Measurement invariance of Rosenberg Self-Esteem Scale between British and Chinese college students

Exam no. 2997847

Project supervisor: Dr Wendy Johnson

Word counts:

Exam No. 2997847

CONTENTS

Exam No. 2997847

ABSTRACT

The present study examined the factor structure and measurement invariance of the Rosenberg Self-Esteem Scale in college students from Britain (N=150) and China (N=205). Confirmatory factor analyses suggested that the two-factor model, which consisted of a positive self-image factor and a negative self-image factor, could fit the data better than single factor structure especially after deleting the eighth item. Furthermore, factorial structure was invariant across groups in configural level and weak factorial level. After releasing several intercepts, partial strong and strict factorial invariance were certified. Subsequently, DIF (differential item functioning) and response patterns were analyzed, evidence indicates that approximately half items are non-invariant on intercept level. However, literature provides little guidance about the implications for the use of the partially invariant scale. In the end, psychological properties of the nine-item RSES was measured, Cronbach's α were satisfactory and item-total correlations were good for Chinese data, and acceptable for British data. Finally, limitations and implications were discussed.

Measurement invariance of Rosenberg Self-Esteem Scale between British and Chinese college students

1. Introduction

1.1 What is self-esteem?

Self-esteem is a term used in psychology to reflect an individual's overall evaluation or appraisal of his or her own worth (Shavelson, Hubner, & Stanton, 1976). Nowadays, it is a highly discussed topic because it plays an important role in understanding human behavior, and is an essential personality construct. Lack of healthy self-esteem is related to many psychological dysfunctions (Donnellan, Trzesniewski, Robins, Moffitt & Caspi, 2005). There is abundant researches from diverse cultural backgrounds that has examined the relation between self-esteem and subjective well-being, depression and even romantic behaviors (Orth, Robins, & Meier, 2009).

However, evidence suggests that levels of self-esteem may vary from culture to culture. It has been largely reported that Asian people have lower average self-esteem than people from U.S, U.K. and other individualistic countries (Heine, Lehman, Markus, & Kitayama, 1999; Schmitt & Allik, 2005; Wang & Ollendick, 2001). Some researchers even claimed that it was a less important construct for Asian cultures (Heine, et al., 1999) . Why does this happen? This is because self-esteem is individuals' evaluation about their self-concept and self-image. Generally, people can form them by two ways: feedback from others as well as social interactions in different social roles (González-Pumariega, & García, 1997, Martín-Albo, Núñez, Navarro, & Grijalvo, 2007).

Criteria for self-esteem are different from culture to culture. According to Tsai, Ying, & Lee (2001) and Marks & Kitayama (1991), culture can also form one's view

and standpoint in assessing himself. For instance, some people view themselves to be separate from others whilst others see themselves to be part of a group, and are strongly connected with friends, families and co-workers. The former situation is typical in America and West European countries, where individualism is the dominant culture, the latter one is the mainstream in Asian countries, such as China, Japan and Korea. In a literature review, Tsai, et al. (2001) summarized that people in individualistic countries tend to demonstrate their uniqueness in front of others, and thus are more likely to announce that they are "superior to others", whereas Chinese and Japanese people may put more emphasis on maintaining relationship with others, accordingly, they are more likely to act interdependently, and inclined to show modesty and make people think they are "inferior to others". It is also widely reported that due to these reasons, Asian people in general show lower self-esteem than people from western cultures (Cai, Brown, Deng, & Oakes, 2007). It is even said that self-esteem is unique to western cultures (Heine, et al., 1999).

As a matter of fact, supposing that translation error does not exist, discrepancies of self-esteem in different cultural backgrounds can be understood in several ways. First, people develop different levels of self-esteem under different cultural context. Second, it is also possible that the inconsistency might due to measurement non-invariance across cultures. Third, discrepancies of self-esteem reported in literature may not due to real divergences of how people feel about themselves, but is because of self-presentation (Tsai, et al., 2001). No matter what reason is, there is a fierce debate about whether self-esteem is universal across culture, details will be illustrated below.

1.2 Is there a universal need for self-esteem?

There is much controversy over whether the concept of self-esteem is meaningless and measurement is futile in Asian countries. Proponents, such as Heine, et al. (1999), claimed that self-esteem was "not a universal, but rather is rooted in significant aspects of North American culture" (p. 766). They express their ideas through psychological, philosophical and anthropological aspects. As to the psychological point, first, they find that the distributions of the global self-esteem score in western world are profoundly positively skewed; medians and means are all above the theoretical midpoint (e.g. Heine & Lehman, 1999). However, the situation is strikingly different from Japan; the scores there formed a normal distribution. Other psychologists also found similar results, for instance, Diener & Diener (1995) reported that Japanese people's average score is close to conceptual midpoint, significant lower than people from north America. Second, self-esteem has been well proved in western cultures that it was significantly correlated with psychological distress, however, in Heine, et al.'s(1999) study, they fail to find this relationship in Japanese culture. Hence, they concluded that self-esteem was not a necessary concept for Asian countries.

There are plenty of scholars objected to this idea, they believe that self-esteem can be universal although slight discrepancies may exist (e.g. Sedikides, Gaertner, & Vevea, 2005; Cai, Wu & Brown ,2009). In a research by Cai et al. (2009), they found an opposite results in People's Republic of China. In view of the points proposed by Heine, et al.(1999), they did a meta-analysis and collected hundreds of empirical research in which self-esteem was assessed. These self-esteem scales including RSES, Coopersmith Self-esteem Inventory (CSEI), and other frequent used questionnaires. The findings suggested that just like individualistic countries, Chinese people revealed a positive bias of self-esteem distribution. As Rosenberg Self-Esteem Scale (RSES;

Rosenberg, 1965) is the most used self-esteem measurement in China, researchers did a similar research by examining studies that only used RSES scale, and again, the distribution was positive. Besides, subjective well-being was also found to be strongly related with self-esteem. This relation was also confirmed by other researchers, who suggested that self-esteem was not only correlated with mental and physical health, but also could be a good predictor of one's health in Hong Kong Chinese young people (Li, Chan, Chung, & Chui, 2010). Thus, it seems that Heine, et al.'s argument has been refuted, the concept can also be applied to some Asian cultures.

There are quite a lot instruments that were designed to measure one's global self-esteem, the most used one is Rosenberg Self-Esteem Scale.

1.3 The Rosenberg Self-Esteem Scale

RSES has only ten items, and is quite convenient to measure one's global self-esteem. Item 1, 3, 4, 7, 10 are positively worded, and the rest are negatively worded. Participants respond to the items on a four-point Likert scale: from strongly agree, agree to disagree and strongly disagree. In positive items, expressing "strongly agree" scores 4, whilst choosing "strongly disagree" gets 1 point. In negative-worded items, it scores in a reversed way (get 4 points for "strongly disagree", 3 points for "disagree", 2 points for "agree" and 1 point for "strongly agree"). There are also some researchers tend to extend the range from 1 to 6, or squeeze it from 0 to 3, but the theory is the same.

Validity and reliability of RSES have been widely discussed by researchers. Empirical evidences suggest that although RSES is brief, it can well test one's self-esteem. Most studies were conducted in individualist cultural background. The original sample for this scale was 5024 high school students randomly selected in 10 schools in New York, and gradually, researchers extended the area from high school

students to adults and from United States to other countries, such as Canada and UK. Researches find RSES to be highly reliable and valid, especially in North America and some European countries (Carmines & Zeller, 1979; Gray-Little, Williams & Hancock, 1997, Martín-Albo, Núñez, Navarro, & Grijalvo ,2007). In a report by Gray-Little et al. (1997), they concluded that "(RSES) deserves its widespread use and continued popularity" (p. 450).

Along with measurement equivalence of the scale, dimensionality of RSES has also been widely discussed. The original scale was designed to be a single factor structure, however, more and more researches indicates that it might be two-dimensional (Goldsmith,1986; Owens, 1993; Greenberger, Chen, Dmitrieva, & Farruggia, 2003). Besides, there are two axes regard to the factors: (1) self-competence factor and self-liking factor. Self-competence is the sense of an individual's ability derived from experiences of successful life experience (Tafarodi & Swann, 2001). Self-liking, on the other hand, is a kind of subjective evaluation of personal worth, not related with external circumstances, but is purely due to internalized criteria of social worth (Tafarodi & Swann, 2001). (2) Positive self-image factor and negative self-image factor. The negative self-image is reported to relate with depression, eating disorders, suicidal attempts and other relative dysfunctional behaviors, whilst positive self-image is associated with positive views about oneself (Bjorck, Clinton, Sohlberg, & Norring, 2007; Friedman, Terras, Zhu, & McCallum, 2004; Hamm, 2009). In cross cultures studies, it is possible that average positive score or negative score might be different, because it is less socially appropriate to express self-enhancement in Asian countries than Western culture due to Confucianism. Thus, average negative score might be lower in these countries (Farruggia, et al., 2004).

