Modeling Extreme Response Style in Psychological Data: A Cross-Cultural Investigation

Leanne M. Stanley

This work addresses two related research questions: (1) How does psychological well-being compare among college students in the US and China? (2) Do these conclusions change when controlling for (vs. ignoring) cultural differences in extreme response style? Researchers in psychology and other social sciences routinely use item responses on multi-point rating scales to estimate individuals' standing on constructs of theoretical and practical interest. For example a researcher might ask participants to rate their overall happiness on a scale from 1 to 5, where 1 is *very unhappy* and 5 is *very happy*. However, in addition to focal constructs (such as happiness), item responses may be meaningfully influenced by response styles, which are characteristic ways in which participants use rating scales that may reflect personality differences which are stable across measures and over time (Bolt & Newton, 2011; Cronbach, 1946; Weijter, Geuens, & Schillewaert, 2010a; 2010b) and may interfere with the accurate measurement of focal constructs (Morren, Gelissen, & Vermunt, 2012).

Research on response styles has consistently demonstrated that individuals in Asian countries, such as China, are less likely to respond in the extremes (e.g., use the endpoints of a multi-point rating scale, such as 1 or 5 on a 5-point response scale) relative to individuals from western cultures, such as the US (de Jong et al., 2008; Hui & Triandis, 1989). This could complicate the interpretation of group differences if individuals with the same underlying levels on a construct (like happiness) appear to differ because of these cultural differences in extreme response style, which may instead reflect differences in how it is socially acceptable to portray oneself in western cultures (which tends to promote bold individualism) and Asian cultures (which tend to emphasize the importance of group harmony). This finding has become a

common motivating example for considering the potential impact of extreme response style in cross-cultural comparisons because ignoring its impact could negatively affect the validity of test score interpretations and uses.

Method

The present research focuses on statistical models for six correlated dimensions of subjective well-being (positive relationships with others, mastery over one's environment, personal growth, purpose in life, self-acceptance, and a sense of autonomy) and a general extreme response style dimension for college students in the US (N = 407) and China (N = 398). Participants completed the Ryff Scales of Psychological Well-Being (Ryff, 1989; Ryff & Keyes, 1995), rating nine items (e.g., "In general, I feel confident and positive about myself.") for each well-being dimension on a six-point Likert-type scale (1 = *strongly disagree*, 6 = *strongly agree*). A six-factor correlated traits model was fit to the data (separately for each country) using categorical confirmatory factor analysis (a psychometric model that is appropriate for item responses in ordered categories; Wirth & Edwards, 2007) to provide evidence that the six-dimensional structure was plausible in both countries based on the root mean square error of approximation (RMSEA; Steiger & Lind, 1983).

The model that is the focus of this investigation can be viewed as a multi-trait cross-cultural extension of the Two-Decision Model (TDM; Thissen-Roe & Thissen, 2013), which posits that two distinct processes underlie responses to items on Likert-type scales, as a person first chooses whether to respond positively, neutrally (if a middle option is available, which was not the case here), or negatively, and then determines the extremity of the response. It represents a departure from the typical assumption that Likert-type scales result in ordinal- or interval-level measurement because different patterns of item responses are linked to distinct constructs (Jeon

& De Boeck, 2016; Johnson & Bolt, 2010). Before fitting the item response model to the data, the item responses were recoded into two pseudo items for each original item response, a necessary first step in fitting IRTrees to data. The TDM pseduo item codes are are shown in Figure 1. Each of the original response options has a code for each of two pseudo item types – positive (agreement) and extreme response style (extremity) pseudo items. Rather than fitting the IRT model to the item responses directly, IRT models are instead fit to the pseudo item responses representing different parts of the item response.

