .84$; Singapore, $\alpha = .83$; Taiwan, English-speaking, $\alpha = .67$; and Taiwan, Chinese-speaking, $\alpha = .78$.

Data Analysis

Data screening. Michalos' (1991) research focused on well-being, and not self-esteem. As a result, most countries participating in the larger project did not collect data on the RSES (Rosenberg, 1965). Out of the countries that did administer the RSES, we first examined data for unusual response patterns before conducting any analyses. Specifically, countries revealing an abnormal majority of responses coded as *no opinion* were excluded. Also, a number of data-collection sites used a 7-point scale, rather than the original 4-point scale plus the *no opinion* option that were used for the rest of the countries.

We only excluded countries from which these unusual response patterns appeared to be administration errors. For example, the Japan sample marked a range of 0.4%, 0.0%, and 1.6% of their responses as *no opinion* for Items 1, 2, and 3, respectively; whereas Items 4 through 10 had 30.8% to 37.5% percent of responses marked as *no opinion*. A possible explanation for this is that Items 1, 2, and 3 were administered to all participants; while Items 4 through 10 were administered only to some participants, with the missing data being coded as *no opinion*. Although a person's choice to indicate *no opinion* is interesting, because of the ambiguity of why this was listed as a response in many cases, we thought that it was inappropriate to include these countries. Therefore, we excluded data collected in Mexico, Bangladesh, Finland, Japan, Korea, and Sweden from analyses using this criterion. This left the eight countries that are included in the present study.

The surveys administered in Germany and in Taiwan among the Chinese-speaking students were written in German and Chinese, respectively. Surveys in the remaining countries were administered in English. Michalos (1991) provided evidence for a number of successful translations of the RSES (Rosenberg, 1965), including the German translation that is reported here.

In order for IRT to be used appropriately, a measure must be unidimensional. This is critical in the analysis of the RSES (Rosenberg, 1965) because of the complex history of the dimensionality of the scale, with some researchers arguing that the scale is unidimensional and others arguing that the scale measures two distinct constructs (e.g., Marsh, 1996; Tafarodi & Milne, 2002). Simulation studies (Drasgow & Parsons, 1983) have shown that data do not need to be strictly unidimensional in order to use IRT. For most practical applications, IRT can be conducted as long as the data have a single, dominant factor.

To investigate if the eight countries had a single, dominant factor, we conducted exploratory factor analysis (EFA) using maximum likelihood extraction on the RSES (Rosenberg, 1965) for each of the samples. Results from the EFAs were promising, as each country had one dominant factor emerge, similar to Schmitt and Allik's (2005) findings. However, we also noticed that among some countries, Item 8 had much lower factor loadings than did the other items. For example, Item 8 for the South African sample had a factor loading of .08, compared to the lowest factor loading among the other items, which was .35 for Item 7. Closer inspection of the data revealed that Item 8 exhibited low correlations with other items in the RSES as well. Schmitt and Allik likewise identified Item 8 as an especially problematic item on the RSES. Importantly, because IRT assumes that a dominant factor emerges from the data and is ideally unidimensional, Item 8 would cause violation of this assumption. Our findings, combined with those by Schmitt and Allik and by Farruggia et al. (2004), reveal that Item 8 is pervasively

Table 4

Results From Exploratory Factor Analysis of Rosenberg's Self-Esteem Scale for Each Country

	Factor 1		Factor 2	
	Eigenvalue	Variance explained	Eigenvalue	Variance explained
1. United States	4.05	45.0%	1.23	13.7%
2. Canada	4.05	44.9%	1.24	13.8%
3. Germany	3.86	42.9%	1.14	12.7%
4. Kenya	3.68	40.8%	1.40	15.5%
5. New Zealand	3.68	40.9%	1.24	13.8%
6. Taiwan, Chinese-speaking	3.45	38.3%	1.37	15.2%
7. Taiwan, English-speaking	3.12	34.7%	1.34	14.9%
8. Singapore	3.87	43.0%	1.51	16.7%
9. South Africa	2.83	31.4%	1.43	15.9%

problematic, and lead us to believe that this item is inappropriate for cross-cultural research. Therefore, we decided to exclude Item 8 from all subsequent analyses.

The results in Table 4 show that although there was variability in the size of the eigenvalues and the variance extracted with each eigenvalue across countries, all countries had one dominant factor emerge from the data, consistent with Rosenberg's (1965) theory concerning the self-esteem construct. Frequency distributions reveal that Items 1, 2, 3, 4, and 5 were negatively skewed, with the majority of responses falling in the *agree* and *strongly agree* categories. For negatively worded items, the majority of responses fell into the *disagree* and *strongly disagree* categories. Items 6, 7, and 10 were also negatively skewed, but respondents used three, rather than two, of the response options. Finally, Item 9 was approximately normally distributed, with respondents using all four of the response options. Across all of the items, very few participants selected *no opinion*.

Given that IRT requires a large number of responses in each response category being analyzed to provide an accurate estimate of the trait (Reise & Yu, 1990), we treated item responses of *no opinion* as missing data, and

collapsed responses to the 4-point scale according to the frequency distributions. These steps follow recommendations put forth by Gray-Little et al. (1997). For example, Item 1 was coded as dichotomous, with responses of 1, 2, and 3 coded as "0" and responses of 4 coded as "1." Fortunately, we found similar patterns of responses across all eight countries so that we were able to use a congruent coding scheme. We reverse-scored all negatively worded items before collapsing item responses.

IRT DIF analysis. We used the IRTLRDIF program (Thissen, 2001) to examine MI across the United States and the seven other countries using the GRM. The IRTLRDIF program uses the likelihood ratio test of nested models to examine if item parameters (*a* or *b*) differ across groups. In IRTLRDIF, a baseline model and a comparison model are compared to compute the likelihood ratio test of item parameters for a given item across two groups. The baseline model involves constraining the parameters of all scale items to be equal across groups. Subsequently, a comparison model is estimated for each item separately in which the parameters for only that item are freed to vary across groups. The difference in fit of the two models (G^2 value) is then compared using a chi-square table with the degrees of freedom equal to the number of parameters estimated for each item. If this likelihood ratio test is significant, the item is considered a DIF item. When IRTLRDIF identifies a DIF item, further analyses are conducted on the item to identify whether the source of the DIF is from the *a* parameter, the *b* parameter, or both, providing useful information about how the items are functioning when examining the two groups. The likelihood ratio tests currently can only be conducted in a pairwise fashion. Thus, we conducted tests of MI for the U.S. and for each of the target countries in a series of analyses.