1.4 Mandarin version of RSES

RSES has been translated to dozens of languages and used in a number of countries, in most of which reveals high reliability and validity (Schmitt & Allik, 2005). In China, it is also the most widely used self-esteem questionnaire, using "zi zun (in Chinese characters)", which means self-esteem, as a keyword retrieved about 1100 papers from 1994 to 2008, and 7 out of 10 used Rosenberg Self-esteem Scale as measuring instrument.

The first Mandarin version of the Rosenberg Self-esteem Scale was translated and used by Ji and Yu (1993). Since then, Chinese scholars started to use this scale widely in various areas. However, because the RSES was developed in western cultural framework, and validation process was carried out in the US and European countries (individualistic culture), therefore, it might be problematic when it comes to the collective countries. Researchers in recent years become aware of the measurement equivalent issue of RSES in different cultures.

As to Chinese part, Ji and Yu's (1993) paper is really difficult to find, because it was published in a supplementary issue and do not have an electronic version. Thus it is hard to know how they translated and validated the scale. Other psychologists also did some validating study in China, but most were carried out in Hong Kong and Taiwan, both of which were special administrative regions (Cheng & Hamid, 1995; Schmitt & Allik, 2005). However, due to historical and economic reasons, the two places cannot well represent the whole China.

There are no much empirical researches related to measurement invariance of RSES in Mainland China. The only one I found was conducted by Farruggia, et al.(2004), and results suggested an identical factor structure across four countries (United States, Czech Republic, China and Korea) after deleting the eighth item.

However, other researchers disapproved this idea. For instance, Wang & Ollendick (2001) suggested that there was no equivalent expression of self-esteem in Chinese language, let alone measurement invariance.

Generally speaking, most researchers do believe RSES can be used in China, but one issue should be solved before widespread application of this scale, that is the use and applicability of the eighth item ("I wish I could have more respect for myself"). It seems that because of language differences, people in China have different understandings, which may lead to some potential problems. Three ways were proposed to deal with this item (Shen & Cai, 2008). The first one is to delete it directly, but according to some researchers ,for example, Schmitt & Allik (2005), they indicated that the Rosenberg Self-esteem had been so widely translated and used (53 countries and 28 languages), deleting an item would render alteration in factor structure. The second method is to revise and translate it in an acceptable way, however, this method also met some pressure, as the suggested new translated version make people confused by leading to different understandings (Shen & Cai, 2008). A final one is to score it reversely, that is, if people choose "Strongly agree" on the eighth item in China, they would get 4 points; whilst in other countries like U.S., subjects would get 1, and this reversed scoring is the most used method in China. But this solution might also lead to alteration in factor structure, which can be inferred from standpoint raised by Schmitt & Allik (2005).

Despite of the scoring system and whether or not to keep the eighth item, there seem two slightly different versions in use at present. The discrepancy is due to a minor difference in understanding and translation of a phrase ("at times"). Therefore, two items (item 2 and item 6) which contain this phrase are suspicious of causing problems and debate in research. It is impossible to give exact example which paper use which

version, as the scale usually did not appear in the appendix part. But due to online search, the two versions do exist. In this case, it is reasonable to assume that researchers may use both versions at the same time.

To sum up, Chinese version of RSES has been widely used in China, but there might be some potential problems. First, the eighth item seems to be interpreted in a different way from English-speaking countries. Besides, measurement invariance is rarely examined.

According to a large-scale study by Suh, Diener, Oishi, & Triandis (1998), China is one of the most collectivistic countries, as in a ten point scale (1 – 10, 1 represents the most collectivist countries, and 10 indicates most individualistic countries), China got only 2 points and is a typical collectivistic country, whilst Great Britain, had a score of 8.95 and can well represent highly individualistic culture.

As there was no much measurement invariance study carried out in China, the present research would enrich Chinese literature by examining cross-cultural invariance of the Rosenberg Self-Esteem Scale was analyzed across in Britain and China, which can well represent individualistic culture and collectivistic culture. Samples were college students from United Kingdom and People's Republic of China.

2. Method

2.1 Participants

Participants were college students from United Kingdom (n =150) and People's Republic of China (n =205). The samples were both approximately half male and half female. For Chinese samples, there were 109 females and 96 males; and for U.K. participants, there were 85 females and 65 males. The Chinese data were collected from the northern part of China, mostly from Jilin Province (Jilin University and Northeast Normal University), whilst U.K. data were collected from the University of

Edinburgh, and the majority of them were native British people, the rest were composed of European students.

2.2 Procedure

Ethical approval was approved by the Ethics Committee before data collection. In the study, subjects were required to fill an anonymous and self-report questionnaire, which is the Rosenberg Self-Esteem Scale. It typically takes about 1 to 2 minutes to finish it. Prior to their consent, they were told that they had the right to omit or refuse to answer any questions; and if they did not want to get involved in the study, they could simply not return me the questionnaire. About 95% agreed to participate in. During the collecting process in the University of Edinburgh, because there were a lot of international students, their nationalities would be asked about.

2.3 Measures

2.3.1 Translation of the scale

The translation of the Rosenberg Self-Esteem Scale followed a standard way of cross-cultural study method (Meadows, Bentzen & Touw-Otten, 1997). First, a bilingual person translated the English version of Rosenberg Scale into Mandarin, and then a second person who was not familiar with self-esteem scale back translated it into English. The third step was to let another person to check semantic differences between the original version and the back translated version to see whether they were equivalent. These steps were repeated several times to minimize bias, and finally, a parallel mandarin version established. This new scale was identical to one of the present versions. Thus, the Chinese version of RSES established. Cronbach's alphas for the samples were .832 for Chinese samples and.779 for U.K. participants, indicating the scale was reliable in both groups.

2.3.2 Plan of Analyses

The major purpose of the present research was to study measurement invariance of RSES in different ethnic groups. To make the comparison meaningful, factor structure was first measured. In general, the analysis mainly used Structural Equation Modeling, and it involved the following steps. To begin with, confirmatory factor analysis was performed to each group to test factor structure of the scale, this step aimed to find a baseline model. Afterwards, multi-group analysis between Chinese and British samples would be conducted to see measurement difference in scale level. And finally, because confirmatory factor analysis did not provide thorough analysis at item level, therefore, differential item functioning (DIF) and response patterns would be examined to see whether items were equivalent across groups. In order to do this, I followed Zumbo's (1999) and Byrne & David's (2003) method respectively. Details are illustrated next.

Before analysis was conducted, missing data was checked, as there were only ten items, missing one item means 10% loss, therefore, uncompleted questionnaire was drop from the data pool. Nine cases were dropped from Chinese data due to unfinished issues, and 12 were omitted from British sample. Finally, it left 205 Chinese and 150 U.K. samples.

Second step was to test measurement invariance. Nowadays, scholars tend to use several ways to see whether a measuring instrument can be used across cultures, such as principal component analysis, confirmatory factor analysis and exploratory factor analysis (Milfont & Fischer, 2010). The present study will use multi-group confirmatory analysis to assess measurement equivalence of the scale across the two ethnic groups, and it will be conducted in Amos 16.0.

Confirmatory factor analysis employs the theory and idea of "measurement invariance model" (Anderson & Gerbing, 1988). There are four levels of invariance,

(1) configural invariance, (2) weak factorial invariance, (3) strong factorial invariance and (4) strict factorial invariance (Meredith, 1993). Configural invariance is the basic one which made no assumptions that construct is equivalent across the ethnic groups. More specifically, factor loadings, intercepts and measurement residuals can vary freely (Meredith, 1993; Stein, Lee, & Jones, 2006), but the items load on the same factors in each group. This step is to see whether the underlying factor structure is identical. If this prerequisite is met, increasingly restrictive models will be measured next. Weak factorial invariance is to see the situation if factor loadings can be set invariant across cultural groups without loss of model fit, whereas strong factorial invariance fixes one set of additional limitation on the basis of weak factorial invariance model, and both intercepts and factor loadings are set equal across groups. Strict factor invariance, as can been inferred from its name, is the highest standard for metric invariance test. Factor loadings, intercepts and unique factor variances are all constrained, only factor means is allow being different across groups.

However, in real cases, it is not always so lucky that scales are equivalent across groups on all levels; sometimes researchers find them to be invariant on factor loading level but not on intercept level. If the scale turns to be equivalent after releasing several parameters, in this circumstance, partial measurement invariance is demonstrated.