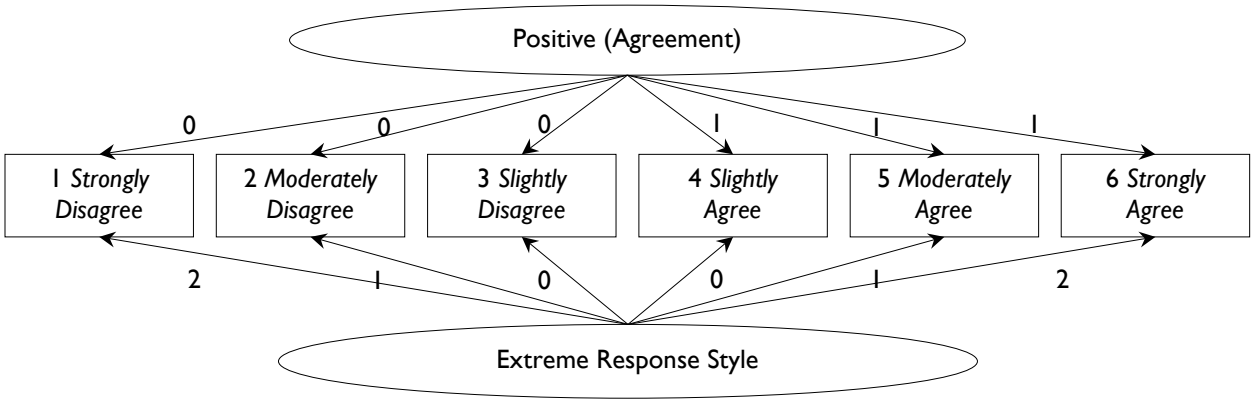

Figure 1. The TDM pseudo item codes for positive (vs. negative) responses and response extremity by rating scale option.

A seven-dimension MIRT model was fit to 108 pseudo items (the recoded item responses). For the 54 binary positive (agreement) pseduo items, a single slope and a single intercept were estimated using a graded response model (Samejima, 1969; see Appendix). For the 54 three-category extremity pseudo items, a single slope and two intercept parameters were estimated. The item parameters were constrained to be equal in the two countries so comparisons would be based on the same underlying scale. The mean of the latent variable prior distributions for the US were fixed to zero and the variances to unity. Correlations between all seven dimensions were estimated for the US responses. The means and variances of the latent variables for China were freely estimated, as were the covariances among all seven dimensions. The

results of this model were compared to a traditional sum score-based model (in which item responses for each dimension are added together to create a total score) in order to gauge the practical impact of fitting the TDM using multidimensional item response theory (MIRT) models relative to traditional classical test theory (sum score-based) methods that ignore the potential impact of extreme response style on test scores.

Results

The ordinal six-factor model provided a close fit to the data, RMSEAs $\leq .056$, suggesting that participants' responses to the 54 items were reasonably well-described in terms of the six well-being dimensions. The item parameters are presented in Table 1 (pseudo items 1-54 are positive (agreement) pseudo items and 55-108 are extremity pseduo items) and the variance-covariance matrices and average country differences for the seven dimensions are presented in Table 2. Extreme response style scores were less variable in China (variance of 0.42, relative to the US variance of 1) and lower on average, $M_{difference} = -0.56$. The correlations between the well-being dimensions and extreme response style in the US varied from 0.49 to 0.63 and these covariances in China ranged from 0.26 to 0.42.

The average differences and their 95% confidence intervals are presented in Figure 2. Results based on the traditional scoring method suggested that well-being among college students in the US was higher than or comparable to well-being among college students in China across all six dimensions. On the other hand, TDM results, which controlled for extreme response style (which was on average higher and also more variable among college students in the US), suggest that the Chinese students reported more positive relationships with others, mastery over their environments, and personal growth, comparable levels of purpose in life, and less self-acceptance and autonomy.