Next, we computed latent trait scores that adjust for differences in item parameters across groups. Schmitt and Allik (2005) elaborated on the importance of comparing latent scores in cross-cultural studies because "comparing the raw scores of the RSES across cultures has somewhat limited value, unless the inherent bias related to the differential functioning of positive and negatively worded items has been taken into account" (p. 638). In order to accomplish this, we first estimated item parameters and latent trait scores separately in each group using MULTILOG (Thissen, 1991).

Next, we used a variant of Stocking and Lord's (1983) characteristic curve method to put the item parameters and latent trait scores on the same metric via Equate 2.1 (Baker, 1995). We used linking constants, which minimize the difference in expected test scores (given the item's parameters) generated from the two groups. Once item parameters are linked onto a common metric, theta scores for one group can be adjusted onto the metric of another group. We used only non-DIF items to link the item parameters to a common metric, and all groups were put onto the metric of the U.S. sample. In short,

we adjusted for problems with DIF, and thus made it possible to conduct more accurate comparisons of self-esteem between the eight countries.

Results

Results from the DIF analyses yield a multitude of parameter estimates for each country, when compared to the U.S. Rather than report the details of each analysis (which can be found in the Appendixes), we will report a detailed description of the DIF analysis of the U.S. versus Canada, first to interpret all of the analyses that we conducted. Then, we will provide a summary of results for the remaining countries.

United States Versus Canada

Results from the DIF analysis between the U.S. and Canada are presented in Table 5. Items 2, 4, and 10 were identified as functioning differentially, as shown by statistically significant G^2 values. These items display uniform DIF, meaning that there were differences in the *b* parameter, but not the *a* parameter. It was easier for U.S. participants to agree with Item 2 ("I feel that I have a number of good qualities"; $b_1 = -.14$) than it was for Canadian participants ($b_1 = .04$). Likewise, U.S. participants agreed more easily with Item 4 ("I am able to do things as well as most people"; $b_1 = .44$) than did Canadian participants ($b_1 = .66$).

However, for Item 10 ("At times, I think I am no good at all"), Canadian participants disagreed more easily ($b_1 = -.72$, $b_2 = .59$) than did U.S. participants ($b_1 = -.46$, $b_2 = .83$). These results indicate that for Items 2 and 4, U.S. participants are more likely to indicate that they have high self-esteem than are Canadian participants, even when individuals from both countries have the same level of self-esteem. Conversely, for Item 10, Canadian participants were more likely to reveal high self-esteem than were U.S. participants. Table 1 provides a summary of these results. In the second column, we summarize the aforementioned information by indicating whether the U.S. or Canadian participants had an easier time reporting high self-esteem on the item, even when self-esteem level across participants in the two countries was equivalent.

To examine these findings more closely, we next conducted a series of one-way ANOVAs on the original observed scores (before being collapsed for IRT analysis), the collapsed observed scale scores, and the DIF adjusted latent trait scores across the eight countries. Given the high power associated with our sample size, the ANOVAs were significant for original observed scores, $F(8,\ 4504) = 24.76$ $p < .001$; collapsed observed scores,

Table 5

Results From Likelihood Ratio Test Comparing United States to Canada

RSES item	Parameters tested	G^2	*df*	United States				Canada			
				a	b_1	b_2	b_3	*a*	b_1	b_2	b_3
1	All	5.1	2	1.86	–0.08			1.55	0.06		
2	All	7.4*	2	1.77	–0.14			1.43	0.04		
	a	2.5	1								
	b	4.9*	1								
3	All	1.2	2	2.57	0.10			2.26	0.08		
4	All	7.0*	2	1.45	0.44			1.56	0.66		
	a	0.0	1								
	b	6.9*	1								
5	All	1.3	2	2.36	0.29			2.40	0.21		
6	All	4.8	3	2.50	–1.05	0.73		2.94	–0.90	0.75	
7	All	4.0	3	2.15	–0.89	1.14		2.36	–0.85	0.96	
9	All	6.2	4	1.56	–1.82	0.09	1.61	1.33	–2.27	0.07	1.66
10	All	15.3*	3	1.53	–0.46	0.83		1.72	–0.72	0.59	
	a	1.3	1								
	b	14.1*	2								

Note. RSES = Rosenberg (1965) Self-Esteem Scale. *a* = alpha parameter; *b* = beta parameter. The beta parameters change from item to item as a function of collapsed response categories. Item 8 was removed from item response theory analyses.
*$p < .05$.

$F(8, 4504) = 12.80$, $p < .001$; as well as theta scores, $F(8, 4504) = 55.24$, $p < .001$. This shows that, regardless of using observed scores or DIF adjusted latent trait scores, there were statistically significant differences in self-esteem across the countries. However, effect sizes were larger for the DIF adjusted theta scores ($\eta_p^2 = .09$) than for original observed scores ($\eta_p^2 = .02$) or collapsed observed scores ($\eta_p^2 = .02$), as were mean differences (see Table 6). The results indicate that cultural differences were more readily apparent when examined by DIF adjusted self-esteem scores.

Using Tukey's pairwise comparisons, we found no statistically significant differences between the U.S. ($M = 3.13$, $SD = 0.50$) and Canada ($M = 3.14$, $SD = 0.47$). Despite the differentially functioning self-esteem items, in comparing Canada to the U.S., there were no statistically significant differences within observed or latent scores between these two countries. One reason the DIF did not manifest in mean differences may be because two of the items favored U.S. participants and one item favored Canadian participants, which

Table 6

Descriptive Statistics and Results from Tukey's HSD Comparisons by Country

		Original observed scores		Collapsed observed scores		Latent trait scores	
	n	*M*	*SD*	*M*	*SD*	*M*	*SD*
United States	496	3.13	0.50	1.75	0.41	–0.04	0.89
Canada	1573	3.14	0.47	1.75	0.40	–0.07	0.81
Germany	794	3.14	0.45	1.76	0.38	–0.04	0.68
Kenya	273	2.93*	0.47	1.61*	0.35	–0.58*	0.55
New Zealand	322	3.04	0.45	1.66*	0.37	–0.45*	0.66
Singapore	173	3.19	0.40	1.76	0.36	0.05	0.56
Taiwan (E)	255	2.88*	0.40	1.56*	0.29	–0.76*	0.37
Taiwan (C)	334	2.86*	0.57	1.67	0.38	–0.41*	0.66
South Africa	293	3.08	0.43	1.75	0.34	–0.06	0.43

Note. The midpoint for original observed scores was 2.50 and for collapsed observed scores it was 0.77. Because latent scores for each country are calculated relative to the other countries, a midpoint at the latent level cannot be calculated. HSD = honestly significant difference. Taiwan (E) = English-speaking; Taiwan (C) = Chinese-speaking.

