RUNNING HEAD: DIF

The Concept of Life Satisfaction Across Cultures: An IRT Analysis

Shigehiro Oishi
University of Virgnia

January 19, 2005
Journal of Research in Personality (in press)

Key words: life satisfaction, culture, measurement, well-being judgments
Correspondence:
Shigehiro Oishi
Department of Psychology
University of Virginia
P.O. Box 400400
Charlottesville, VA22904-4400
Phone (434) 243-8989
E-mail: soishi@virginia.edu

Abstract

The present study examined measurement equivalence of the Satisfaction with Life Scale (SWLS) between American and Chinese samples using multigroup Structural Equation Modeling (SEM), Multiple indicator multiple cause model (MIMIC), and Item Response Theory (IRT). Whereas SEM and MIMIC identified only one biased item across cultures, the IRT analysis revealed that four of the five items had differential item functioning (DIF). According to IRT, Chinese whose latent life satisfaction scores were quite high did not endorse items such as "So far I have gotten the important things I want in life" and "If I could live my life over, I would change almost nothing." The IRT analysis also showed that even when the unbiased items were weighted more heavily than the biased items, the latent mean life satisfaction score of Chinese was substantially lower than that of Americans. The differences among SEM, MIMIC, and IRT are discussed.

The Concept of Life Satisfaction Across Cultures: An IRT Analysis

Kitayama and Markus (2000) presented a theoretical analysis of cultural differences in well-being, and argued that (a) well-being comes from cultural participation, and (b) to the extent that cultural participation requires different forms across cultures, well-being feels different and means something different across cultures. For instance, the Item Response Theory (IRT) analysis of the positive affect (PA) subscale of the Positive and Negative Affect Schedule (PANAS: Watson, Clark, & Tellegen, 1988) showed that "pride" was not endorsed by Chinese who endorsed other "positive" emotions, whereas it was endorsed by Americans who endorsed other "positive" emotions (Oishi, in press). This measurement discrepancy indicates that "pride" is not conceived as "positive" among Chinese and reveals the conceptual difference of "positive" emotions between Chinese and Americans (see Huang, Church, & Katigbak, 1997 on anxiety between Philippines and Americans). A main implication of Kitayama and Markus' theoretical analysis for culture and personality research is that it is crucial to examine not only mean-level differences in a construct (e.g., self-esteem) and the nomological net of this construct across cultures, but also the deeper structure of the construct because the traditional questions of mean-level difference across cultures presuppose conceptual equivalence.

The Satisfaction with Life Scale (SWLS; Diener, Emmons, Larsen, & Griffin, 1985) has been one of the most widely used scales for the measurement of global life satisfaction. Life satisfaction is one of the central constructs of well-being (Diener, 1984) and has been of great interest to both cultural and personality psychologists (Diener, Oishi, & Lucas, 2003; Diener, Suh, Lucas, & Smith, 1999 for review). Its psychometric properties have been well-established in the United States (Pavot & Diener, 1993). In contrast, the psychometric properties of the SWLS in non-American samples have not been extensively examined (see Vittersø, Røysamb, & Diener, 2002, however, for an initial effort in this direction). Therefore, although previous research found large international differences in the mean levels of life satisfaction (e.g., Diener, Suh, Smith, & Shao, 1995), it is unclear exactly how these mean

differences can be interpreted because of the lack of information concerning measurement equivalence. The present study examines measurement equivalence of the SWLS between Chinese and American college student samples, using the structural equation modeling (SEM), multiple indicator multiple cause (MIMIC) modeling, and Differential Item Functioning (DIF) analysis.

DIF Analysis

To examine measurement equivalence of the SWLS among Chinese and American college students, I employed the IRT analysis with a model-testing approach (Thissen, Stenberg, & Gerrard, 1986) using the Multilog 7.03 program. IRT is different from classical test theory (CTT) in several important ways (see Embretson & Reise, 2000; Hambleton & Swaminathan, 1985 for details). The most significant difference between CTT and IRT in the present context is concerned with the standard error of measurement. Whereas the standard error of measurement is assumed to apply to the whole sample in CTT, the standard error of measurement in IRT varies depending on the latent trait score (typically, there is less reliability for those with extreme latent scores). In other words, whereas the source of errors in CTT is either occasion (in the case of test-retest reliability) or item sampling (in the case of internal consistency), additional sources of error can be considered in IRT (as in Generalizability theory by Cronbach, Gleser, Nanda, & Rajaratnam, 1972; Shavelson, Webb, & Rowley, 1989), such as a person's latent score and person-by-item interaction. Traditional reliability indices such as Cronbach's alpha and test-retest reliability coefficient do not provide information about person-by-item interaction, namely, whether some items measured some individuals better than others. In IRT, this interaction is considered. In addition, classical item parameters (e.g., item-total correlation) are sample-specific, whereas IRT parameters are not sample dependent. The score computed in IRT, therefore, can be readily compared across different test forms. By and large, IRT parameters have a greater degree of generalizability than classical item parameters.

Second, in CTT, if two individuals answered the same number of items "correctly" (or gave the same number of "yes" responses), these two individuals would have the same total score. In contrast, in IRT, even if two individuals answered the same number of items "correctly" (or gave the same number of "yes" responses), the person who correctly answered more difficult items (or those who said "yes" to the items less frequently endorsed) would receive a higher total score than the other who correctly answered less difficult items in IRT. Virtually all previous cross-cultural research in well-being has neglected item difficulty parameters. Thus, it is of great interest whether the scoring method of IRT, which takes into account the item difficulty parameters, would reveal a different result than the conventional scoring method.