In order to evaluate model fits, several fit indexes were inspected: (1) the Tucker-Lewis Index (TLI), (2) comparative fit index (CFI), (3) Chi-square value, (4) the ratio of the chi-square to degree of freedom ($\chi 2 / df$), (5) standardized root mean squared residual (SRMR), (6) root mean square error of approximation (RMSEA) and (7) the Akaike information criterion (AIC) and Bayesian information criterion (BIC), which indicate the simplicity of a model. TLI and CFI range from 0 to 1, and is the

higher the better, they can represent good fit if it equals or larger than .95 (Hu & Bentler, 1999). Actually, .90 would be an acceptable level (Marsh, Hau, & Wen, 2004). With respect to the ratio of the chi-square to degree of freedom ($\chi 2 / df$), if the value is smaller than 2, then the model can well fit the data. Some researchers also consider it acceptable if it is less than 3 (Mavondo & Farrell, 2000). For SRMR and RMSEA, they test how bad a model fit the data from different aspects. The majority of research materials suggest that they should be less than .06 and .05 respectively, but there are also some literature indicate that .08 and .06 is more practical in daily research (Marsh, et al., 2004).

3. Results

3.1 Descriptive statistics

Preliminary study of the data was conducted in SPSS 16.0. For item scores, Skewness values were smaller than |.73| and kurtosis values were smaller than |1.16| for both groups (see Table 1). With regard to the full scores, skewness value was -.23 and kurtosis was .14 for British group, and -.16， -.19 for Chinese group respectively, falls below the cut-off values for severe nonnormality (skewness values > 2 and kurtosis values > 7) proposed by Curran, West, & Finch (1996) . Thus, maximum likelihood can be used in the frame of confirmatory factor analysis. The Chinese version is presented in the appendix part (Appendix 1).

3.2 Multi-group confirmatory factor analysis

3.2.1 Establishment of baseline model for each group

Rosenberg Self-Esteem Scale was originally conceptualized as single-factor structure, yet later on, some of scholars suggested it to be two dimensional. Therefore, models that represent these ideas were first tested (in Amos 18.0).

This step is to find a model that can best fit both groups separately. Confirmatory factor analysis is a theory-based method, and according to the literature, RSES is reported to be one-dimensional or two-dimensional. Specifically, there are two axes regard to the factors: (1) a positive self-image factor (item 1, 3, 5, 7, 10) and a negative self-image factor (item 2, 5, 6, 8, 9); and (2) self-competence factor (item 1,

2, 3, 4, 5) and self-liking factor (item 6, 7, 8, 9, 10). Four models were established and assessed (factor structure showed in Figure 1).

Table 1 *Descriptive Statistics of the two countries*

Items	China				U.K.			
	M	SD	Skewness	Kurtosis	M	SD	Skewness	Kurtosis
1. On the whole, I am satisfied with myself.	3.12	.69	-.25	-.56	3.16	.52	.19	.28
2. At times, I think I am no good at all.	2.82	.89	-.33	-.61	2.64	.73	.46	-.66
3. I feel that I have a number of good qualities.	3.12	.57	-.14	.76	3.42	.53	-.08	-1.16
4. I am able to do things as well as most other people.	3.16	.59	-.06	-.29	3.31	.57	-.09	-.59
5. I feel I do not have much to be proud of.	2.66	.70	.14	-.40	3.27	.66	-.50	-.09
6. I certainly feel useless at times.	2.97	.82	-.32	-.63	2.58	.74	.34	-.43
7. I feel that I'm a person of worth, at least on an equal plane with others	3.11	.61	-.20	-.17	3.34	.53	.12	-.87
8. I wish I could have more respect for myself.	1.5	.57	.73	.49	2.69	.46	-.47	.32
9. All in all, I am inclined to feel that I am a failure.	3.21	.70	-.66	-.50	3.36	.62	-.41	.65
10. I take a positive attitude toward myself	3.18	.59	-.07	-.32	3.11	.59	-.02	.13
Total score	28.85	4.29	-.16	-.19	30.88	3.60	-.23	.14

Model 1 tested all the 10 items, and assumed self-esteem to be one-dimensional. Model 2 kept all the items, but supposed the scale was consisted of two factors (self-competence and self-liking), whilst Model 3 assessed negative and positive self-image factor. Model 4 added one error covariance between item 2 ("At times, I think I am no good at all.") and item 6 ("I certainly feel useless at times.") on the basis of model 3. According to Brown (2006), correlated item can be specified

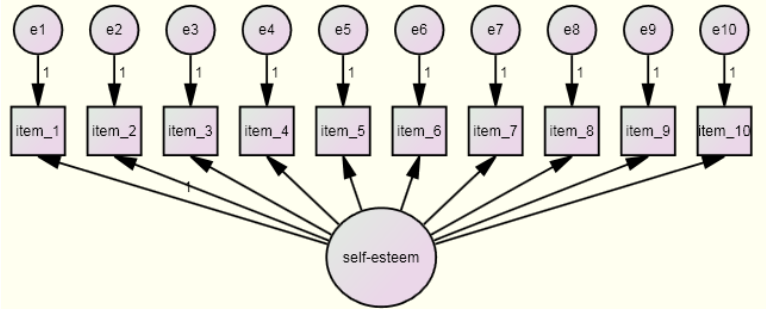
when items are "very similarly worded, reverse-worded, or differentially prone to social desirability, and so forth" (p. 181). In this case, "no good at all" and "useless" were quite similarly worded, thus it complies the requirement.

When comparing which model better fit the data, Researchers usually examine (1) chi-square difference ($\Delta\chi2$) between models, if P<.01, then the latter model is better. However, there are more and more researchers noted that $\chi2$ is too sensitive, thus (2) ΔCFI is estimated as well, if ΔCFI>.01, it suggest that the second model is a significant improvement (Cheung & Rensvold, 2002).Besides, (3) AIC and BIC which represent model parsimony were also compared between models. Sometimes, these criterions give inconsistent results; it is up to researchers to decide which parameter to rely on. In Byrne's book (2009), he said: "the decision of which one ($\Delta\chi2$ or ΔCFI) to accept is purely an arbitrary one and rests solely with each individual researcher. It seems reasonable to assume that such decisions might be based on both the type of data under study and/or the circumstances at play." (P 223).

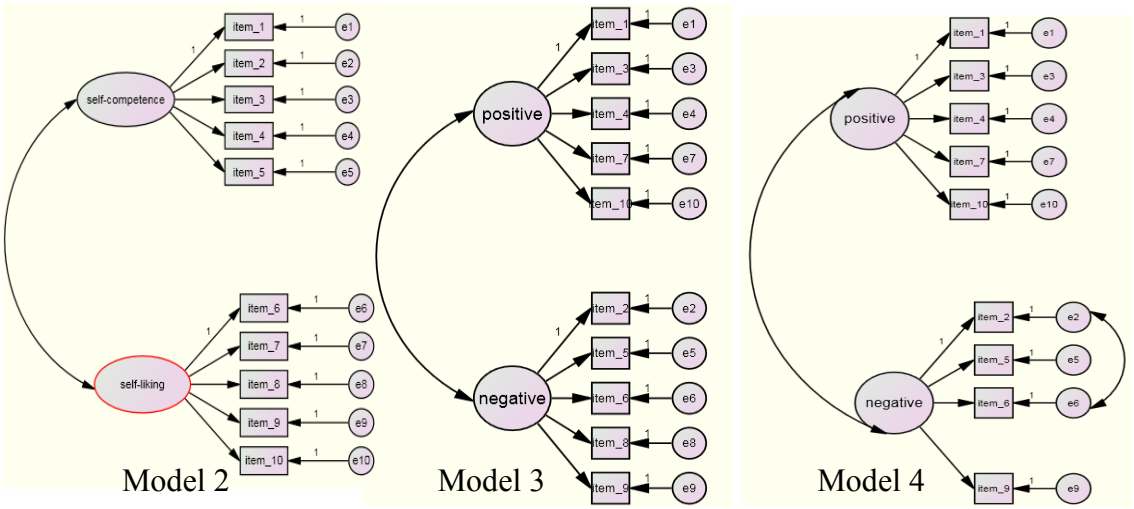
Details are listed in Table 2. Single-factor model (model 1) fitted both Chinese and British awfully. Chi-square values ($\chi2$) for the model fit were significant, $\chi2_{Chinese}$ (35) = 129.20, P < .001, $\chi2_{U.K.}$ (35) = 77.05, P < .001. The ratio of the chi-square to degree of freedom ($\chi2$ / *df*) was 3.691. As $\chi2$ is very sensitive to sample size, other fit indexes were also considered. CIF = 8.69, RMSEA = .115, SRMR = .067, all them fell to meet the minimum acceptable criterion. The situation for British sample was similar, although $\chi2$ / *df* was close to 2, other goodness of fit-indices were far from

good (CFI = .856, RMSEA= .090 and SRMR = .077). Therefore, suggesting a lack of fit between the one-factor model and data for both ethnic groups.