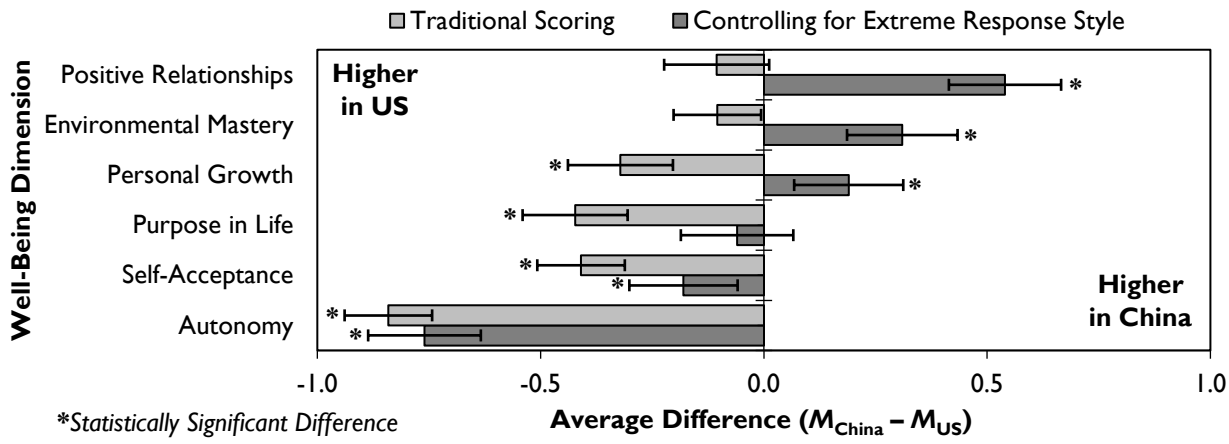

Figure 2. Average differences on six dimensions of psychological well-being. Scores are standardized for the US sample ($M = 0$, $SD = 1$). The error bars are 95% confidence intervals.

Discussion

Scientific inferences regarding differences in average well-being across six dimensions in the US and China changed dramatically when modeling (vs. ignoring) extreme response styles. This work highlights the importance of thinking critically about group differences that are identified based on traditional sum scores, and the potential usefulness of psychometric models that can take into account the influences of response styles in psychological data. The use of such models may lead to more scientifically sound inferences in cross-cultural comparisons. In this case, the results from the traditional method of comparing the groups resulted in dramatically different results regarding differences in six dimensions of psychological well-being for college students in the US and China. It is not possible to know whether the IRTree model studied here is the best possible model for the data because, as in all data analysis, the true model is unknown. However, the results suggest that cultural differences in the mean and variance of rating scale data could potentially bias scientific inferences based on statistical group comparsions, and that thus it may be a good idea for reasearchers to take extreme response style into account when comparing scores from research participants from countries for which there is *a priori* reason to believe that such cultural differences may exist, as was true here. In general, IRTrees may be

useful for testing assumptions regarding the measurement properties of rating scale data that have long been viewed as untestable. Other potentially useful models for response style based on a different set of data analytic techniques was recenlty introduced by Falk and Cai (2016).

References

Bolt, D. M., & Newton, J. R. (2011). Multiscale measurement of extreme response style. *Educational and Psychological Measurement, 71,* 814-833.

Cronbach, L. J. (1946) Response sets and test validity. *Educational and Psychological Measurement, 6,* 475-494.

de Jong, M. G., Steenkamp, J.-B. E. M., Fox, J.-P., & Baumgartner, H. (2008). Using item response theory to measure extreme response style in marketing research: A global investigation. *Journal of Marketing Research, 65,* 104-115.

Falk, C.F., & Cai, L. (2016). A flexible full-information approach to the modeling of response styles. *Psychological Methods, 21,* 328-347.

Hui, C. H., & Triandis, H. C. (1989). Effects of culture and response formal on extreme response style. *Journal of Cross-Cultural Psychology, 20,* 296-309.

Jeon, M., & De Boeck, P. (2016). A generalized item response tree model for psychological assessments. *Behavioral Research Methods, 48,* 1070-1085

Johnson, T. R., & Bolt, D. M. (2010). On the use of factor-analytic multinomial logit item response models to account for individual differences in response style. *Journal of Educational and Behavioral Statistics, 35,* 92-114.

Morren, M., Gelissen, J., & Vermunt, J. (2012). The impact of controlling for extreme responding on measurement equivalence in cross-cultural research. *Methodology, 8,* 159-217.