There are several IRT models. For personality research, the 2 parameter logistic model (2PL model) is perhaps most relevant as it estimates an item discrimination parameter and item difficulty parameters (see Embretson & Reise, 2000 for review; see Gray-Little, Williams, & Hancock, 1997; Reise & Waller, 1990 for examples). An item discrimination parameter ("a" parameter) indicates how well a given item captures the latent trait that it is supposed to measure. This is conceptually equivalent to item-total score correlation in CTT and item-factor correlation in factor analysis. An item difficulty parameter ("b" parameter) indicates the likelihood of "passing" the item (or saying "yes" to the item), given different levels of the latent score. If a person with a latent score of -.75 (.75 sd below mean) had a 50% probability of passing item X, then the b for item X is -.75. Similarly, if a person with a latent score of .50 (.50 sd above the mean) had a 50% probability of passing item Y, then the b for item Y is .50. The 2PL model is typically expressed as follows: $Pix = \exp(\alpha(\theta x-\beta i))/[1+\exp(\alpha(\theta x-\beta i)]$, where Pix indicates the probability of person X passing item I, θx indicates person X's latent score, α indicates the item discrimination, and βi indicates item I's difficulty.

Suppose item 1 had the item discrimination, $\alpha = 1$, whereas item 2 had the item discrimination, $\alpha = 2$. Let's further suppose that both items had the item difficulty, $\beta = .50$. The person with a latent trait score of 1.00 has a 62.24% probability of passing item 1, whereas the same person has a

73.11% probability of passing item 2. In other words, even when θx and βi are the same, Pix can be different, depending on an item discrimination parameter. Most important, when estimating a latent trait score, the 2PL model takes into account item discrimination as well as item difficulty, and, therefore, the better items (those with higher item discrimination) have greater weight in estimating the latent trait score than do other items (this is an advantage of the 2PL model over the 1PL and CTT).

So far, all the examples assumed that items were dichotomous (i.e., pass or fail; yes or no). In many scales used in psychological research, however, items are multiple ordered-response categories (e.g., strongly disagree, disagree, neither agree nor disagree, agree, strongly agree). Samejima (1969) developed the graded-response model of IRT that is appropriate for polytomous items. Whereas there was one item difficulty parameter in the dichotomous model, there are k-1 item difficulty parameters, where k is the number of response categories (e.g., there are six b parameters for a 7-point scale). In the case of the 3-point scale, the "b1" parameter indicates the level of latent score that has the equal probability of endorsing 1 and 2, whereas the "b2" parameter indicates the level of latent score in which the probability of endorsing 2 and 3 is equal. If b1 is -.33 and b2 is .67, then people whose latent trait score is below -.33 are most likely to endorse the response category 1, whereas people whose latent score is between -.33 and .67 have the greatest probability of endorsing the response category 2, and those higher than .67 are most likely to endorse the response category 3.

DIF analysis under the IRT framework examines whether item discrimination and item difficulty parameters are comparable between two groups (see Holland & Wainer, 1993; Milsap & Everson, 1993; Thissen et al., 1986 for review). In essence, this analysis provides an index for whether the conditional probability of getting an item "correct" differs across two groups. A large DIF indicates group differences in the conditional probability; namely, an item has DIF if the probability of endorsing an item is different across groups, given the same level of latent score. The IRT analysis also provides the estimate of latent mean score for a scale for each participant. As articulated by Reise, Widaman, and Rugh (1993), IRT generates the estimate of latent mean scores even when there are items exhibiting DIF. Specifically, Reise et al. state "Our basic requirement in IRT is that at least one item be invariant across groups. The invariant item can then be used as an anchor to estimate values for individuals within both groups concurrently on a common scale"(p. 561).

A general approach for identifying DIF items is very similar to model comparisons often used in SEM. As in SEM, measurement invariance will be tested on an item-by-item basis. A model that constrains target item's parameters to be equivalent in two groups will be compared with the baseline model, in which all the parameters are allowed to differ between groups, using G^2 value (G^2 is the difference in -2 times the log of the likelihood function in two models; -2 items the log of the likelihood function is provided by Multilog for each model). If constraining the item parameters across samples does not result in a significant increase in G^2, then the target item is deemed non-DIF. If constraining the item parameters results in a significant increase in G^2, then this item is deemed DIF.

Despite the vigor of IRT, to our knowledge, it has not yet been used in cross-cultural research on well-being. Traditionally, multigroup SEM has been used to establish measurement equivalence (e.g., Schimmack, Radhakrishnan, Oishi, Dzokoto, & Ahadi, 2002). Although this is a great advancement over traditional models that do not control for measurement errors, multigroup SEM tests the equivalence of item discrimination parameters between groups only, and does not test the equivalence of item difficulty parameters between groups. Another SEM-based technique is multiple indicator multiple cause (MIMIC) model (Jöreskog & Goldberger, 1975). In the MIMIC modeling, two samples are combined to one data set, and a group variable is included as an exogenous variable. As can be seen in Figure 1, the baseline model that does not allow any association between the group variable and errors for items will be compared with the model that allows for association between group and error terms for

five items. Because the factor loadings are assumed to be equivalent across groups in the MIMIC, the association between error terms and the group variable indicates the lack of measurement equivalence between two samples. As in SEM, MIMIC focuses on item discrimination, and does not take into account potential differences in item difficulty. The 2PL model of IRT provides a more rigorous test of measurement equivalence across groups because it tests the equivalence of both item discrimination and item difficulty parameters.