*Country had a significantly different self-esteem score when compared to U.S. at $p < .05$.

effectively served to balance the overall mean score (Raju, van der Linden, & Fleer, 1995).

United States Versus Other Countries

Specific IRT parameter estimates for the other countries compared to the U.S. are reported in the Appendixes. These values can be interpreted in the same manner as the U.S. versus Canada estimates. Table 1 shows that, overall, there was less DIF when participants from other independent countries (e.g., Germany, New Zealand) were compared to those from the U.S. than when participants from more collectivistic countries, including both those from Asia and Africa, were compared to the U.S. Also, the pattern of DIF in Table 1 generally shows that, given the same level of self-esteem, U.S. participants had an easier time indicating that they had high self-esteem for RSES (Rosenberg, 1965) Items 1 through 5. However, participants from more collectivistic countries had an easier time indicating that they had high self-esteem for RSES Items 6 through 10.

Although the results presented in Table 1 clearly demonstrate a large occurrence of DIF when comparing the U.S. to other countries, it is interesting to note that after we adjusted the self-esteem scores for DIF, the ANOVA results (Table 6) did not change as much as would be expected with such a large number of differentially functioning items. Tukey's pairwise comparisons conducted on each of the countries compared to the U.S. reveal that, using the original observed scores, participants in Kenya ($M = 2.93$, $SD = 0.47$) and Taiwan (English-speaking, $M = 2.88$, $SD = 0.40$; Chinese-speaking, $M = 2.86$, $SD = 0.57$) had lower self-esteem scores than did participants from the U.S. ($M = 3.31$, $SD = 0.50$). Tukey's pairwise comparisons conducted on the DIF adjusted latent scores, replicated these findings; and also identified New Zealand participants ($M = –0.45$, $SD = 0.66$) as having lower self-esteem than U.S. participants ($M = –0.04$, $SD = 0.89$). Again, these results indicate that despite the large number of differentially functioning items on the RSES (Rosenberg, 1965), overall self-esteem scores obtained from the current sample were not biased to the extent that would be expected as a result of DIF balancing out across items (Raju et al., 1995).

Discussion

Self-esteem, or individuals' feelings concerning their own sense of self-worth, has long been studied in psychology under the belief that it is a much sought after and beneficial trait. However, most research examining the

construct of self-esteem has been conducted from a Western ideological standpoint, using Western participants. This fact has propelled researchers to investigate self-esteem cross-culturally in order to examine the possible universal nature of this construct. Although some researchers (e.g., Heine et al., 1999) believe that self-esteem is dependent on the North American culture in which it was conceptually derived, others (e.g., Brown, 1986; Crocker & Wolfe, 2001; Sedikides & Strube, 1997) believe that the drive for positive self-esteem is a universal human motive that transcends cultural boundaries. Before researchers can address the ubiquity of self-esteem, however, a necessary first step is to establish that the scale used to measure self-esteem functions adequately across cultures. In the current study, we analyzed the psychometric properties of the most widely used measure of self-esteem, Rosenberg's (1965) Self-Esteem Scale.

We found that the descriptive statistics and EFA results from the current study are similar to results found by Schmitt and Allik (2005), who also attempted to examine the cross-cultural validity of using the RSES (Rosenberg, 1965). Specifically, we confirmed some of Schmitt and Allik's findings in demonstrating that observed mean self-esteem scores across all countries were, in fact, above the midpoint of the response scale. As expected, more individualistic countries had higher mean level self-esteem scores, whereas more collectivistic countries had lower self-esteem scores (although Singapore participants, who are considered to be more collectivistic, had unexpectedly high self-esteem scores). Likewise, EFA yielded one predominant factor across all countries. It is important to note that we excluded Item 8 from all analyses because of low factor loadings between the item and the other RSES items on the scale. Other researchers (e.g., Farruggia et al., 2004; Schmitt & Allik, 2005) identified similar problems with Item 8. Based on past research and the findings from the current study, Item 8 appears to be pervasively problematic and inappropriate for use in cross-cultural comparisons.

However, unlike Schmitt and Allik (2005), we conducted a series of IRT analyses to examine the measurement invariance (MI) of the RSES (Rosenberg, 1965) between the U.S. and each of the other countries. Importantly, we found DIF in every analysis, including countries that are theoretically very similar (at least in terms of cultural makeup, and the extent to which they theoretically related in reference to Western ideals of esteeming the self), such as the U.S. and Canada. Moreover, we found that all of the examined RSES items functioned differentially for at least one of the IRT comparisons, although some items appear to be much more differentially functioning than others.

Overall, the results from the IRT analyses reveal a pattern whereby U.S. participants evidenced an easier time indicating that they have high self-esteem on the first five RSES (Rosenberg, 1965) items, whereas participants from

collectivistic countries had an easier time indicating that they have high self-esteem on the last five items. These findings underscore suggestions from proponents of distinct self-esteem dimension, such as Tafarodi and colleagues (Tafarodi et al., 1999; Tafarodi & Milne, 2002; Tafarodi & Swann, 1996), who proposed that the first five items on the RSES represent a self-assessment or self-competence dimension, whereas the last five items represent a self-acceptance or self-liking dimension. In their view, the self-competence component is similar to self-efficacy (cf. Bandura, 1989) and refers to feelings of self-worth that come from personal experiences concerning an individual's abilities, talents, and moments where one has (or has not) been able to act as an independent, causal agent. The instrumental value inherent to this dimension requires an external referent for comparison. That is, high self-esteem concerning self-competence stems from external appraisals concerning that one is doing better than someone else on some evaluative dimension.