Figure 1



Model 1



Model 2 Model 3 Model 4

Two-factor model was estimated next (model 2, self-competence and self-liking factors), and the results did not show a significant improvement, as $\chi 2$ was almost the same with model 1 ($\chi 2$ = 129.159). Other parameters also expressed the same idea, for British group, CFI = .868, RMSEA= .117 and SRMR = .067, as to Chinese group, CFI = .857, RMSEA= .091 and SRMR = .077. Both the two ethnic groups had a larger AIC and BIC values in this model than that of the previous single-factor model.

With respect to Model 3 (positive/ negative self-image factors), compared with model 1, this model proved to be a better choice for both Chinese and British groups. For Chinese group, $\chi 2_{\text{Chinese}}$ (34) = 85.24. CFI, SRMR and RMSEA all reached to an acceptable value (CFI = .93, GFI = .92, SRMR = .06, RMSEA = .09). The change of chi-square ($\Delta\chi 2$) exceeded critical Chi-square values ($\Delta\chi^2_{\text{Chinese}}$ (1) = 43.96, P < .01). ΔCFI = .06 > .01, again showed significant improvement. The two-factor solution greatly enhanced model fit for British data as well ($\Delta\chi^2_{\text{U.K.}}$ (1) = 5.244, P < .05; ΔCFI = .01). These findings were taken as supporting the two-factor structure of Rosenberg Self-Esteem Scale, besides, AIC and BIC were smaller in this model for both group, thus, all the evidence suggested that model 3 was significantly better than model 1. However, it still did not reach an ideal level; therefore, I examined factor loadings and modification indices for both groups.

Factor loadings were moderate for British samples, Item1, 3, 4, 7, 10 were set to load on the positive factor, λ_1 = .44, λ_3 = .58, λ_4 = .46, λ_7 = .63 and λ_{10} = .56, and , Item 2, 5, 6, 8, 9 were set to load on the negative factor, λ_2 = .40, λ_5 = .67, λ_6 = .45, λ_8 = .45 and λ_9 = .75. Factor loadings seemed well for Chinese group except for the eighth item, λ_1 = .59, λ_3 = .66, λ_4 = .73, λ_7 = .66 and λ_{10} = .77; whereas on the negative factor, λ_2 = .75, λ_5 = .64, λ_6 = .72, λ_8 = -.10 and λ_9 = .77. Because item 8 loaded oppositely on negative self-image factor (British group: λ = .45, Chinese group: λ = -.10), and modification indices also suggest this item profoundly hampered the overall fit for the model, thus it was deleted. Besides, there was a large error covariance

between item 2 and item 6 for both groups, 16.99 for Chinese data and 22.31 for British data. In this case, the covariance between these two items needed to be set free. So, aside from deletion of the 8th item, an error covariance was added to the new model (model 4).

Table 2 Summary of fit statistics for Measurement Models

Model	df	χ2	χ2/ df	CFI	RMSEA	SRMR	AIC	BIC
Model 1								
China	35	129.201	3.691	.869	.115	.067	169.201	235.661
U.K.	35	77.050	2.201	.856	.090	.077	117. 050	177.262
Model 2								
China	34	129.159	3.799	.868	.117	.067	171.159	240.942
U.K.	34	75.909	2.233	.857	.091	.077	117.909	181.133
Model 3								
China * *	34	85.240	2.507	.929	.086	.058	127.240	197.024
U.K. *	34	71.766	2.111	.871	.086	.075	113.766	176.989
Model 4								
China * *	25	61.744	1.871	.960	.065	.048	108.598	-
U.K. * *	25	47.605	1.443	.950	.052	.055	91.287	-

Model comparisons	Δdf	Δχ2	P value	ΔCFI
Models 1 and 2				
China	1	.042	P< .85	.001
U.K.	1	2.141	P< .15	.000
Models 1 and 3				
China	1	43.961	P< .001	.060
U.K.	1	5.244	P< .05	.015
Models 3 and 4				
China	9	23.496	P< .01	.031
U.K.	9	24.161	P< .01	.079

* estimates of model comparison are significant at .05 level

* * estimates of model comparison are significant at 0.01 level

"-" Amos did not calculate BIC for model 4

Model 4 was the best fit model for both Chinese and British college students. For Chinese group, $\chi 2_{Chinese}(25) = 50.598$, and $\chi 2 / df < 3$. Besides, goodness of fit

indices also indicated that the model fitted Chinese data well (e.g. CIF = .964, SRMR = .044 and RMSEA = .071). Besides, $\Delta\chi^2_{Chinese}$ (9) = 34.642, P < .01 and ΔCFI = .045, strongly suggested it significantly better than model 3. As to the British data, $\chi 2_{U.K.}$ (25) = 49.833, and $\chi 2 / df = 1.577$, other indices also attained a good level, CIF =. 975, SRMR = .055, RMSEA = .052. Once again, both $\Delta\chi^2$ and ΔCFI fell into the criteria, $\Delta\chi^2_{U.K.}$ (9) = 21.933, P < .01 and ΔCFI = .094; AIC had the smallest value among all these models, $AIC_{U.K.} = 91.287$ and $AIC_{Chinese} = 108.958$. From this sense, model 4 could fit both groups well; therefore, configural invariance can be demonstrated.

In short, Rosenberg Self-Esteem Scale could better fit the data in a two-dimensional structure, and it is consisted of a positive self-image factor and a negative self-image factor. Besides, model 4 has the best fit and thus would be used as baseline model in the following analysis.

3.2 Multi-group Confirmatory Factor Analysis

3.2.1 Research model and design

The findings of the above analysis indicated that the two groups have a similar factor pattern on Rosenberg Self-Esteem Scale, thus the configural invariance was demonstrated. To further estimate invariance, several hierarchical nested models were tested and compared: (1) model A: unconstrained model, (2) Model B, in which factor loadings (measurement weight) were constrained to be equal in the two groups, (3) Model C, in which factor loadings and intercept were set equal across the two

countries, and finally (4) Model D, in which factor loadings, intercept, and indicator variances are constrained to be the same.

At the very beginning, validity of baseline model was tested. According to Byrne (2004, 2009), this step is strongly recommended, as even if the baseline model can well fit every dataset respectively, it may not be the case for simultaneous testing. Thus, some researchers use the simultaneous assessment to determine configural invariance (e.g. Michaels, Barr, Roosa, & Knight, 2007; Bryne, 2008). Besides, the results would be different from single group tests, as in multi-group analysis no matter how many groups are measured, only one set of statistics for overall fit would be generate. It is worth noting that Chi-square values and degree of freedom in multi-group analysis are summative, which means their values should equal to the sum of Chi-square and degree of freedom gained in separate analysis for each group (Byrne, 2004). If the results indicate that the baseline model can well fit both datasets, more stringent model tests would be formed and compared.

3.2.2 Model Fit

Several indices could be used to estimate model fit, (1) Chi-square difference (*Δχ2*) (Jöreskog, 1971, 1993; Byrne, 2001, 2009). For example, if *Δχ2* between Model A and model B reaches a significant level (i.e. $\Delta\chi2$ exceeds critical chi-square value), then it represents that the measurement is not equivalent across groups. In this case, researchers need to check relevant parameters to see where the problems lie, and make accordant alterations. Otherwise, if weak factorial invariance can be

demonstrated, model comparison continues. As $\chi 2$ is very sensitive to sample size, other indices are employed to test model fit. (2) ΔCFI, if ΔCFI < .01, it suggests no significant change between models, thus measurement equivalence can also be demonstrated (Byrne, 2008) (3) RMSEA, it stands for closeness of fit, and if is smaller than .08, the model is acceptable. Satisfaction of model B is the most important one, as it provide minimal evidence for measurement equivalence (Marsh,1994; Dishman et al., 2002). Strong factor invariance is also essential, as it "provide evidence that scale scores are on the same metric…and allows meaningful mean difference comparisons to be made across groups" (P 277).

3.2.3 Measurement invariance

The results of hierarchical nested models were presented in Table 3. The first step was to test the baseline model simultaneous in two groups with no constraints imposed. The model fit the data quite well (χ^2 (50) = 83.881, p = .002; CFI = 0.965; TLI = .952 and RMSEA = .044 (90% Confidence interval of .027 to .060). The unconstrained model showed good fit features, and again supported the basic requirement of configural invariance.