In short, the main goals of the present investigation were to examine (a) whether the items in the SWLS would show DIF among Americans and Chinese, (b) if there are DIF items, whether they are conceptually different from non-DIF items, and (c) whether item biases would affect the magnitude of mean differences between Americans and Chinese.

Method

Participants. In China, 556 students at Zhejiang Institute of Technology participated in this study. They completed a questionnaire package in the classroom. In the United States, 442 students at the University of Illinois enrolled in an introductory psychology course participated in this study. They completed the same package in groups of 15 to 20 in a large lecture room.

Measures. The Satisfaction with Life Scale (SWLS: Diener et al., 1985) was used to assess global life satisfaction. The SWLS consists of five items ("In most ways my life is close to my ideal," "The conditions of my life are excellent," " I am satisfied with my life," " So far I have gotten the important things I want in my life," and " If I could live my life over, I would change almost nothing"). The participants responded to these items using a 7-point scale, ranging from 1 (*strongly disagree*) to 7 (*strongly agree*). The English version of the SWLS was translated into Chinese by a bilingual research assistant. Another bilingual research assistant back-translated the Chinese version of the SWLS. The back-translation version of the SWLS was compared with the original SWLS by Ed Diener. The back-translation confirmed the accuracy of the translation.

Results

Traditional Analyses

First, I conducted traditional tests of psychometric properties of the SWLS in China and the US. Cronbach's alpha was .71 for China, and .88 for the US. In both samples items 4 and 5 had the lowest correlations with other items. To test structural equivalence of the SWLS, I next conducted confirmatory factor analysis. I compared the baseline model without any constraint of factor loadings with a series of constrained models. In the baseline model, factor loadings can be different across the two samples, whereas in the constrained model, factor loadings were set to be exactly the same value across two samples for one item at a time. The baseline model yielded the chi-square value of 50.26 ($df = 10$, $p < .01$) with the fit indices of *GFI* = .98, *AGFI* = .94, and *RMSEA* = .06 (see Table 1 for factor loadings). Despite the significant chi-square, the fit indices suggest that the one factor model fits the data fairly well in both samples. Next, I compared the unconstrained model with a series of constrained models that constrained each item one at a time. Table 2 shows that only the model that constrained item 5 was significantly different from the unconstrained model. All other models did not differ from the baseline model. Thus, according to SEM, only item 5 showed an item bias across the two cultural groups.

Next, I conducted a series of MIMIC analyses. The baseline model, in which no association between the group variable (China was coded as 0, and U.S was coded as 1) and the error terms was allowed, yielded the chi-square value of 236.75 ($df = 9$, $p < .01$), *GFI* = .921, *AGFI* = .821, and *RMSEA* = .161. This model is conceptually equivalent to constraining all the factor loadings between two samples in the multigroup SEM analysis, and the fit indices suggest room for improvement. Next, we allowed an error term for each item and the group variable to be associated one at a time. Allowing the error term for item 4 and the group variable to be associated resulted in a dramatic improvement in fit, χ^2 ($df = 8$) = 48.24, $p < .01$, $\Delta \chi^2$ = 188.51, $p < .01$, *GFI* = .983, *AGFI* = .956, and *RMSEA* = .072 in this model. Although chi-square is still significant, the fit indices are acceptable. Allowing any additional error term to be associated with the group variable did not improve the fit indices nor result in a

significant change in chi-square value. Thus, the MIMIC analyses indicate that item 4's relation to the latent trait life satisfaction was systematically different between Chinese and American samples. The final MIMIC model (with item 4's error and the group variable associated) yielded the standardized regression coefficient from the group variable to latent trait life satisfaction of .458, p, < .01, d = 1.03, or a large mean difference between Chinese and American samples.

The DIF Analysis

The reliability, multigroup SEM, and MIMIC analyses provided information about the dimensional and structural equivalence across the two samples. However, these analyses do not provide critical information concerning item difficulty. It is unclear from the previous analyses, for instance, whether the probability of endorsing item 1 "In many ways my life is close to my ideal" would be different between Chinese and Americans with the same latent life satisfaction score. To address this issue, it is important to go beyond the traditional approach and use IRT. DIF analysis is the most relevant criterion in addressing this issue. Following Reise et al. (1993), I compared the baseline model, which allowed all parameters (the item discrimination parameter and item difficulties parameters) to differ between the two groups with a series of constrained models, in which item parameters were set to be equal between the groups. Table 3 shows that the baseline model revealed that item difficulty parameter estimates for the first three items were fairly comparable, whereas these estimates were markedly different between the two groups for Items 4 and 5. To illustrate cultural similarities and differences in an Item Characteristic Curve (ICC), Figure 2 shows two ICC's. The first panel shows the ICC of item 2 for Chinese participants. The X axis indicates latent life satisfaction in a standardized unit, and the Y axis indicates the probability of choosing a particular response category. If an item is functioning as it is supposed to be, then the probability of choosing a response category should go up from 1 to 2, 2 to 3, and so on until 7, as the latent life satisfaction score goes up from 3 SDs below the mean to 3 SDs above the mean. Figure 2 indicates that item 2 meets this criterion in both groups. For instance, a Chinese whose latent life satisfaction is 3 standard deviations below the mean has about 62% probability of choosing category 1 "strongly disagree" for item 2, about 36% probability of choosing category 2 "disagree" for item 2, almost zero % probability of choosing categories 4 or higher. A Chinese whose latent life satisfaction is 2 standard deviations below the mean is most likely to choose category 2, followed by categories 1 and 3. Similarly, a Chinese whose latent life satisfaction is 2 SDs above the mean is most likely to choose category 6 "agree," followed by categories 7 "strongly agree" and 5 "slightly agree." In contrast, the ICC of item 5 for Chinese (Figure 3) clearly indicates that this item is not functioning as it is supposed to be. For instance, a Chinese whose latent life satisfaction score is 1 SD above the mean is most likely to choose category 2 "disagree" for item 5, followed by category 6 "agree." Also, a Chinese with latent life satisfaction score of 2 SDs above the mean is mostly likely to choose category 6, followed by the equal probability of categories 2 "disagree" and 7 "strongly agree"! The ICC of item 4 was very similar to item 5. Thus, it is clear that the majority of Chinese whose latent life satisfaction score was quite high (e.g., 2 SDs above the mean) did not endorse these two items. In contrast, the ICC of items 2 (Figure 2) and 5 (Figure 3) for the American participants follows the expected patterns with a shift of the dominant response as the latent life satisfaction.