Self-liking, on the other hand, refers to feelings of self-worth that emerge from an inherent recognition that one has value. As Tafarodi and Milne (2002) noted, this value stems, at least to some degree, from representation of the self as a social entity or stemming from one's own social value. Individualistic cultures, like the U.S., encourage autonomy, control, and pursuing one's own goals over the goals of the community. The current data suggest that the Westernized cultural system readily facilitates the promotion of self-esteem bred through self-competence. Collectivistic cultures, on the other hand, encourage deference and the pursuit of personal goals that are aligned with community goals. Our data suggest that this may serve to promote genuine self-liking for these individuals (Tafarodi et al., 1999). This is evidenced by the pattern of DIF found in the current study, revealing that U.S. participants had an easier time agreeing with self-competence-related items, but a harder time agreeing with items relating to self-liking, whereas we found the converse for individuals from collectivistic countries (Kenya, South Africa, Singapore, and Taiwan). That is, participants from collectivistic countries generally reported an easier time agreeing with self-liking items, but a harder time agreeing with self-competence items. Likewise, other researchers have provided empirical support for this cultural distinction (e.g., Schmitt & Allik, 2005; Tafarodi & Walters, 1999).

Despite the presence of DIF among comparisons between the U.S. and other independent countries (Canada, Germany, and New Zealand), the overall magnitude of DIF was minimal when comparing independent countries. The exception to this pattern was Germany, which had six differentially functioning items when compared to the U.S. One possible reason for this finding is that the RSES (Rosenberg, 1965) was administered in German, rather than English. In this case, the DIF could be a product of a poor translation, different use of the response scale, or differences in participants'

conceptualization of self-esteem (Drasgow & Probst, 2004). Future research should examine this possibility more closely.

Another interesting finding was the nature of the DIF when comparing the U.S. to New Zealand. Note that only one item functioned differentially, and observed mean self-esteem scores were not statistically different from one another. However, when the bias as a result of the DIF was taken into account, the mean self-esteem score for New Zealand was significantly lower than the U.S. We highlight this finding because it shows that even one differentially functioning item on a scale can distort mean level self-esteem results, which exemplifies the importance for cross-cultural researchers to test MI using IRT or CFA before making mean comparisons across groups (cf. van de Vijver & Leung, 2001).

DIF results concerning the two African countries (Kenya and South Africa) indicate that both countries were quite similar in their responses on the RSES (Rosenberg, 1965). That is, in a generally congruent fashion, self-liking items were easier for individuals from both African countries to indicate having high self-esteem, while self-competence items were easier for the U.S. participants to indicate having high self-esteem. Moreover, both the English and Chinese versions of the RSES had congruent parameter estimates when tested on the Taiwanese sample, providing evidence of a successful translation of the RSES into Chinese. Estimates from the African countries and Taiwan displayed the most DIF congruency and, although results from Singapore did align with those from the other communal countries, Singapore displayed the most dissimilar pattern of DIF among all of the communal countries. Considering that Singapore also had an unexpectedly high mean on the RSES, it is important for future research to investigate individuals' self-esteem in Singapore in more detail before strong conclusions are drawn. Singapore also had the lowest number of participants, and a larger sample would lend more confidence in these results.

Although the current study identified many differentially functioning items on the RSES (Rosenberg, 1965) when comparing the U.S. to other countries, there were only a few cases in which bias as a result of DIF caused latent mean self-esteem scores to be different from observed mean self-esteem scores. One explanation for this finding is that bias on the test balances out when the items are formed into a composite (see Raju et al., 1995). During every comparison (except the U.S. vs. New Zealand), some items were easier for U.S. participants to indicate that they had high self-esteem, whereas other items were easier for the other country's participants to indicate that they had high self-esteem. Although fewer differences between latent mean scores and observed mean scores were found than expected, the effect size increased when conducting an ANOVA on latent mean scores versus observed mean scores, indicating that there was more variability on self-esteem between

countries when the bias as a result of DIF was taken into account. Furthermore, there may be two distinct manners by which self-esteem is fostered and understood—namely, self-competence and self-liking—that are measured by the RSES. Summing all of the items on the RSES may make countries look more similar on mean level self-esteem scores than they actually are and may obscure meaningful differences.

When doing this type of cross-cultural research, it is imperative that researchers provide evidence that participants across the countries are responding to the instrument used to measure the construct in a congruent manner. In other words, researchers first must establish MI before making observed score comparisons. Without this evidence, there are inherent questions that arise concerning the validity of interpretations of mean score differences when comparing countries to one another (Horn & McArdle, 1992; Raju et al., 2002; Rensvold & Cheung, 1998; Steenkamp & Baumgartner, 1998; Vandenberg & Lance, 2000). It is important to note, however, that IRT analyses do not definitively answer why items function differentially. That is, although our findings align well with previous research and theory (e.g., Tafarodi & Walters, 1999), we cannot say definitively that the detected DIF was a result of different cultures defining self-esteem in different ways, or that it was a function of the RSES functioning differently across cultures.

Given that MI was generally not well established in our study, cross-cultural researchers have several methods by which they may try to proceed. First, we recommend that researchers begin cross-cultural research with an investigation of MI. If DIF is found, as it was in this study, researchers should consider the nature and severity of the DIF. Minor DIF, particularly as a result of *b* parameters, is less cause for concern than DIF as a result of *a* parameters. Like factor loadings, *a* parameters are indicative of the nature of the relationship between the item and the latent construct. Thus, large differences in *a* parameters across groups imply that the scale or item is functioning especially differently and, therefore, probably should not be used with its current wording. Minor differences in *a* parameters and differences in *b* parameters can be managed by estimating latent trait (i.e., θ) scores for each group and making comparisons on the basis of those scores, as we did in the current study. In the end, the researchers must use their best judgment as to the severity and nature of the DIF encountered.

It is important to note that a limitation to our study was the extent to which we collapsed categories as a result of low frequency of response. As with previous studies (e.g., Schmitt & Allik, 2005), we found that, in general, participants tended to agree with statements indicative of high self-esteem. Although it would have been preferable to have had some respondents in each response category, which would have allowed for more item parameters to be estimated, we were fortunate that the general frequency of response

patterns was highly similar across countries. Collapsing categories to the extent that we did was necessary for adequate estimation of the item parameters examined in the study.