Weak factorial invariance between the two ethnic groups was estimated by constraining factor loadings of the same item to be equal across groups (model B, χ^2 (57) = 99.929, p < .001; CFI = 0.956; TLI = .945). Model fit was not significantly different from unconstrained model. Three criterions mentioned above were satisfied, $\Delta\chi 2_{(7)}$ = 16.048, P < .05; Δ CFI = .009 and RMSEA = .046 (90% Confidence interval of .031

to .061), therefore, null hypothesis was accepted, and weak factorial invariance was demonstrated.

Table 3 The results of measurement invariance tests of RSES for the two groups

Model	χ2	df	CFI	TLI	RMSEA (90% CI)
Model A (configural invariance)	83.881	50	.965	.952	.044(.027 - .060)
Model B (weak factorial invariance)	99.929	57	.956	.945	.046 (.031 - .061)
Model C (strong factorial invariance)	246.158	64	.814	.791	.090 (.078 - .120)
Model C_partial (partial strong factorial invariance)	105.078	59	.953	.943	.047 (.032 - .061)
Model D (strict factorial invariance)	270.399	74	.799	.805	.090 (.078 - .102)
Model D_partial (partial strict factorial invariance)	126.722	69	.941	.938	049 (.035 - .062)

Model C estimated the invariance of factor loadings and intercept across the two ethnic groups by placing equality constraints on these parameters. Goodness-of-fit indices suggest a poor fit χ^2 (64) = 246.158; p < .001; CFI = .814, RMSEA = .090 (90% confidence interval of .078 to .120). Compared to model B, $\Delta\chi2_{(7)}$ = 146.229, P< .001; Δ CFI = .142, suggesting some intercepts were not equivalent across countries and needed to be set free to reach a partial factorial invariance.

Model D in the sequence included constraining factor loadings, intercept and unique factor variances (measurement residuals) was rejected ($\Delta\chi2_{(10)}$ = 24.241, P< .001; CFI = .799, TLI = .805, RMSEA = .090 (90% confidence interval of .078 to .102)).

From above, the assumption of measurement invariance was not sufficiently supported across the two countries, as full equivalence only exist in weak factorial

level. Both strong and strict factorial models were rejected due to significant $\Delta\chi 2$ between continuous nested models, as well as inadequate CFI, TLI and RMSEA values. Therefore, partial strong and strict factorial invariance were examined next.

3.2.3 Partial measurement invariance

The intercept model above indicated that some intercepts should be freely estimated. In order to see which item intercept(s) was not equivalent across groups, we need to release intercept restrictions item by item, and each time only release one item's intercept. For example, after removing intercept constraint of the first item, 8 items' intercept left (model C_2). Compared to model C, which set 9 pairs of intercept equal across groups, model C_2 had only 8 pairs of restrictive intercept. Then multi-group confirmatory factor analysis was performed to estimate chi-square of model C_2. If the difference of chi-square between model C and model C_2 is insignificant, it suggests that item is not statistically different across the two groups. This procedure was repeated item by item, until all the items were examined.

The results were displayed in Table 4. The chi-square differences implied 5 items out of 9 were not equivalent on intercept level across groups (item 3, 5, 6, 9, 10; $P< .01$). Based on these findings, intercept restrictions of the five items were relaxed to meet partial strong factorial invariance (model C_partial, table 3). In this model, Goodness-of-fit indexes reached a good level, CFI = .953, TLI = .943, RMSEA = .047 (90% confidence interval of .032 to .061); compared with model B, $\Delta\chi 2_{(10)} = 5.149$, $P< .88$. All of the measurement invariance criteria were complied, thus partial

strong factorial invariance was demonstrated. Partial strict factorial invariance was also accepted, and the results are listed in Table 4 (CFI = .941, TLI = .938, RMSEA = .049 (90% confidence interval of .035 to .062)).

Table 4 Equivalent and Nonequivalent Intercepts of Items across Countries

Item	Related Factor	$\chi 2$	$\Delta\chi 2$(df = 1)	Probability
1	Positive self-image	243.532	2.626	P < .10
3	Positive self-image	229.870	26.288	P < .001 *
4	Positive self-image	246.157	.001	P < .98
7	Positive self-image	241.951	4.207	P < .05
10	Positive self-image	222.061	24.097	P < .001 *
2	Negative-image	242.145	4.013	P < .05
5	Negative-image	175.977	70.181	P < .001 *
6	Negative-image	211.636	34.522	P < .001 *
9	Negative-image	233.343	12.815	P < .001 *

*Significant at p<.01

Provided with the findings of partial measurement equivalence by confirmatory factor analysis, next step is to see why certain items are not equivalent, item-level analyses were performed. First, item bias was estimated by ordinal logistic regression modeling, following the procedures recommended by Zumbo (1999). Then, response patterns for the non-invariant items were assessed and compared between the two countries (Cheung & Rensvold, 2000).

3.3 Ordinal Logistic Regression Modeling to test DIF

Test items are expected to function equivalently across groups, that is, irrelevant to gender, ethnic group of the test takers. If people with similar abilities from diverse groups tend to give different responses to a measuring instrument at the same level of the trait， then it means this item functions differently, and has differential

item function (DIF) (Hambleton, Swaminathan, & Rogers, 1991; Zumbo, 1999). According to Johnson, Spinath, Krueger, Angleitner, & Riemann (2008), "This (DIF) takes place because the measurement instrument is not completely unidimensional: It means that individuals from the two samples will have different probabilities of endorsement for the DIF items and thus are likely to have different sum scores on the measure creating a potentially misleading indication of sample differences in the trait when evaluated using sum scores." P 673

In order to make a better understanding on which item did not performed equivalently across groups, DIF analyses would be carried out. Besides, according to Roth, Decker, Herzberg, & Brähler (2008), confirmatory factor analysis makes normal curve assumptions, however, some evidence suggested that item distribution of Rosenberg Self-esteem Scale was suffer from the debate of bimodal distribution. Form this sense; (DIF) analyses were carried out to see measurement invariance on item level. Compared with confirmatory factor analysis, DIF can provide detailed information on item-level than confirmatory factor analysis. DIF refers to one single item, and when a cluster of items are examined, the Differential item functioning can be extended to differential test functioning (DTF) (Abad, Francis & Hills, 2008).

In order to perform DIF, the test should be unidimensional, whereas in the previous part, it has been demonstrated that RSES can better fit the data use a two dimensional solution. But in this case, as there are only nine items left in RSES, it is impractical to split it up and tested the two dimensions separately, because this would

lead to extreme narrow construct and thus make the analyses meaningless (Johnson et al., 2008). Besides, according to Tate (2003) and Abad et al. (2008), "there is no single, recognized test for unidimensionality testing". Therefore, Stout (1987) suggested that DIF could be performed if there was a single dominant factor. In order to prove this DIF analyses can be performed in RSES, a principal component analysis was conducted, it turned out that the first factor accounted for 39.24% variance and the second one took up to 14.26%. Therefore, it suited the criterion, and DIF of Rosenberg Self-Esteem Scale can be analyzed.

The analysis followed Zumbo's (1999) method, ordinal Logistic Regression Modeling was conducted in SPSS 16.0. According to Zumbo (1999), this method is especially suitable to detect non-uniform DIF. There are two forms of DIF: uniform DIF and non-uniform DIF. Uniform DIF occurs when the probability of endorsing an item is different, but this discrepancy holds constant over ability levels, whilst in non-uniform DIF items, there are interactions between group and capacity level. For instance, if people from different social groups with the same extraversion level tend to response differently to an extraversion-introversion item, then this item shows a DIF. If people with lower extraversion consistently have the same odds in endorsing this item, then uniform DIF occurs, whilst if this item favors people in one group on certain level, but a different level in another group, then, it exhibit non-uniform DIF.

There are 3 steps to estimate DIF: "Step #1: One first enters the conditioning variable (i.e., the total score), Step #2: enter group variable, and finally Step #3: The

interaction term is entered into the equation" (P26). Effect size can be calculated by the deducting R-squared$_{step1}$ from R-squared$_{step3}$ to measure uniform and non-uniform DIF (Zumbo, 1999)

Table 5 listed R^2 which were derived from the three steps mentioned above: (1) total score, (2) total score and group and (3) total score, group and the interactions between test score and group (Zumbo, 1999). According to cut-points raised by Jodoin and Gierl (2001), DIF can be divided into three levels: Negligible or A-level DIF (ΔR^2 < .035), Moderate or B-level DIF（$.035 \leq \Delta R^2 < .07$）and Large or C-level DIF（$\Delta R^2 \geq .07$）. Null hypothesis should be rejected if change of R^2 reached moderate level.