Although ICC provides valuable information visually, whether each item is functioning differently between the two groups has to be formally tested using the model comparison method. Table 2 shows the G^2 for each constrained model. For instance, when item 1's item parameters were set equal in two groups, G^2 index was 37.7 with 7 degrees of freedom. Because the G^2 is distributed as chi-square, significance can be tested against the chi-square distribution. The last column of Table 2 indicates the p-value for each constrained model. Essentially, the larger the G^2, the larger the DIF. As expected, items 4 and 5 had very large DIF. According to the G^2 indices, item 2 was the only item that did not have significant DIF.

Because item 2 was the only item without DIF, the latent score for each group was computed based on this model. The latent life satisfaction score was .71 lower among the Chinese participants than among the American participants. Because the mean of American participants was set to zero with a variance of 1, this can be interpreted as Cohen's d (d = .71), or a large effect size. The t-test on the simple sum of the five items yielded t (979) = 18.35, $p < .01$, d = 1.18. The mean difference is therefore smaller based on the latent score than the observed score, suggesting that the sum of the five observed scores somewhat exaggerated the cultural difference in life satisfaction. T-tests on each item resulted in large differences between Chinese and Americans (d = .70, .67, .83, 1.55, .70 for items 1 to 5, respectively). Interestingly, the degree of DIF and the size of the mean difference seem to be independent, as the item with the largest DIF (item 5) was not much different in the size of the mean difference from the item with the smallest DIF (item 2). The latent life satisfaction score was computed with items that had more information (i.e., items with higher discrimination) weighted more heavily than others, and therefore was a more accurate reflection of the true score than the simple sum of the five items. Although cultural differences were smaller with the IRT scoring, IRT analyses still revealed substantial differences in mean life satisfaction between Chinese and Americans.

Discussion

The present paper examined measurement equivalence of the SWLS between Chinese and American college students. The reliability analysis revealed that the SWLS among the American college students had a higher degree of internal consistency than among the Chinese college students. The multigroup SEM and the MIMIC analyses showed that items 5 and 4, respectively, in the Chinese sample were not as good indicators of life satisfaction as they were in the American sample. The DIF analysis revealed that four of the five items had significant DIFs. Replicating Reise et al. (1993), I found more item biases when I used the 2PL model of IRT than multigroup SEM and MIMIC. The differences between two SEM-based analyses and DIF analyses demonstrated that DIF present a more conservative test of measurement equivalence between groups than multigroup SEM and MIMIC. This indicates the importance of taking into account both item discrimination and item difficulty parameters.

Items 4 and 5 had particularly large DIFs. It should be also noted these two items were identified as different across the two samples by multigroup SEM and MIMIC analyses as well. Neither SEM, nor MIMIC, nor IRT explain why these two items had largest cross-cultural differences. Therefore, the interpretation of these findings must rely on the post-hoc analysis of item contents and/or insights from cultural psychology and anthropology. In terms of item content, these two items assess one's satisfaction with past accomplishments ("So far I have gotten the important things I want in life" and "If I could live my life over, I would change almost nothing"), whereas the first three items focus on external living conditions or the present level of satisfaction ("In most ways my life is close to my ideal," "The conditions of my life are excellent," "I am satisfied with my life"). A number of cultural and cross-cultural psychology studies have found that East Asians do not evaluate their personal accomplishments or task performance as positively as do Americans (e.g., Markus & Kitayama, 1991; Heine, Lehman, Markus, & Kitayama, 1999, for review). Markus and Kitayama (1991) laid out two general motivational syndromes; self-enhancement in an independent society, and self-criticism in an interdependent society. In the present context, the positive evaluation of their own past accomplishments seems more self-enhancing than the positive evaluation of external living conditions. One of the reasons why generally satisfied Chinese participants did not endorse items 4 and 5 might be, then, that the endorsement of these two items seems too self-enhancing, and not socially desirable.

Another explanation is that Chinese concept of life satisfaction is based primarily on external conditions and the current status rather than past accomplishments. Even if Chinese view the conditions of their lives as excellent and are satisfied with their lives in general, they may still feel that they have not accomplished important goals in life and that

they would change something if they lived their lives again. In a self-critical society, where self-improvement is highly valued, then, past accomplishments might not guarantee satisfaction because the old standard that they just achieved is constantly upgraded to a newer, higher standard. Unfortunately, the present study does not present any information concerning the relative utility of the two interpretations. Future research should test which account is more plausible. In addition to these two accounts, however, translation equivalencies should be examined as one of the possible causes for DIF. For instance, item 5, which uses counterfactual reasoning, might be particularly difficult to translate in a culture where counterfactual thinking is not employed as often as in Western cultures (e.g., Bloom, 1981). It is important to examine measurement artifacts, because they will allow researchers to make theoretical interpretations of the obtained results.