Morren, M., Gelissen, J., & Vermunt, J. (2012). The impact of controlling for extreme responding on measurement equivalence in cross-cultural research. *Methodology, 8,* 159-217.

Ryff, C. (1989). Happiness is everything, or is it? Explorations on the meaning of psychological well-being. *Journal of Personality and Social Psychology, 57,* 1069–1081.

Ryff, C., & Keyes, C. (1995). The structure of psychological well-being revisited. *Journal of Personality and Social Psychology, 69,* 719–727.

Samejima, F. (1969). Estimation of latent ability using a response pattern of graded scores. *Psychometrika Monograph No. 17,* 34.

Steiger, J. H., & Lind, J. C. (1980, June). *Statistically based tests for the number of factors.* Paper presented at the annual meeting of the Psychometric Society, Iowa City, IA.

Thissen-Roe, A., & Thissen, D. (2013). A two-decision model for responses to Likert-type items. *Journal of Educational and Behavioral Statistics, 38,* 522-547.

Weijters, B., Geuens, M., & Schillewaert, N. (2010a). The individual consistency of acquiescence and extreme response style in self-report questionnaires. *Applied Psychological Measurement, 34,* 105-121.

Weijters, B., Geuens, M., & Schillewaert, N. (2010b). The stability of individual response styles. *Psychological Methods, 15,* 96-111.

Wirth, R. J., & Edwards, M. C. (2007). Item factor analysis: Current approaches and future directions. *Psychological Methods, 12,* 58-79.

Table 1

Two-Decision Model Estimated Item Parameters

Pseudo Item	Dimension	α_1	α_2	α_3	α_4	α_5	α_6	α_7	β_1	β_2
1	Positive	1.08	0	0	0	0	0	0	2.33	1
2	Relationships	1.43	0	0	0	0	0	0	0.66	2
3	with Others	1.59	0	0	0	0	0	0	0.23	3
4		0.70	0	0	0	0	0	0	1.91	4
5		1.88	0	0	0	0	0	0	1.46	5
6		1.93	0	0	0	0	0	0	-0.39	6
7		0.93	0	0	0	0	0	0	1.92	7
8		1.70	0	0	0	0	0	0	1.46	8
9		1.84	0	0	0	0	0	0	2.54	9
10	Environmental	0	1.40	0	0	0	0	0	2.07	10
11	Mastery	0	1.58	0	0	0	0	0	0.35	11
12		0	1.05	0	0	0	0	0	1.33	12
13		0	1.11	0	0	0	0	0	1.98	13
14		0	0.80	0	0	0	0	0	-0.38	14
15		0	0.77	0	0	0	0	0	1.51	15
16		0	1.03	0	0	0	0	0	0.74	16
17		0	2.60	0	0	0	0	0	0.59	17
18		0	0.97	0	0	0	0	0	0.57	18
19	Personal	0	0	1.12	0	0	0	0	3.11	19
20	Growth	0	0	0.24	0	0	0	0	1.43	20
21		0	0	1.01	0	0	0	0	3.61	21
22		0	0	1.65	0	0	0	0	1.27	22
23		0	0	2.92	0	0	0	0	4.19	23
24		0	0	0.58	0	0	0	0	-0.46	24
25		0	0	1.61	0	0	0	0	4.33	25
26		0	0	1.54	0	0	0	0	2.35	26
27		0	0	0.41	0	0	0	0	1.32	27
28	Purpose in Life	0	0	0	0.65	0	0	0	1.08	28
29		0	0	0	0.83	0	0	0	1.31	29
30		0	0	0	0.95	0	0	0	1.05	30
31		0	0	0	1.57	0	0	0	1.14	31
32		0	0	0	1.25	0	0	0	1.89	32
33		0	0	0	1.96	0	0	0	2.77	33
34		0	0	0	1.66	0	0	0	1.85	34
35		0	0	0	1.59	0	0	0	1.87	35
36		0	0	0	0.41	0	0	0	2.07	36

Table 1 (Continued)