Examining the difference between step 3 and step 1, it is obvious that items 2, 5, 6, 10 exhibited large DIF effect sizes, as ΔR^2=.115, .072, .189, .491, 070 respectively (Table 5). Comparing ΔR^2 value between step 2 and step 3, the data suggested that item 2, 5, 6 and 10 showed predominant uniform DIF. It seems college students with lower self-esteem in China tend to have the same probability in endorsing three out of four DIF items.

In general, DIF mostly exist in reversed scoring items (items 2, 5, 6 and 10), and all of them showed Uniform DIF but not Non-uniform DIF. Items 2, 6 and 10 favored Chinese students and item 10 favored British students. However, it should be pointed out that conducting DIF analysis requires large sample size; whereas the sample size in the present study is not adequately large, especially for the British group.

Table 5 summary for Logistic Regression Modeling

item	1	2	3	4	5	6	7	9	10
(1)	.388	.446	.501	.490	.577	.369	.556	.667	.571
(2)	.403	.520	.512	.491	.647	.550	.557	.680	.641
(3)	.416	.561	.521	.497	.649	.558	.558	.680	.641
ΔR2	.027	.115	.020	.007	.072	.189	.002	.013	.070
Category of DIF	A	C	A	A	C	C	A	A	C
Favored group	N	CN	N	N	U.K.	CN	N	N	CN
DIF type	-	uni	-	-	uni	uni	-	-	uni

(1) Only total score in the model, (2) Total score and Uniform DIF variable (group) in the model and (3) total score, Uniform DIF and Non-uniform DIF variable (interactions) in the model

CN "China"; A "no or negligible DIF"; B "slight to moderate DIF"; C "moderate to large DIF"

uni "Uniform DIF"; non "Non-uniform"

3.4 Psychometric Properties of the 9-item Version Scale

Item and scale features were measured again after the deletion of the eighth item, results are displayed below (Table 6 for British students and Table 7 for Chinese students and). With respect to the British group, both item-subscale correlations, item-scale correlations and internal consistency (Cronbach's α) reached to acceptable levels, $\alpha_{\text{positive subscale}} = .666$, $\alpha_{\text{negative subscale}} = .680$ and $\alpha_{\text{total}} = .768$, suggesting that the nine item scale was reliable. For Chinese group, both item-subscale correlations and item-scale correlations were in the upper range (greater than .5), internal consistencies for the two dimensions were high, indicating the scale is reliable, $\alpha_{\text{positive subscale}} = .811$, $\alpha_{\text{negative subscale}} = .809$ and $\alpha_{\text{total}} = .832$.

Table 6 Psychometric properties of British version of RSES (N=150)

Scale/item	α		r_{it}	Frequency of scores (%)			
				1	2	3	4
RSES-positive	.666						
Item 1	-	.340	.336	0	6.7	70.7	22.7
Item 3	-	.458	.426	0	2.0	54.0	44.0
Item 4	-	.393	.388	0	5.3	58.7	36.0
Item 7	-	.492	.495	0	2.7	60.7	36.7
Item 10	-	.416	.504	0	12.0	64.7	23.3
RSES-negative	.680						
Item 2	-	.451	.354	1.3	46.7	38.7	13.3
Item 5	-	.443	.559	.7	10.0	51.3	38.0
Item 6	-	.488	.418	3.3	46.7	38.7	11.3
Item 9	-	.462	.604	0	7.3	49.3	43.3
Total score	.768			-	-	-	-

r_{is} is the item-subscale correlation, r_{it} is the item-scale correlation

scores 1,2,3,4 represent Liker-type response. "1" strongly agree, "2" agree, "3" disagree and "4" strongly disagree for positive worded items (items 1,3, 4, 7, 10), whilst for negative worded items (items 2, 5, 6, 9). "1" strongly disagree, "2" disagree, "3" agree and "4" strongly agree

Table 7. *Psychometric properties of Chinese version of RSES (N=205)*

Scale/item	α	r_{is}	r_{it}	Frequency of scores (%)			
				1	2	3	4
RSES-positive	.811						
Item 1	-	.530	.504	.5	17.1	52.7	29.8
Item 3	-	.628	.548	.5	9.3	68.3	22.0
Item 4	-	.650	.622	0	10.7	62.4	26.8
Item 7	-	.571	.584	.5	12.2	62.9	24.4
Item 10	-	.634	.694	0	10.2	62.0	27.8
RSES-negative	.809						
Item 2	-	.698	.625	7.8	25.9	42.4	23.9
Item 5	-	.520	.593	2.4	39.5	47.3	10.7
Item 6	-	.643	.613	3.4	24.9	43.4	28.3
Item 9	-	.664	.672	2.0	10.2	52.7	35.1
Total score	.866	-	-	-	-	-	-

r_{is} is the item-subscale correlation, r_{it} is the item-scale correlation

scores 1,2,3,4 represent the same as above

4. Discussion

4.1 A summary of the findings

The purpose of this study is to examine factor structure and measurement invariance of Rosenberg Self-Esteem Scale in China and U.K., which represent two typical cultures: collectivism and individualism. The results suggested the scale to be two-dimensional and partially invariant across college students from Britain and China.

4.2 Compared with the results of previous studies

With respect to the dimensionality, the findings of present study were consistent with some of previous researches, including those performed in America, where the RSES scale was developed and originally validated (e.g. Goldsmith, 1986; Owens, 1993). And it is also accordant with some studies conducted in British and China (Farruggia, et al., 2004; Han, Jiang, Yang, & Wang, 2005Paterson, Power, Yellowlees, Park, & Taylor, 2007). However, some scholars believed that RSES was one-dimensional; the reason why two factors could be detected was due to the item wording effect (Cai, et al., 2007; Greenberger, et al., 2003). In Greenberger et al.'s (2003) research, they let undergraduate students finished one of three versions of RSES. These three versions were: original version with five positively worded items and five negatively worded items; the other two were scales in which all of the ten items were positively or negatively worded. Results showed that only the original scale displayed a dual self-image structure, the other two was found to be

one-dimensional. From this sense, it might be possible that two-factor construct might be "an artifact of the two types of item-wording (positive and negative) used in that scale." (P1252). Besides, it is also reported in their article that, somehow, both revised versions can reduce social desirable responding to some extent, and the negative worded scale was especially the case.

When estimating measurement equivalence, confirmatory factor analysis is the most common way that has been used. It assesses four levels of invariance: configural invariance, weak factorial invariance, strong factorial invariance and strict factorial invariance. Ideally, strict factorial invariance is expected to be satisfied, whereas, practical experience indicates that strong factorial invariance is more attainable (Steenkamp & Baumgartner, 1998; Byrne, Shavelson & Muthén, 1989). If an instrument cannot exhibit a partial strong factorial invariance, then it would be of little value to do cross-cultural comparison (Steven & Gregorich, 2006). In the present study, partial strong factorial invariance has been demonstrated, in which half of the item intercepts were released. This indicates British college students and Chinese students used a different metric in responding for about half of the items. Thus, it is not applicable compare mean levels across the two groups.

Generally speaking, there are three ways to deal with this partial measurement invariance situation (Millsap & Millsap, 2004). (1) Omit nonequivalent items; however, this might yield many diverse versions across different countries. (2) Compare the whole scale regardless of non-invariance, this method ignores the magnitude of

non-invariance. Besides, in order to make the scale comparable across different cultures, there should be a criterion in which minimum proportion of invariant parameters is specified. But it is also not practical to specify a certain cut-off point, as it might be quite arbitrary based on extant research. (3) Simply do not use the scale in cross-cultural comparisons. In this study, RSES is good for testing individual differences within each group, but is problematic to estimate differences between countries.

Taking a close look at the item level, both CFA and DIF analyses suggested several items were non-equivalent, and the most severe one was the 8th item, and thus, it was deleted in the final model. Concerning the possible reasons that might lead to non-invariant of this item, it may probably due to language discrepancies and thinking habits. People in China have a different understanding in this word "wish" (Farruggia, et al., 2004; Shen & Cai, 2008) . "I wish I could have more respect for myself" uses subjunctive mood, which suggest the opposite situation in reality in English. That is, only if someone does not have enough respect in real life, then he will crave more. In this case, there is no problem employing reversed scoring method in English. However, Chinese language does not have the subjunctive mood. In this case, "wish" simply means some kind of hope or desire, and has nothing to do with situations in actual life. In view of this, participants, no matter they are highly respected or not in real world, tend to choose "agree" or "strongly agree"(Shen & Cai, 2008). Therefore, some Chinese scholars tried to rephrase this word in the hope of expressing the same

meaning with the original item. In Shen & Cai's (2008), they also let subjects filled one of three versions of RSES, but different from the prior example which had all the items worded in a different way, Shen's study only changed the eighth item. One questionnaire used original expression, and the other two conveyed the 8th item in a slightly different way. It turned out the expression of "I think it is difficult for me to get more respect in the future" is cloest to the English meaning. However, the change was not quite accepted by Chinese scholars, as this version never appeared in later studies. It is possible that researchers believe it did not well rephrase the original item. As to the other non-invariant items (item 2, item 5, item 6 and item 10), they all exhibit Uniform DIF, that is, people with the trait level response differently.