Before closing, the limitations of IRT should be noted. As suggested by the self-presentation account, IRT does not address the issue of response style nor social desirability. For instance, a key reason why IRT identified more DIR than other methods might be that item difficulty parameters are largely affected by response tendencies such as a mid-point use or extreme scale use (see Chen, Lee, & Stevenson, 1995 for cultural difference sin response tendencies). Similarly, desirability of the item might be different across cultures. As discussed above, items 4 and 5 seem particularly socially undesirable to endorse in a culture where modesty is highly valued. Thus, researchers should keep in mind that IRT analyses are affected by response tendencies and social desirability. Equally important, the identification of DIF items using the IRT should not be strictly determined by G^2 values alone. These indices should be used as a guideline rather than strict rules. Finally, the latent means in the IRT were derived based on only one non-DIF item. Although this is acceptable, this is far from ideal. In order to obtain truly unbiased latent mean differences, a greater number of non-DIF items should exist.

Conclusion

The DIF analyses provided a different perspective from the traditional approaches to the measurement issue in culture and well-being research. Equally important, IRT analyses revealed a substantive level of mean differences between Chinese and Americans, even when the biased items were weighted less in scoring. Thus, previously found mean differences between Americans and Chinese (e.g., Diener et al., 1995) might not be due simply to item biases. Finally, IRT analysis provided invaluable information concerning the concept of life satisfaction. The present research illuminates the importance and benefit of employing DIF analyses in other constructs (e.g., self-esteem, depression). I hope that DIF analysis and other IRT models will be utilized in the future in various research topics in culture and personality.

References

Bloom, A. (1981). *The linguistic shaping of thought*. Hillsdale, NJ: Earbaum.

Chen, C., Lee, S-Y., & Stevenson, H.W. (1995). Response style and cross-cultural comparisons of rating scales among East Asian and North American students. *Psychological Science*, *6*, 170-175.

Cronbach, L. J., Gleser, G. C., Nanda, H., & Rajaratnam, N. (1972). *The dependability of behavioral measurements: Theory of generalizability of scores and profiles*. New York: Wiley.

Diener, E. (1984). Subjective well-being. *Psychological Bulletin*, 95, 542-575.

Diener, E., Emmons, R. A., Larsen, R. J., & Griffin, S. (1985). The Satisfaction With Life Scale. *Journal of Personality Assessment*, *49*, 71-75.

Diener, E., Oishi, S., & Lucas, R. E. (2003). Personality, culture, and subjective well-being: Emotional and cognitive evaluations of life. *Annual Review of Psychology*, *54*, 403-425.

Diener E, Suh EM, Lucas RE, Smith HE. 1999. Subjective well-being: Three decades of progress. *Psychological Bulletin, 125*, 276-302

Diener, E., Suh, E. M., Smith, H., & Shao, L. (1995). National differences in reported subjective well-being: Why

do they occur? *Social Indicators Research*, *34*, 7-32.

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.

Hambleton, R. K., & Swaminathan, H. (1985). *Item response theory: Principles and applications*. Boston: Kluwer-Nijhoff.

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. (1993). *Differential item functioning*. Hillsdale, NJ: Lawrence Erlbaum.

Huang, C. D., Church, A. T., & Katigbak, M.S. (1997). Identifying cultural differences in items and traits: Differential item functioning in NEO personality inventory. *Journal of Cross-Cultural Psychology*, *28*, 192-218.

Jöreskog, K. G., & Goldberger, A. S. (1975). An estimation of multiple indicators and multiple causes of a single latent variable. *Journal of the American Statistical Association*, *58*, 1048-1053.

Kitayama, S., & Markus, H. R. (2000). The pursuit of happiness and the realization of sympathy: Cultural patterns of self, social relations, and well-being. In E. Diener, & E. M. Suh (Eds.), *Cultural and subjective well-being* (pp. 113-161).

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

Milsap. R. E., & Everson, H. T. (1993). Methodology review; Statistical approach for assessing measurement bias. *Applied Psychological Measurement*, *17*, 297-334.

Oishi, S. (in press). The Application of structural equation modeling and item response theory to cross-cultural positive psychology research. In A. D. Ong & M. van Dulmen (Eds), *Handbook of methods in positive psychology*. NY: Oxford University Press.

Pavot, W., & Diener, E. (1993). Review of the Satisfaction with Life Scale. *Psychological Assessment, 5,* 164-171.

Reise, S. P., & Waller, N. G. (1990). Fitting the two-parameter model to personality data: The parameterization of the Multidimensional Personality Questionnaire. *Applied Psychological Measurement*, *14*, 45-58.

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

Samejima, F. (1969). Estimation of latent ability using a response pattern of graded scores. *Psychometric Monograph*. No. *17*.

Schimmack, U., Radhakrishnan, P., Oishi, S., Dzokoto, V., & Ahadi, S. (2002). Culture, personality, and subjective well-being: Integrating process models of life satisfaction. *Journal of Personality and Social Psychology, 82*, 582-593.

Shavelson, R. J., Webb, N. M., & Rowley, G. L. (1989). Generalizability theory. *American Psychologist*, *44*, 922-932.

Thissen, D., Stenberg, L., & Gerrard, M. (1986). Beyond group-mean differences: The concept of item bias. *Psychological Bulletin*, *99*, 118-128.