Two-Decision Model Estimated Item Parameters

Pseudo Item	Dimension	α_1	α_2	α_3	α_4	α_5	α_6	α_7	β_1	β_2
37	Self-	0	0	0		1.68	0	0	0.92	
38	Acceptance	0	0	0		2.03	0	0	2.65	
39		0	0	0		1.26	0	0	0.44	
40		0	0	0		1.74	0	0	2.36	
41		0	0	0		1.20	0	0	2.35	
42		0	0	0		2.51	0	0	1.53	
43		0	0	0		1.36	0	0	-0.08	
44		0	0	0		1.46	0	0	1.97	
45		0	0	0		1.11	0	0	1.11	
46	Autonomy	0	0	0	0	0	1.42	0	1.58	
47		0	0	0	0	0	1.27	0	0.64	
48		0	0	0	0	0	1.54	0	-0.76	
49		0	0	0	0	0	1.01	0	2.15	
50		0	0	0	0	0	1.02	0	0.27	
51		0	0	0	0	0	1.79	0	2.75	
52		0	0	0	0	0	0.92	0	1.37	
53		0	0	0	0	0	1.32	0	0.75	
54		0	0	0	0	0	0.89	0	2.21	
55	Positive	0	0	0	0	0	0	0.87	1.40	-0.87
56	Relationships	0	0	0	0	0	0	0.96	1.17	-0.93
57	with Others	0	0	0	0	0	0	0.99	1.04	-0.86
58		0	0	0	0	0	0	1.31	1.79	-0.28
59		0	0	0	0	0	0	1.18	1.45	-0.86
60		0	0	0	0	0	0	1.00	0.97	-1.12
61		0	0	0	0	0	0	1.46	1.36	-0.87
62		0	0	0	0	0	0	1.28	1.43	-0.65
63		0	0	0	0	0	0	1.29	1.75	-0.45
64	Environmental	0	0	0	0	0	0	1.26	1.16	-1.72
65	Mastery	0	0	0	0	0	0	0.92	0.50	-1.78
66		0	0	0	0	0	0	1.00	1.17	-0.93
67		0	0	0	0	0	0	1.42	0.78	-1.96
68		0	0	0	0	0	0	0.52	0.30	-1.77
69		0	0	0	0	0	0	1.07	1.10	-1.30
70		0	0	0	0	0	0	1.12	0.21	-2.01
71		0	0	0	0	0	0	1.43	0.67	-2.04
72		0	0	0	0	0	0	1.52	0.44	-1.89

Table 1 (Continued)

Two-Decision Model Estimated Item Parameters

Pseudo Item	Dimension	α_1	α_2	α_3	α_4	α_5	α_6	α_7	β_1	β_2
73	Personal	0	0	0	0	0	0	0.98	2.30	0.00
74	Growth	0	0	0	0	0	0	0.78	0.57	-1.57
75		0	0	0	0	0	0	0.82	1.78	-0.31
76		0	0	0	0	0	0	1.72	1.41	-0.99
77		0	0	0	0	0	0	1.89	1.48	-0.99
78		0	0	0	0	0	0	0.75	0.36	-2.03
79		0	0	0	0	0	0	1.38	2.87	0.39
80		0	0	0	0	0	0	1.66	1.80	-0.43
81		0	0	0	0	0	0	0.60	1.16	-0.62
82	Purpose in Life	0	0	0	0	0	0	0.43	0.82	-0.86
83		0	0	0	0	0	0	1.21	0.72	-1.65
84		0	0	0	0	0	0	1.38	0.87	-1.54
85		0	0	0	0	0	0	1.52	1.36	-0.68
86		0	0	0	0	0	0	1.76	1.60	-0.74
87		0	0	0	0	0	0	1.71	1.37	-1.17
88		0	0	0	0	0	0	1.78	0.72	-1.75
89		0	0	0	0	0	0	1.23	1.13	-1.14
90		0	0	0	0	0	0	0.84	1.77	-0.05
91	Self-	0	0	0	0	0	0	1.25	0.53	-1.68
92	Acceptance	0	0	0	0	0	0	1.35	1.32	-1.21
93		0	0	0	0	0	0	1.13	0.59	-1.67
94		0	0	0	0	0	0	1.66	1.53	-1.64
95		0	0	0	0	0	0	1.61	1.30	-1.31
96		0	0	0	0	0	0	2.06	1.27	-1.24
97		0	0	0	0	0	0	0.96	0.61	-1.42
98		0	0	0	0	0	0	1.32	1.69	-0.14
99		0	0	0	0	0	0	1.17	0.99	-1.16
100	Autonomy	0	0	0	0	0	0	1.06	0.61	-1.56
101		0	0	0	0	0	0	0.99	0.18	-2.02
102		0	0	0	0	0	0	0.41	0.65	-1.55
103		0	0	0	0	0	0	1.07	0.93	-0.83
104		0	0	0	0	0	0	0.69	0.36	-1.65
105		0	0	0	0	0	0	1.66	0.39	-1.98
106		0	0	0	0	0	0	1.08	0.75	-1.38
107		0	0	0	0	0	0	0.95	0.27	-1.83
108		0	0	0	0	0	0	1.52	1.19	-1.23