Another consideration concerning interpretation of the present findings is the demographic variability among participants. When conducting MI studies, it is preferable to have participants be as homogeneous as possible, except for the condition that is being examined as a cause of MI (Vandenberg & Lance, 2000). Although all participants were college students, the samples from Taiwan and Singapore had lower proportions of first-year students than did the other countries. Additionally, samples from the U.S., Canada, Germany, New Zealand, Kenya, and South Africa were collected earlier than were samples from Taiwan and Singapore. Age and gender proportions were fairly constant; however, it is possible that, for example, the lower proportion of females in the Kenyan sample and higher mean age in the German sample or translation issues may have confounded the results.

Finally, it is interesting that we did not find a meaningful pattern of DIF for the positively and negatively worded items on the RSES (Rosenberg, 1965). Rather, the current research shows that the RSES may have a self-competence and self-liking component, suggesting that a cross-cultural confirmatory factor analysis examining a two-dimensional factor structure focusing on the self-competence and self-liking dimension may be a particularly useful direction for future research.

In summary, every item on the RSES (Rosenberg, 1965) exhibited DIF at some point, and every country likewise had at least one differentially functioning item, with some countries exhibiting up to eight differentially functioning items. The pattern of DIF demonstrates that those from individualistic cultures (e.g., U.S.) may have an easier time indicating that they have high self-esteem when the item focuses on assessment of objective qualities relating to self-competence, whereas individuals from collectivistic countries evidence higher self-esteem when the item focuses on more subjective qualities relating to self-acceptance. Although more research is needed, the results from our analyses nonetheless suggest that self-esteem, as measured by the RSES, is not necessarily conceptualized in congruent ways across countries and that this difference may be especially pronounced when comparing those from countries with individualistic values to people from countries with more communal values.

References

Baker, F. B. (1995). *Equate 2.1: Computer program for equating two metrics in item response theory*. Madison, WI: University of Wisconsin, Laboratory of Experimental Design.

Bandura, A. (1989). Human agency in social cognitive theory. *American Psychologist*, *44*, 1175–1184.

Baumeister, R. F., & Tice, D. M. (1988). Metatraits. *Journal of Personality*, *56*, 571–598.

Blascovich, J., & Tomaka, J. (1991). Measures of self-esteem. In J. P. Robinson, P. R. Shaver, & L. S. Wrightsman (Eds.), *Measures of personality and social psychological attitudes* (pp. 115–160). San Diego, CA: Academic Press.

Brown, J. D. (1986). *Social psychology* (2nd ed.). New York: Free Press.

Brown, J. D. (1993). Self-esteem and self-evaluation: Feeling is believing. In J. Suls (Ed.), *Psychological perspectives on the self* (Vol. *4*, pp. 27–58). Hillsdale, NJ. Lawrence Erlbaum.

Brown, J. D., Collins, R. L., & Schmidt, G. W. (1988). Self-esteem and direct versus indirect forms of self-enhancement. *Journal of Personality and Social Psychology*, *59*, 538–549.

Brown, J. D., & Dutton, K. A. (1995). The thrill of victory, the complexity of defeat: Self-esteem and people's emotional reactions to success and failure. *Journal of Personality and Social Psychology*, *68*, 712–722.

Byrne, B. M., & Campbell, T. L. (1999). Cross-cultural comparisons and the presumption of equivalent measurement and theoretical structure: A look beneath the surface. *Journal of Cross-Cultural Psychology*, *30*, 555–574.

Campbell, J. D. (1990). Self-esteem and clarity of the self-concept. *Journal of Personality and Social Psychology*, *59*, 538–549.

Carmines, E. G., & Zeller, R. A. (1979). *Reliability and validity assessment*. Beverly Hills, CA: Sage.

Crocker, J., & Wolfe, C. T. (2001). Contingencies of self-worth. *Psychological Review*, *108*, 593–623.

Drasgow, F., & Parsons, C. K. (1983). Application of unidimensional item response theory models to multidimensional data. *Applied Psychological Measurement*, *7*, 189–199.

Drasgow, F., & Probst, T. M. (2004). The psychometrics of adaptation: Evaluating measurement equivalence across languages and cultures. In R. K. Hambleton, P. F. Merenda, & C. D. Spielberger (Eds.), *Adapting educational and psychological tests for cross-cultural assessment* (pp. 265–298). Mahwah, NJ: Lawrence Erlbaum.

Embretson, S. E., & Reise, S. P. (Eds.). (2000). *Item response theory for psychologists*. Mahwah, NJ: Lawrence Erlbaum.

Farruggia, S. B., Chen, C., Greenberger, E., Dmitrieva, J., & Macek, P. (2004). Adolescent self-esteem in cross-cultural perspective: Testing measurement equivalence and a mediating model. *Journal of Cross-Cultural Psychology*, *35*, 719–733.

Geertz, C. (1975). On the nature of anthropological understanding. *American Scientist*, *63*, 47–53.

Gray-Little, B., Williams, V. S. L., & Hancock, T. D. (1997). An item response theory analysis of the Rosenberg Self-Esteem Scale. *Personality and Social Psychology Bulletin*, *23*, 443–451.

Harter, S. (1993). Causes and consequences of low self-esteem in children and adolescents. In R. R. Baumeister (Ed.), *Self-esteem: The puzzle of low self-regard* (pp. 87–116). New York: Plenum.

Heine, S. J., Lehman, D. R., Markus, H. R., & Kitayama, S. (1999). Is there a universal need for positive self-regard? *Psychological Review*, *106*, 766–794.

Holland, P. W., & Wainer, H. (Eds.). (1993). *Differential item functioning*. Hillsdale, NJ: Lawrence Erlbaum.

Horan, P. M., DiStefano, C., & Motl, R. W. (2003). Wording effects in self-esteem scales: Methodological artifact or response style? *Structural Equation Modeling*, *10*, 435–455.

Horn, J. L., & McArdle, J. J. (1992). A practical and theoretical guide to measurement invariance in aging research. *Experimental Aging Research*, *18*, 117–144.