Vittersø, J., Røysamb, E., & Diener, E. (2002). The concept of life satisfaction across cultures: Exploring its diverse meaning and relation to economic wealth. In E Gullone & R. Cummins (Eds.), *The Universality of subjective wellbeing indicators*, (pp. 81-103). Dordrecht: Kluwer Academic Publishers.

Watson, D., Clark, L. A., & Tellegen, A. (1988). Development and validation of brief measure of positive and negative affect: The PANAS scales. *Journal of Personality and Social Psychology*, *54*, 1063-1070.

Table 1
Factor Loadings in the Baseline (unconstrained) Model

China Item	Unstandardized	Standardized
1	1.00	.62
2	1.05	.66
3	1.17	.72
4	.69	.49
5	.68	.38
US Item		
1	1.00	.84
2	.97	.80
3	.96	.85
4	.84	.67
5	1.02	.71

Table 2
Model Comparisons by Confirmatory Factor Analysis (CFA) and DIF analysis

	CFA	DIF	
Item Constrained	$\Delta\chi^2$	-2 LK	G^2
1	.53	3098.7	37.7**
2	.53	3075.7	14.7
3	3.72	3095.5	34.5**
4	2.27	3155.7	94.7**
5	8.32**	3156.8	95.8**

Note. **$p < .01$ df = 1 for CFA analyses, df = 7 for DIF analyses.

Table 3
The Baseline Model of IRT analysis

Item	Parameter	China	US
1	a	1.53	2.89
	b1	-2.68	-2.23
	b2	-0.87	-1.25
	b3	-0.35	-0.82
	b4	-0.10	-0.49
	b5	1.08	0.29
	b6	2.89	1.82
2	a	1.89	2.62
	b1	-2.68	-2.46
	b2	-1.01	-1.32
	b3	-0.35	-0.82
	b4	0.08	-0.41
	b5	0.86	0.31
	b6	2.46	1.54
3	a	2.18	3.13
	b1	-2.55	-2.36
	b2	-1.09	-1.55
	b3	-0.36	-1.07
	b4	0.00	-0.76
	b5	0.61	-0.06
	b6	1.96	1.44
4	a	1.04	1.79
	b1	-1.53	-2.50
	b2	0.62	-1.56
	b3	1.26	-0.99
	b4	1.91	-0.44
	b5	2.64	0.25
	b 6	4.43	1.63
5	a	0.75	2.08
	b1	-2.02	-1.71
	b2	0.18	-0.84
	b3	0.81	-0.20
	b4	1.55	0.06
	b5	2.35	0.65
	b6	4.16	1.76

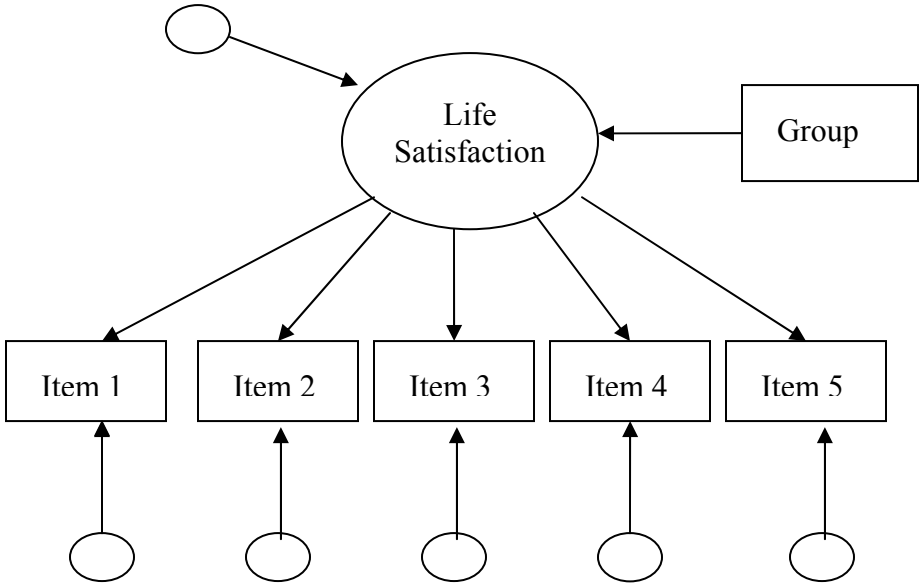
Note. “a” denotes an item discrimination parameter. “b” denotes an item difficulty parameter for each threshold.