Table 2

Variance-Covariance Matrices and Average Differences for the Dimensions by Country

Country	Dimension	PR	EM	PG	PL	SA	AU	ERS
US	Positive Relationships (PR)	1						
	Environmental Mastery (EM)	0.73	1					
	Personal Growth (PG)	0.58	0.68	1				
	Purpose in Life (PL)	0.51	0.73	0.72	1			
	Self-Acceptance (SA)	0.73	0.84	0.72	0.78	1		
	Autonomy (AU)	0.34	0.51	0.46	0.49	0.51	1	
	Extreme Response Style (ERS)	0.49	0.56	0.54	0.63	0.61	0.49	1
China	Positive Relationships (PR)	1.11						
	Environmental Mastery (EM)	0.82	1.08					
	Personal Growth (PG)	0.61	0.75	1.17				
	Purpose in Life (PL)	0.55	0.81	0.99	1.44			
	Self-Acceptance (SA)	0.67	0.72	0.54	0.54	0.75		
	Autonomy (AU)	0.37	0.47	0.38	0.47	0.48	0.96	
	Extreme Response Style (ERS)	0.32	0.27	0.32	0.33	0.26	0.28	0.42
Average Difference (θ_{China} - θ_{US})			0.54	0.31	0.19	-0.06	-0.18	-0.76
Standard Error			0.06	0.05	0.06	0.06	0.05	0.05

Appendix

The graded response model is the most popular IRT model for Likert-type item responses in psychology. For item responses in M ordered categories ($m = 0, \ldots, M - 1$) and a single latent construct, the probability of that person p will give a response to item i (Y_{pi}) in category m or higher can be modeled after conditioning on that person's latent trait score (θ_p):

$$P(Y_{pi} \geq m|\theta_p) = g^{-1}(\alpha_i\theta_p + \beta_{im})$$

where g^{-1} is the inverse logit link function,[1] α_i is an item-specific slope that summarizes the strength of the relationship between an item response pattern and the latent trait (θ_p), and β_{im} is an item- and category-specific intercept. By definition, $P(Y_p \geq 0|\theta_p) = 1$ and $P(Y_p \geq M|\theta_p) = 0$. The probability of observing a response in category m is the difference between the probability of responding in that category and the next highest category ($m + 1$):

$$P(Y_p = m|\theta_p) = P(Y_p \geq m|\theta_p) - P(Y_p \geq m + 1|\theta_p).$$

This slope-intercept parameterization of the unidimensional graded response model is easily extended to accommodate multiple latent constructs:

$$P(Y_{pi} \geq m|\boldsymbol{\theta}_p) = g^{-1}(\boldsymbol{\alpha}_i'\boldsymbol{\theta}_p + \beta_{im}),$$

where $\boldsymbol{\alpha}_i$ and $\boldsymbol{\theta}_p$ are D-length vectors ($d = 1, \ldots, D$) of item slopes and person scores on the D dimensions, respectively.

[1] A probit link function may also be used.