Kernis, M. H., Brockner, J., & Frankel, B. S. (1989). Self-esteem and reactions to failure: The mediating role of overgeneralization. *Journal of Personality and Social Psychology*, *57*, 707–714.

Larrick, R. P. (1993). Motivational factors in decision theories: The role of self-protection. *Psychological Bulletin*, *113*, 440–450.

Lord, F. (1980). *Applications of item response theory to practical testing problems*. Hillsdale, NJ: Lawrence Erlbaum.

Markus, H. R., & Kitayama, S. (1991). Culture and the self: Implications for cognition, emotion, and motivation. *Psychological Review*, *98*, 224–253.

Markus, H. R., Mullally, P. R., & Kitayama, S. (1997). Selfways: Diversity in modes of cultural participation. In U. Neisser & D. A. Jopling (Eds.), *The conceptual self in context: Culture, experience, self-understanding* (pp. 13–61). New York: Cambridge University Press.

Marsh, H.W. (1996). Positive and negative global self-esteem: A substantively meaningful distinction or artifacts? *Journal of Personality and Social Psychology*, *70*, 810–819.

Maslow, A. H. (1943). A theory of human motivation. *Psychological Review*, *50*, 370–396.

Maurer, T. J., Raju, N. S., & Collins, W. C. (1998). Peer and subordinate performance appraisal measurement equivalence. *Journal of Applied Psychology*, *83*, 693–702.

McDonald, R. P. (1999). *Test theory: A unified treatment*. Mahwah, NJ: Lawrence Erlbaum.

Meade, A. W., & Lautenschlager, G. J. (2004). A comparison of item response theory and confirmatory factor analytic methodologies for establishing measurement equivalence/invariance. *Organizational Research Methods*, *7*, 361–368.

Michalos, A. C. (1991). *Global report on student well-being*. New York: Springer-Verlag.

Quilty, L. C., Oakman, J. M., & Risko, E. (2006). Correlates of the Rosenberg Self-Esteem Scale method effects. *Structural Equation Modeling*, *13*, 99–117.

Raju, N. S., Laffitte, L. J., & Byrne, B. M. (2002). Measurement equivalence: A comparison of methods based on confirmatory factor analysis and item response theory. *Journal of Applied Psychology*, *87*, 517–529.

Raju, N. S., van der Linden, W. J., & Fleer, P. F. (1995). IRT-based internal measures of differential item functioning of tests and items. *Applied Psychological Measurement*, *19*, 353–368.

Reise, S. P., & Yu, J. (1990). Parameter recovery in the graded response model using MULTILOG. *Journal of Educational Measurement*, *27*, 133–144.

Rensvold, R. B., & Cheung, G. W. (1998). Testing measurement models for factorial invariance: A systematic approach. *Educational and Psychological Measurement*, *58*, 1017–1034.

Riordan, C. M., & Vandenberg, R. J. (1994). A central question in cross-cultural research: Do employees of different cultures interpret work-related measures in an equivalent manner? *Journal of Management*, *20*, 643–671.

Robie, C., Zickar, M. J., & Schmit, M. J. (2001). Measurement equivalence between applicant and incumbent groups: An IRT analysis of personality scales. *Human Performance*, *14*, 187–207.

Rosenberg, M. (1965). *Society and the adolescent child*. Princeton, NJ: Princeton University Press.

Rosenberg, M. (1979). *Conceiving the self*. New York: Basic Books.

Rosenberg, M., Schoenbach, C., Schooler, C. & Rosenberg, F. (1995). Global self-esteem and specific self-esteem: Different concepts, different outcomes. *American Sociological Review*, *60*, 141–156.

Samejima, F. (1969). Estimation of latent ability using a response pattern of graded scores. *Psychometrika Monograph*, 17.

Scheier, M. F., & Carver, C. S. (1985). Optimism, coping, and health: Assessment and implications of generalized outcome expectancies. *Health Psychology*, *4*, 219–247.

Schmitt, D. P., & Allik, J. (2005). Simultaneous administration of the Rosenberg Self-Esteem Scale in 53 nations: Exploring the universal and culture-specific features of global self-esteem. *Journal of Personality and Social Psychology*, *89*, 623–642.

Sedikides, C., Gaertner, L., & Toguchi, Y. (2003). Pancultural self-enhancement. *Journal of Personality and Social Psychology*, *84*, 60–79.

Sedikides, C., & Strube, M. J. (1997). Self-evaluation: To thine own self be good, to thine own self be sure, to thine own self be true, and to thine own self be better. In M. P. Zanna (Ed.), *Advances in experimental social psychology* (Vol. *29*, pp. 209–269). San Diego, CA: Academic Press.

Steenkamp, J. E. M., & Baumgartner, H. (1998). Assessing measurement invariance in cross-national consumer research. *Journal of Consumer Research*, *25*, 78–90.

Stocking, M. L., & Lord, F. M. (1983). Developing a common metric in item response theory. *Applied Psychological Measurement*, *7*, 201–210.

Tafarodi, R. W., Lang, J. M., & Smith, A. J. (1999). Self-esteem and the cultural trade-off: Evidence for the role of individualism–collectivism. *Journal of Cross-Cultural Psychology*, *30*, 620–640.

Tafarodi, R. W., & Milne A. B. (2002). Decomposing global self-esteem. *Journal of Personality*, *70*, 443–483.

Tafarodi, R. W., & Swann, W. B., Jr. (1995). Self-liking and self-competence as dimensions of global self-esteem: Initial validation of a measure. *Journal of Personality Assessment*, *65*, 322–342.

Tafarodi, R. W., & Swann, W. B., Jr. (1996). Individualism–collectivism and global self-esteem: Evidence for a cultural trade-off. *Journal of Cross-Cultural Psychology*, *27*, 651–672.

Tafarodi, R. W., & Walters, P. (1999). Individualism–collectivism, life events, and self-esteem: A test of two tradeoffs. *European Journal of Social Psychology*, *29*, 797–814.

Tesser, A. (1988). Toward a self-evaluation maintenance model of social behavior. In L. Berkowitz (Ed.), *Advances in experimental social psychology* (Vol. *21*, pp. 181–227). San Diego, CA: Academic Press.

Thissen, D. (1991). *MULTILOG users guide: Multiple categorical item analysis and test scoring using item response theory*. Chicago: Scientific Software International.