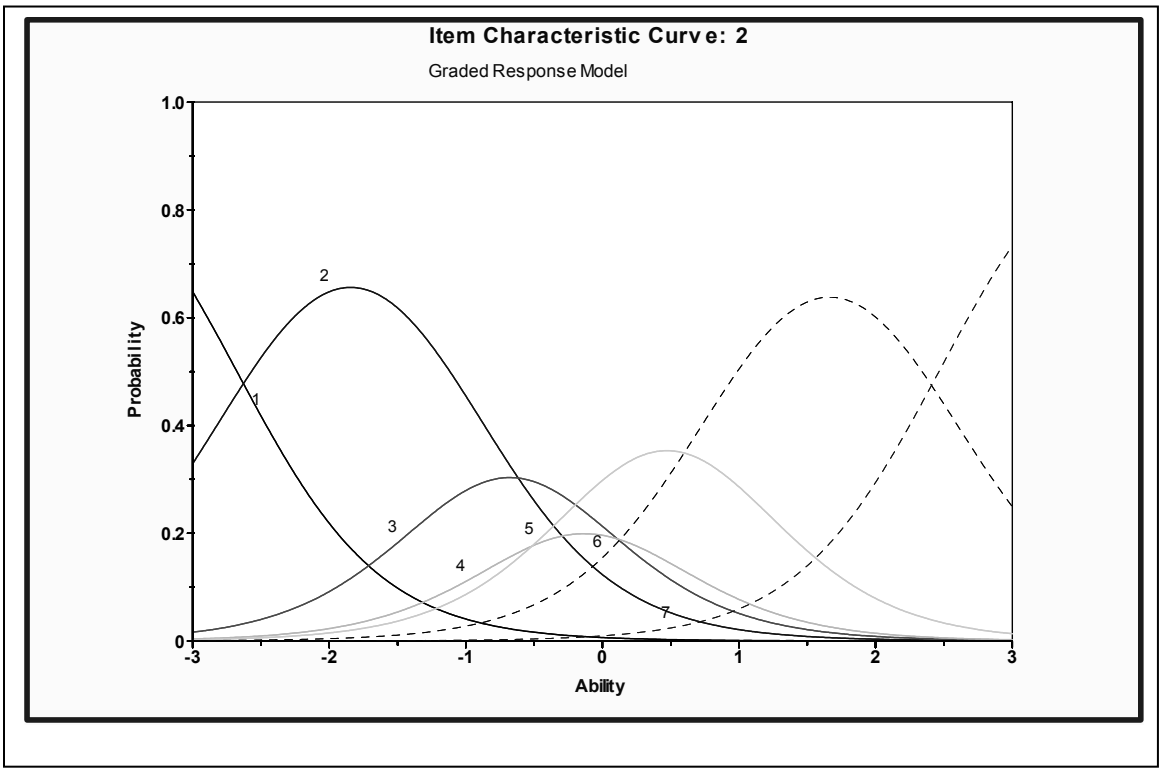
Figure Captions

Figure 1. The multiple cause multiple indicators (MIMIC) model

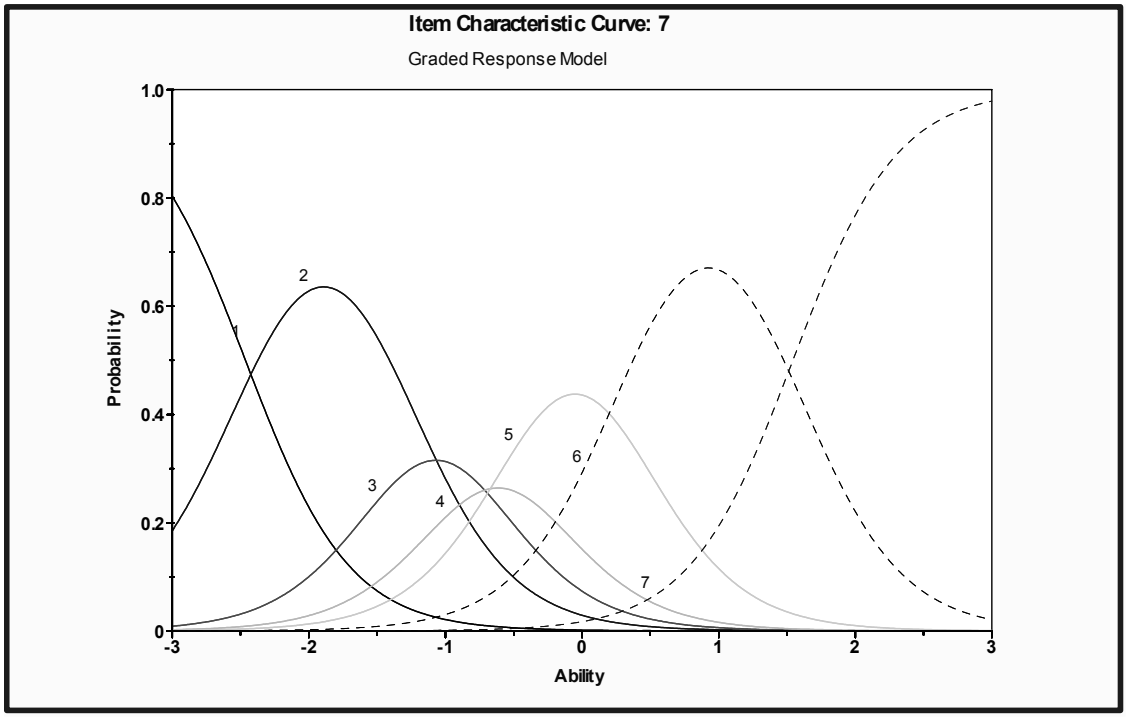
Figure 2. *Item characteristic curves for item 2 for Chinese (upper) and Americans (lower). Note.* Item characteristic curves: 2 indicates item 2 for Chinese; Item characteristic curves: 7 indicates item 2 for Americans. Ability indicates latent life satisfaction score in the standard deviation unit. Each of the curves plots the probability of endorsement, as the function of latent life satisfaction score.

Figure 3. *Item characteristic curves for item 5 for Chinese (upper) and Americans (lower). Note.* Item characteristic curves: 5 indicates item 5 for Chinese; Item characteristic curves: 10 indicates item 5 for Americans. Ability indicates latent life satisfaction score in the standard deviation unit. Each of the curves plots the probability of endorsement, as the function of latent life satisfaction score.