Most cross-cultural researches of self-esteem related to Chinese people were carried out in Hong Kong and Taiwan, where some beliefs and values are different from Mainland China to some extent due to historical reasons. For instance, Schmitt & Allik (2005) found that people in Hong Kong, United Kingdom and United States responded comparably to Rosenberg self-esteem Scale. Whereas Cheng & Hamid (1995) found item 8 did not work well for People in Hong Kong. With respect to Taiwan, a research using differential item analysis examiend RSES across eitght countries, the findings suggested that self-esteem was not conceptualized in congruent ways, which was particularly the case between individualistic cultures (e.g. U.S.) and collectivistic cultures (e.g. Taiwan) (Baranik et al., 2008). Sometimes, immigrants were also tested, such as Chinese Americans (Russell, Crockett, Shen, & Lee, 2008).

However, their situations and ideas are also different because of unavoidable cultural assimulation.

Measurement equivalence of the RSES was rarely examined in Mainland China. As far as I searched, only one empirical study was found (Farruggia, et al., 2004). It was conducted in the 11th grade students across four countries (United States, Czech Republic, China and Korea). The results suggested a strict factorial invariance across four countries after omitting the eighth item. Thus, the findings of the present study were partial supported, as it showed configural invariance and weak factorial invariance, whilst failed to exhibit full strong and strict factorial invariance.

4.3 Strength and Weakness of the present study

This study estimated measurement invariance of RSES by several methods. First, Confirmatory factor analysis was used to test the factor structure across two countries. After detecting metric non-invariance (i.e. lack of strong factorial invariance), intercepts of certain items were set free to satisfy partial measurement equivalence. Besides, invariance test was also performed on item level, DIF were analyzed, and they can provide more detailed information, and were complementary to CFA.

There are several weaknesses of this study, but two of which are quite problematic. (1)Insufficient sample size. It is recommended that ratio of sample size to number of freely estimated parameters should be greater than 20:1 (Bentler, 1987), from this sense, each groups should have approximately 400 participants, whereas in this study,

only half of subjects attended. If sample size is not enough, there will be no sufficient power to detect measurement invariance. (2) Representativeness. The British data was collected from library during the holiday; it might be more representative for hard-working students rather than the whole British college students.

CONCLUSION

RSES has been widely used to measure one's global self-esteem across many countries, yet the issue of measurement equivalence was not raised until recent decades. Little research in this aspect has been carried out in China, and the findings were inconsistent as well. Besides, there were also some debates with regard to its dimensionality. This study suggested that two-factor solution was better than single dimension structure for both British and Chinese college students. The two factors are: positive self-image factor and negative self-image factor. Beside, evidence of this research indicated that RSES had the same underlying factor structure for the two ethnic groups; nevertheless, students from the two countries used a different metric in responding to about half of the items.

To, sum up, this study indicated RSES were not fully identical across Britain and China, thus, it should be cautious in future cross-cultural studies.

REFERENCES

Anderson, J. C., & Gerbing, D. W., (1988). An updated paradigm for scale development incorporating unidimensionality and its assessment, *Joumal of Marketing Research, XXV*, pp. 186-192.

Bentler, P. M. (2005). *EQS 6: Structural equations program manual.* Encino, CA: Multivariate Software.

Bjorck, C., Clinton, D., Sohlberg, S., & Norring, C. (2007). Negative self-image and outcome in eating disorders: results at 3-year follow-up. *Eat Behav, 8*(3), 398-406.

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-CulturalPsychology, 30,* 555-574.

Byrne, B.M., Shavelson, R.J., Muthén, B. (1989).Testing for the equivalence of factor covariance and mean structures: the issue of partial measurement invariance. *Psychological Bulletin, 105,* 456 – 466.

Byrne, B. M. (2004). Testing for Multigroup Invariance Using AMOS Graphics: A Road Less Traveled. *Structural Equation Modeling, 11*(2), 272-300.

Byrne, B. M. (2008). Testing for multigroup equivalence of a measuring instrument: A walk through the process. *Psicothema, 20*(4), 872-882.

Byrne, B. M. (Ed.). (2009). *Structural equation modeling with AMOS: Basic Concepts, Applications, and Programming* (2 ed.): routledge.

Byrne, B. M., & David, W. (2003). The Issue Of Measurement Invariance Revisited. *Journal of Cross-Cultural Psychology, 34*, 155 -175.

Cai, H., Wu Q., & Brown J. D. (2009). Is self-esteem a universal need? Evidence from The People's Republic of China. *Asian Journal of Social Psychology, 12*, 104-120.

Cai, H. J., Brown, J. D., Deng, C. D., & Oakes, M. A. (2007). Self-esteem and culture: Differences in cognitive self-evaluations or affective self-regard? [Article]. *Asian Journal of Social Psychology, 10*(3), 162-170.

Cheng, S. T., & Hamid, P. N. (1995). An error in the use of translated scales: The Rosenberg Self-Esteem Scale for Chinese. *Perceptual and Motor Skills, 81*, 431-434.

Chen, F. F., & West, S. G. (2008). Measuring individualism and collectivism: The importance of considering differential components, reference groups, and measurement invariance. *Journal of Research in Personality*, *42,* 259-294.

Cheung, G. W., & Rensvold, R. B. (2000). Assessing extreme and acquiescence response sets in cross-cultural research using structural equation modeling. *Journal of Cross-Cultural Research, 31*, 187-212.

Cheung, G. W., & Rensvold, R. B. (2002). Evaluating goodness-of-fit indexes for testing measurement invariance. *Structural Equation Modeling: A Multidisciplinary Journal*, *9*, 233–255.

Curran, P.J., West, S.G., & Finch, J.F. (1996). The robustness of test statistics to nonnormality and specification error in confirmatory factor analysis. *Psychological Methods, 1,* 16-29.

Diener, E., & Diener, M. (1995). Cross-cultural correlates of life satisfaction and self-esteem. *Journal of Personality and Social Psychology, 68,* 653–663.

Dishman, R. K., Motl, R. W., Saunders, R. P., Dowda, M., Felton, G., Ward, D. S. (2002). Factorial Invariance and Latent Mean Structure of Questionnaires Measuring Social-Cognitive Determinants of Physical Activity among Black and White Adolescent Girls. [doi: DOI: 10.1006/pmed.2001.0959]. *Preventive Medicine, 34*(1), 100-108.

Donnellan M.B., Trzesniewski K.H., Robins R.W., Moffitt T.E., Caspi A. (2005). Low self-esteem is related to aggression, antisocial behavior, and delinquency. *Psychological Science*, 16 (4), 328-335

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

Friedman, A. S., Terras, A., Zhu, W. Z., & McCallum, J. (2004). Depression, negative self-image, and suicidal attempts as effects of substance use and substance dependence. *Journal of Addictive Diseases, 23*(4), 55-71. doi: 10.1300/J069v23n04_05.

Goldsmith, R. E. (1986).Dimensionality of the Rosenberg Self-EsteemScale. *Journal of Social Behavior and Personality, 1,* 253-264

Greenberger, E., Chen, C., Dmitrieva, J., & Farruggia, S. P. (2003). Item-wording and the dimensionality of the Rosenberg Self-Esteem Scale: do they matter? *Personality and Individual Differences, 35*, 1241-1254.

González-Pienda, J., Núñez, J.C., González-Pumariega, S., & García, M.S. (1997). Autoconcepto, *autoestimay aprendizajeescolar. Psicothema*, *9,* 271-289.

Hamm, R. (2009). Negative will, self-image, and personality dysfunction. [Case Reports; ; Review]. *Psychoanal Rev, 96*(1), 55-82.

Han, X., Jiang, B., Yang, J., & Wang, Y. (2005). The problems and suggestions in using self-esteem scale. *Chinese Journal of Behavioral and Medicine Science (Chinese version), 14*(8), 763.

Heine, S. J., Lehman, D. R., Markus, H. R., & Kitayama, S. (1999). Is There a Universal Need for Positive Self-Regard? *Psychological Review October, 106*(4), 766-794.

Hu, L.-T., & Bentler, P. M. (1999). Cut off criteria for fit indexes in covariance structure analysis: Conventional criteria versus new alternatives. *Structural Equation Modeling, 6*, 1-55.