Thissen, D. (2001). *IRTLRDIF. Version 2.0b: Software for the computation of statistics involved in item response theory likelihood–ratio tests for differential item functioning*. Chapel Hill, NC: University of North Carolina at Chapel Hill.

Vandenberg, R. J., & Lance, C. E. (2000). A review and synthesis of the measurement invariance literature: Suggestions, practices, and recommendations for organizational research. *Organizational Research Methods*, *3*, 4–69.

van der Linden, W. J., & Hambleton, R. K. (Eds.). (1997). *Handbook of modern item response theory*. New York: Springer.

van de Vijver, F. J. R., & Leung, K. (1997). Methods and data analysis of comparative research. In J. W. Berry, Y. H. Poortinga, & J. Pandey (Eds.), *Handbook of cross-cultural psychology* (2nd ed., pp. 257–300). Needham Heights, MA: Allyn & Bacon.

van de Vijver, F. J. R., & Leung, K. (2001). Personality in cultural context: Methodological issues. *Journal of Personality*, *69*, 1007–1031.

Wang, J., Siegal, H. A., Falck, R. S., & Carlson, R. G. (2001). Factorial structure of Rosenberg's Self-Esteem Scale among crack-cocaine drug users. *Structural Equation Modeling*, *8*, 275–286.

Zickar, M. J. (2002). Modeling data with polytomous item response theory. In F. Drasgow & N. Schmitt (Eds.), *Measuring and analyzing behavior in organizations: Advances in measurement and data analysis* (pp. 123–155). San Francisco: Jossey-Bass/Pfeiffer.

Zickar, M. J., & Robie, C. (1999). Modeling faking good on personality items: An item-level analysis. *Journal of Applied Psychology*, *84*, 551–563.

Appendix A

Results From Likelihood Ratio Test Comparing United States With Germany

RSES item	Parameters tested	G^2	*df*	U.S.				Germany			
				a	b_1	b_2	b_3	*a*	b_1	b_2	b_3
1	All	3.3	2	1.80	−0.07			1.39	−0.02		
2	All	53.4*	2	1.72	−0.15			1.21	0.57		
	a	5.2	1								
	b	48.2*	1								
3	All	16.3*	2	2.56	0.12			2.90	−0.13		
	a	0.7	1								
	b	15.6*	1								
4	All	67.6*	2	1.51	0.44			1.63	1.21		
	a	0.2	1								
	b	67.4*	1								
5	All	14.3*	2	2.42	0.26			2.13	0.55		
	a	0.8	1								
	b	13.5*	1								
6	All	7.1	3	2.46	−1.05	0.74		3.22	−0.83	0.69	
7	All	12.9*	3	2.14	−0.88	1.16		2.49	−0.93	0.82	
	a	1.6	1								
	b	11.3*	2								
9	All	4.9	4	1.56	−1.83	0.09	1.61	1.62	−1.88	0.02	1.34
10	All	55.8*	3	1.56	−0.45	0.83		2.11	−0.90	0.35	
	a	6.2	1								
	b	49.6*	2								

Note. RSES = Rosenberg (1965) Self-Esteem Scale. *a* = alpha parameter; *b* = beta parameter. The beta parameters change from item to item as a function of collapsed response categories. Item 8 was removed from item response theory analyses.
*$p < .05$.

Appendix B

Results From Likelihood Ratio Test Comparing United States With New Zealand

RSES item	Parameters tested	G^2	*df*	U.S. *a*	U.S. b_1	U.S. b_2	U.S. b_3	New Zealand	New Zealand b_1	New Zealand b_2	New Zealand b_3
1	All	0.7	2	1.96	–0.07			1.69	–0.07		
2	All	2.9	2	1.83	–0.13			1.65	0.04		
3	All	0.2	2	2.68	0.10			2.84	0.07		
4	All	14.0*	2	1.59	0.44			1.04	1.15		
	a	3.8	1								
	b	10.2*	1								
5	All	0.1	2	2.47	0.28			2.35	0.28		
6	All	5.3	3	2.36	–1.05	0.76		2.74	–1.01	0.51	
7	All	8.3	3	2.03	–0.91	1.16		2.07	–0.86	0.83	
9	All	7.6	4	1.57	–1.82	0.09	1.60	1.49	–2.13	0.33	1.87
10	All	3.3	3	1.58	–0.47	0.80		1.96	–0.54	0.66	

Note. RSES = Rosenberg (1965) Self-Esteem Scale. *a* = alpha parameter; *b* = beta parameter. The beta parameters change from item to item as a function of collapsed response categories. Item 8 was removed from item response theory analyses.
*$p < .05$.

Appendix C

Results From Likelihood Ratio Test Comparing United States With Kenya

RSES item	Parameters tested	G^2	*df*	U.S.				Kenya			
				a	b_1	b_2	b_3	*a*	b_1	b_2	b_3
1	All	26.7*	2	2.07	−0.08			1.43	0.59		
	a	2.9	1								
	b	23.8*	1								
2	All	10.5*	2	2.00	−0.13			1.38	0.25		
	a	3.0	1								
	b	7.6*	1								
3	All	19.2*	2	2.66	0.12			1.96	−0.24		
	a	2.1	1								
	b	17.1*	1								
4	All	4.3	2	1.65	0.44			2.45	0.21		
5	All	6.0*	2	2.41	0.27			1.73	0.60		
	a	2.4	1								
	b	3.6	1								
6	All	37.3*	3	2.44	−1.02	0.76		2.84	−1.29	0.17	
	a	0.6	1								
	b	36.7*	2								
7	All	10.3	3	2.01	−0.92	1.16		2.14	−0.66	0.88	
9	All	16.7*	4	1.53	−1.85	0.09	1.63	1.22	−1.94	0.23	1.34
	a	1.5	1								
	b	15.2*	3								
10	All	5.8	3	1.53	−0.48	0.81		1.52	−0.20	0.89	

Note. RSES = Rosenberg (1965) Self-Esteem Scale. *a* = alpha parameter; *b* = beta parameter. The beta parameters change from item to item as a function of collapsed response categories. Item 8 was removed from item response theory analyses.
*$p < .05$.