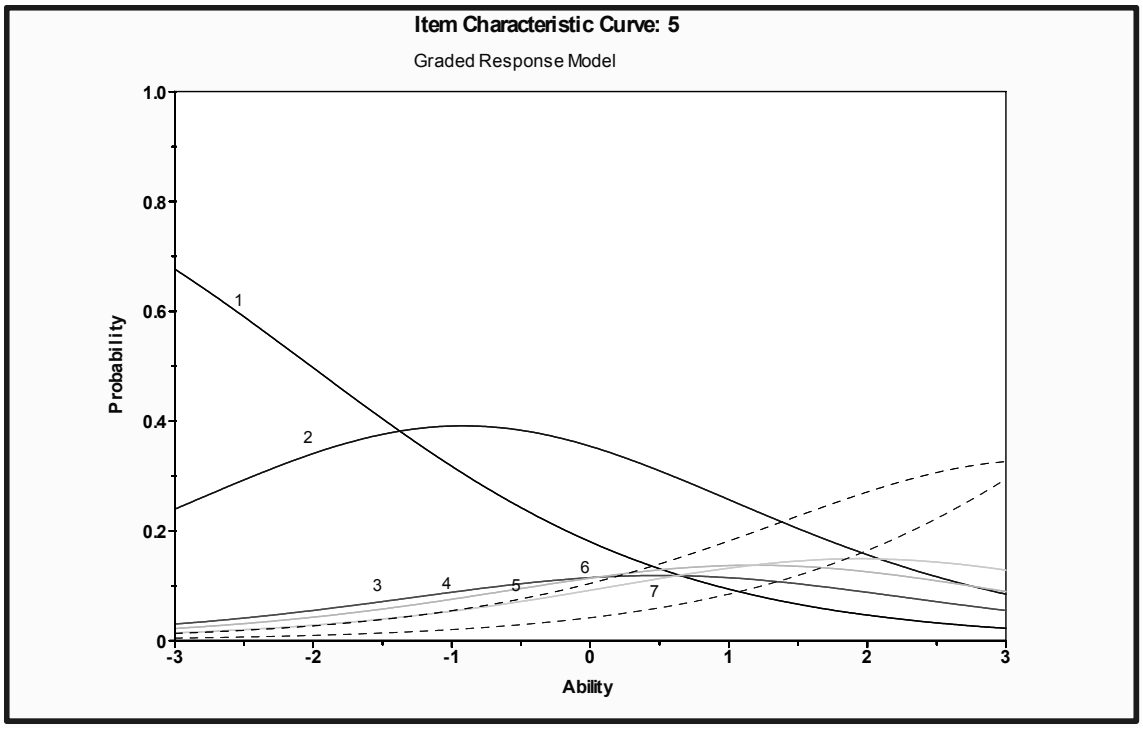

Life
Satisfaction
Group
Item 1
Item 2
Item 3
Item 4
Item 5



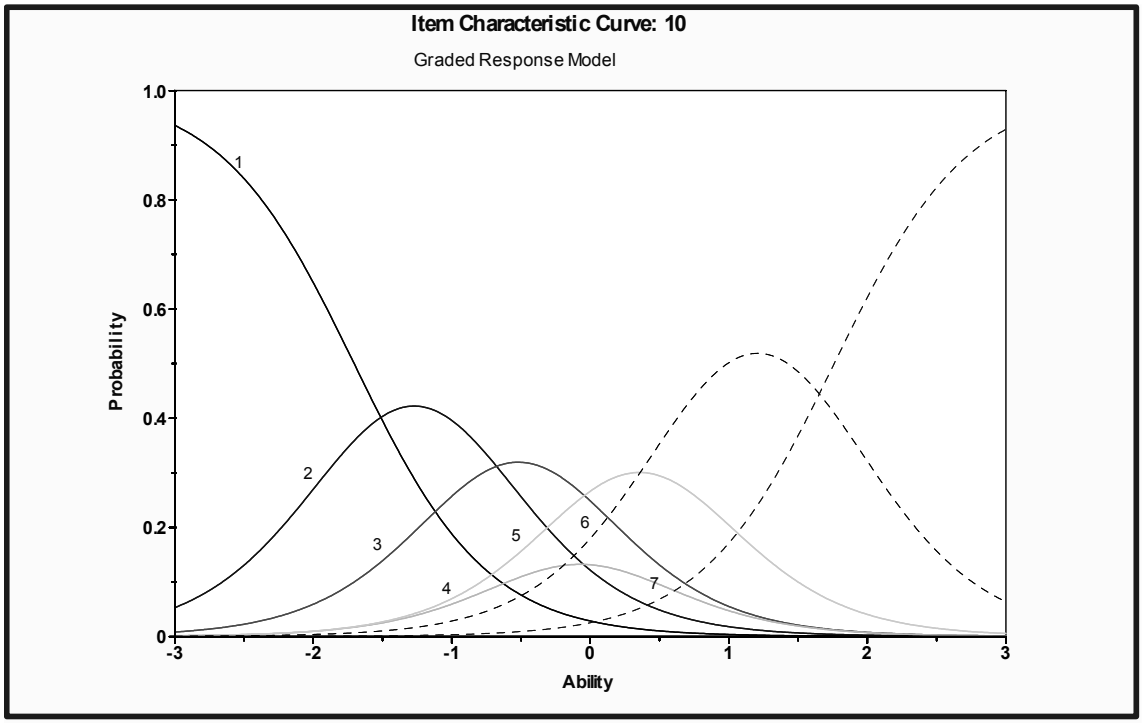
Item 2 for Chinese



Item 2 for Americans



Item 5 for Chinese



Item 5 for Americans