Ji, Y.F., & Yu, X. (1993). The assessment of self-esteem. In X.D. Wang (Ed.), Rating scales for mental health. *Beijing: Journal of Chinese Mental Health.* (pp. 251–252)

Jodoin, M.G., & Gierl, M.J. (2001). Evaluating type I error and power rates using an effect size measure with the logistic regression procedure for DIF detection. *Applied Measurement in Education, 14,* 329-349.

Johnson, W., Spinath, F., Krueger, R. T., Angleitner, A., & Riemann, R. (2008). Personality in Germany and Minnesota: An IRT-Based Comparison of MPQ Self-Reports. *Journal of Personality 76*(3), 667-707. Jöreskog, K. G. (1971). Simultaneous factor analysis in several populations. *Psychometrika, 36,* 409–426

Jöreskog, K. G. (1993). *Testing structural equation models*. Newbury Park, CA: Sage

Li, H. C. W., Chan, S. L. P., Chung, O. K. J., & Chui, M. L. M. (2010). Relationships among Mental Health, Self-esteem and Physical Health in Chinese Adolescents An exploratory study. [Article]. *Journal of Health Psychology, 15*(1), 96-106. doi: 10.1177/1359105309342601

MacCallum, R. C. , Roznowski, M., & Necowitz, L. B. (1992). Model modifications in covariance structure analysis: The problem of capitalization on chance. *Psychological Bulletin*, *111*, 490–504.

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

Marsh, H.W. (1994). Confirmatory factor analysis models of factorial invariance: a multifaceted approach. *Structural Equation Model, 1(1)*, pp. 5–34

Marsh, H. W., Hau, K. T., & Wen, Z. (2004). In search of golden rules: Comment on hypothesis-testing approaches to setting cutoff values for fit indexes and dangers in overgeneralizing Hu and Bentler's findings. *Structural Equation Modeling, 11*, 320-341.

Martín-Albo, J., Núñez, J. L., Navarro, J. G., & Grijalvo, F. (2007). The Rosenberg Self-Esteem Scale: Translation and Validation in University Students. *The Spanish Journal of Psychology, 10*(2), 458-467

Mavondo, F. T. & Farrell, M.A. (2000). Measuring Market Orientation: Are There Differences Between Business Marketers and Consumer Marketers? *Australian Journal of Management, 25 (2)*, 223-244

Meadows, K., Bentzen, N. & Touw-Otten, F. (1997). Cross-cultural issues: an outline of the important principles in establishing cross-cultural validity in health outcome assessment. In: Hutchinson A, Bentzen N, Konig-Zahn C, editors. Cross cultural health outcome assessment: a user's guide. *UK: European Research Group on Health Outcomes.* 34– 40.

Meredith, W. (1993). Measurement invariance, factor analysis, and factorial invariance. *Psychometrika, 58*, 525-543.

Michaels, M. L., Barr, A., Roosa, W., & Knight, G. P. (2007). Assessing Measurement Equivalence in a Multiethnic Sample of Youth. *Journal of Early Adolescence, 27*, 269-297.

Milfont, T. L., & Fischer, R. (2010). Testing measurement invariance across groups: applications in cross-cultural research. [i]. *International Journal of psychological research, 3*(1), 2011-2079.

Millsap, R.E., Kwok, O.M. (2004). Evaluating the impact of partial factorial invariance on selection in two populations. *Psychological Methods.*; *9,*93–115

Orth, U., Robins, R. W., & Meier, L. L. (2009). Disentangling the Effects of Low Self-Esteem and Stressful Events on Depression: Findings From Three Longitudinal Studies. [Miscellaneous Article]. *Journal of Personality & Social Psychology August, 97*(2), 307-321.

Owens, T. J. (1993). Accentuate the positive and the negative: Rethinking the use of self-esteem, self-deprecation, *Social Psychology Quarterly, 56*, 288-299

Paterson, G., Power, K., Yellowlees, A., Park, K., & Taylor, L. (2007). The relationship between two-dimensional self-esteem and problem solving style in an anorexic inpatient sample. *European Eating Disorders Review, 15*(1), 70-77. doi: 10.1002/erv.708

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

Reise, S. P., Widaman, K. F. & Pugh R H. (1993). Confirmatory factor analysis and item response theory: two approaches for exploring measurement invariance. *Psychological Bulletin, 114,* 552~566

Rosenberg M. (Ed.). (1965). *Society and adolescent self-image.*: New Jersey Princeton University Press.

Roth, M., Decker, O., Herzberg, Y., & Brähler, E. (2008). Dimensionality and Norms of the Rosenberg Self-esteemScale in a German General Population Sample. *European Journal of Psychological Assessment 24*(3), 190-197.

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 & Social Psychology October, 89*(4), 623-642.

Sedikides, C., Gaertner, L., & Vevea, J. L. (2005). Pancultural Self-Enhancement Reloaded: A Meta-Analytic Reply to Heine (2005). [Miscellaneous]. *Journal of Personality & Social Psychology October, 89*(4), 539-551.

Shavelson, J., Hubner, J.J., & Stanton, G.C. (1976). Self-concept: Validation of construct interpretations. *Review of Educational Research, 46,* 407-442.

Shen, Z., & Cai, T. (2008). Disposa l to the 8th Item of Rosenberg Self-Esteem Scale (Chinese version). *Chinese mental health magazine, 12*(9), 661-663.

Steenkamp, J-BEM & Baumgartner, H. (1998). Assessing measurement invariance in cross-national consumer research. *J Consum Res, 25,* 78 – 90.

Stein, J. A., Lee, J. W., & Jones, P. S. (2006). Assessing Cross-Cultural Differences Throug Use of Multiple-Group Invariance Analyses. *Journal of personality assessment, 87*(3), 249 - 258.

Steven, E.,& Gregorich (2006). Do Self-Report Instruments Allow Meaningful Comparisons cross Diverse Population Groups? Testing Measurement Invariance Using the Confirmatory Factor Analysis Framework. Medical Care, 44(3), 78–94.

Suh, E., Diener, E., Oishi, S., & Triandis, H. (1998). The shifting of life satisfaction judgments across cultures: Emotions versus norms. *Journal of Personality and Social Psychology, 74,* 482-493.

Swaminathan, H., & Rogers, H. J. (1990). Detecting differential item functioning using logistic regression procedures. *Journal of Educational Measurement, 27,* 361-370.

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,* 4-70.

Velde, S.V., Bracke, P., Levecque, K. (2008). The psychometric properties of the CES-D 8 depression inventory and the estimation of cross-national differences in the true prevalence of depression. *Retrieved from:* http://www.csdiworkshop.org/pdf/3mc2008_proceedings/session_34/Bracke.pdf

Tsai, J. L., Ying, Y. W., & Lee, P. A. (2001). Cultural predictors of self-esteem: A study of Chinese American Female and Male Young Adults. *Culturual Diversity and Ethnic Minority Psychology, 7(3),* 284-297.

Wang, Y., & Ollendick, T. (2001). A Cross-Cultural and Developmental Analysis of Self-Esteem in Chinese and Western Children. *Clinical Child and Family Psychology Review, 4*(3), 254-271.

Zumbo, B. D. (1999). A Handbook on the Theory and Methods of Differential Item *Functioning (DIF): Logistic Regression Modeling as a Unitary Framework for Binary and Likert-type (Ordinal) Item Scores.* Ottawa ON: Directorate of Human Resources Research and Evaluation, Department of National Defense

Appendix 1

我了解了以上信息，并且同意参加研究。我已经年满 18 周岁

_____是　　_____ 否　日期:_____

填表注意事项：这个量表是用来了解您是怎样看待自己的。请仔细阅读下面的句子，在最符合您情况的数字上划“✓”（**1-非常符合；2-符合；3-不符合；4-非常不符合**）。请注意，这里要回答的是您实际上认为您自己怎样，而不是回答您认为您应该怎样。

	非常符合	符合	不符合	非常不符合
1. 总的来说，我对自己是满意的。	1	2	3	4
2. 我有时认为自己一无是处。	1	2	3	4
3. 我感到我有许多好的品质。	1	2	3	4
4. 我能像大多数人一样把事情做好。	1	2	3	4
5. 我感到自己值得自豪的地方不多。	1	2	3	4
6. 我有时确实感到自己毫无用处。	1	2	3	4
7. 我感到我是一个有价值的人，至少与其他人在同一水平上。	1	2	3	4
8. 我希望我能为自己赢得更多尊重。	1	2	3	4
9. 归根结底，我倾向于认为自己是一个失败者。	1	2	3	4
10. 我对自己持肯定态度。	1	2	3	4