Appendix D

Results From Likelihood Ratio Test Comparing United States With South Africa

RSES item	Parameters tested	G^2	*df*	U.S.				South Africa			
				a	b_1	b_2	b_3	*a*	b_1	b_2	b_3
1	All	75.5*	2	2.14	–0.08			1.73	0.86		
	a	0.9	1								
	b	74.6*	1								
2	All	20.3*	2	1.98	–0.13			2.22	0.27		
	a	0.3	1								
	b	20.0*	1								
3	All	0.0	2	2.74	0.11			2.76	0.12		
4	All	1.3	2	1.67	0.44			2.14	0.36		
5	All	16.8*	2	2.58	0.27			2.27	0.67		
	a	0.4	1								
	b	16.5*	1								
6	All	53.5*	3	2.19	–1.06	0.79		2.42	–1.33	0.10	
	a	0.4	1								
	b	53.1*	2								
7	All	40.4*	3	1.94	–0.92	1.18		0.99	–1.22	1.05	
	a	10.3*	1								
	b	30.1*	3								
9	All	57.5*	4	1.55	–1.82	0.11	1.63	1.98	–1.67	–0.51	0.65
	a	2.0	1								
	b	55.5*	3								
10	All	5.0	3	1.57	–0.46	0.80		1.98	–0.58	0.57	

Note. RSES = Rosenberg (1965) Self-Esteem Scale. *a* = alpha parameter; *b* = beta parameter. The beta parameters change from item to item as a function of collapsed response categories. Item 8 was removed from item response theory analyses.
*$p < .05$.

Appendix E

Results From Likelihood Ratio Test Comparing United States With Singapore

RSES item	Parameters tested	G^2	*df*	U.S.				Singapore			
				a	b_1	b_2	b_3	*a*	b_1	b_2	b_3
1	All	0.8	2	2.06	–0.06			1.94	–0.16		
2	All	4.4	2	1.95	–0.13			2.27	0.09		
3	All	11.9*	2	2.58	0.10			1.60	0.51		
	a	3.9*	1								
	b	8.0*	1								
4	All	1.3	2	1.65	0.44			2.12	0.43		
5	All	26.1*	2	2.38	0.27			2.67	0.80		
	a	0.2	1								
	b	25.9*	1								
6	All	11.5	3	2.38	–1.03	0.76		2.61	–1.27	0.42	
7	All	44.0*	3	2.02	–0.90	1.17		3.69	–0.96	0.31	
	a	7.9*	1								
	b	36.1*	3								
9	All	6.1	4	1.52	–1.86	0.09	1.63	1.01	–2.23	0.02	2.08
10	All	1.4	3	1.52	–0.48	0.81		1.70	–0.38	0.90	

Note. RSES = Rosenberg (1965) Self-Esteem Scale. *a* = alpha parameter; *b* = beta parameter. The beta parameters change from item to item as a function of collapsed response categories. Item 8 was removed from item response theory analyses.
*$p < .05$.

Appendix F

Results From Likelihood Ratio Test Comparing United States With Taiwan (English-Speaking)

RSES item	Parameters tested	G^2	*df*	U.S. *a*	U.S. b_1	U.S. b_2	U.S. b_3	Taiwan (English) *a*	Taiwan (English) b_1	Taiwan (English) b_2	Taiwan (English) b_3
1	All	0.8	2	2.15	−0.05			2.02	0.04		
2	All	30.6*	2	1.92	−0.13			2.33	0.38		
	a	0.5	1								
	b	30.1*	1								
3	All	12.1*	2	2.74	0.10			2.55	0.47		
	a	0.1	1								
	b	12.0*	1								
4	All	7.8*	2	1.66	0.45			2.46	0.03		
	a	2.3	1								
	b	5.5*	1								
5	All	25.0*	2	2.53	0.27			2.29	0.92		
	a	0.1	1								
	b	24.9*	1								
6	All	48.4*	3	2.23	−1.06	0.79		3.19	−1.18	−0.04	
	a	2.8	1								
	b	45.6*	2								
7	All	15.0*	3	1.84	−0.94	1.21		2.70	−0.90	0.45	
	a	2.9	1								
	b	12.1*	2								
9	All	41.5*	4	1.57	−1.81	0.11	1.62	2.05	−1.89	−0.59	1.08
	a	1.4	1								
	b	40.1*	3								
10	All	20.8*	3	1.56	−0.46	0.80		3.16	−0.57	0.54	
	a	10.1*	1								
	b	10.8*	2								

Note. RSES = Rosenberg (1965) Self-Esteem Scale. *a* = alpha parameter; *b* = beta parameter. The beta parameters change from item to item as a function of collapsed response categories. Item 8 was removed from item response theory analyses.
*$p < .05$.

Appendix G

Results From Likelihood Ratio Test Comparing United States With Taiwan (Chinese-Speaking)

RSES item	Parameters tested	G^2	*df*	U.S.				Taiwan (Chinese)			
				a	b_1	b_2	b_3	*a*	b_1	b_2	b_3
1	All	1.6	2	2.20	–0.05			1.73	–0.05		
2	All	90.4*	2	2.11	–0.13			1.79	0.88		
	a	0.7	1								
	b	89.7*	1								
3	All	0.3	2	2.74	0.11			2.50	0.13		
4	All	10.5*	2	1.68	0.45			2.04	0.09		
	a	1.0	1								
	b	9.6*	1								
5	All	54.5*	2	2.64	0.27			1.04	1.51		
	a	18.8*	1								
	b	35.6*	1								
6	All	113.6*	3	2.37	–1.04	0.75		0.06	–1.63	0.80	
	a	48.5*	1								
	b	65.1*	2								
7	All	65.6*	3	1.81	–0.95	1.23		2.32	–0.84	0.26	
	a	2.3	1								
	b	63.4*	2								
9	All	84.5*	4	1.58	–1.81	0.11	1.62	2.37	–1.67	–0.45	0.34
	a	6.6*	1								
	b	77.9*	3								
10	All	6.2	3	1.59	–0.47	0.79		1.97	–0.22	0.72	

Note. RSES = Rosenberg (1965) Self-Esteem Scale. *a* = alpha parameter; *b* = beta parameter. The beta parameters change from item to item as a function of collapsed response categories. Item 8 was removed from item response theory analyses.
*$p < .05$.