Soc Indic Res (2015) 124:249–258
DOI 10.1007/s11205-014-0773-9

An Item Response Theory Analysis of the Subjective Happiness Scale

Brian P. O'Connor · Maxine R. Crawford · Mark D. Holder

Accepted: 20 September 2014 / Published online: 30 September 2014

Abstract Item response theory analyses were conducted on Lyubomirsky and Lepper's (Soc Indic Res 46:137–155, 1999) Subjective Happiness Scale (SHS) in a sample of 1,037 undergraduates. In general, the SHS performed well. However, four, clear, previously unreported findings emerged that suggest modifications for improving the scale. The measure displayed weak discrimination power at the high, happy end of the latent trait continuum. The number of response options (seven) is excessive and unnecessary. One of the SHS items is problematic and provides minimal psychometric information relative to the other three items, apparently because of ambiguity in the wording of the item. And a three-item version of the SHS performed just as well as the original, four-item version. The implications of these findings for the usage and further development of the SHS are discussed.

Keywords Happiness · Wellbeing · Psychometrics · Item response theory

1 Introduction

Happy people enjoy more successful social relationships, have better physical health, better immune systems, they sleep better, have more successful careers, display less racial intolerance, have lower rates of suicide, and they live longer (Lyubormirsky et al. 2005a). Research on the factors that promote happiness requires reliable, valid, and sensitive

B. P. O'Connor (✉) · M. R. Crawford · M. D. Holder
Department of Psychology, University of British Columbia - Okanagan,
ART 330 - 1147 Research Road, 3333 University Way, Kelowna, BC V1V 1V7, Canada
e-mail: brian.oconnor@ubc.ca

M. R. Crawford
e-mail: maxcrawford@shaw.ca

M. D. Holder
e-mail: mark.holder@ubc.ca

measures of wellbeing. Most studies use self-reports (Brunel et al. 2004; DeNeve and Cooper 1998; Lyubomirsky and Lepper 1999), at least partly because happiness is a personal, subjective experience and each person is the best judge of his or her own happiness (Diener 1984; Lyubomirsky et al. 2005b; Myers and Diener 1995). One prominent measure is Lyubomirsky and Lepper's four-item Subjective Happiness Scale (SHS). When using the SHS, respondents indicate both their own levels of happiness and their interpretations of their own happiness in relation to others. The original measure consisted of 13 items, and psychometric analyses reduced the inventory to just four non redundant, unidimensional items (Lyubomirksy and Lepper 1999).

As described by Lyubomirksy and Lepper (1999), the SHS has undergone rigorous testing in large samples of high school students, college students, adults, and retired persons from the U.S. and Russia. Test–retest correlations range between 0.55 and 0.90 for 3-week to 1-year time spans. Convergent validity has been established via correlations between the SHS and five other measures of aspects of wellbeing: the Satisfaction With Life Scale, the Affect Balance Scale, the Delighted-Terrible Scale, the Global Happiness Item, and the Recent Happiness Item. The correlations between the SHS and these five other measures ranged from 0.52 to 0.72. The correlations between the SHS and measures of emotions and personality characteristics that are known to be associated with happiness (self-esteem, optimism, positive emotion, negative emotion, extraversion, neuroticism) ranged from 0.36 to 0.60. These generally moderate convergent validity correlations were viewed by Lyubomirksy and Lepper (1999) as supporting their argument that the SHS measures a construct that is similar to, but sufficiently separate from, other measures.

The present study used item response theory (IRT) methods to provide more precise psychometric information about the SHS than what has been provided in previous research. The SHS was developed using classical test theory (CTT) methods, which have been used in all previously published reports (to our knowledge) on the SHS. CTT methods provide statistics such as Cronbach's alpha, item-total correlations, and factor loadings, that are assumed apply to an entire sample, i.e., to all levels of a latent trait continuum. IRT methods, in contrast, are more precise and informative because they reveal how items and tests perform at all levels of a latent trait (for reviews, see de Ayala 2009; DeMars 2010; Edelen and Reeve 2007; Morizot et al. 2007; Reise et al. 2005).

IRT methods provide latent trait estimates for respondents and two kinds of parameters for the individual items. The item discrimination parameter reflects the magnitude of the association between an item and the latent trait. Higher discrimination values indicate more sensitive discriminations between respondents, whereas discrimination values near zero indicate that an item provides little information about the latent trait. The threshold parameter is an index of item difficulty, indicating the level of the latent trait that is required for an item response option to be endorsed. An information function can be obtained for each item, which reveals an item's ability to differentiate between respondents located at various points along the trait level continuum. The information functions for the test items are then added to obtain the scale information function, which indicates how well the entire test functions in different ranges of the latent trait continuum.

IRT analyses can thus reveal whether a measure is more informative (or reliable) at some sections of a latent trait continuum than it is at other sections. It can expose items that are redundant with other items or that are otherwise not operating properly. And it can reveal potential problems with item response options.

IRT analyses of the SHS should be informative because the measure is very short, with just four items. Is the scale too short? Are any items redundant or problematic? Are all seven response options for each item used properly by respondents? Perhaps most important, IRT analyses can determine whether the SHS has sufficiently strong discrimination power at all levels of the latent trait continuum. Given the extensive use of the SHS in positive psychology research, we were most curious to discover whether the measure performs well at the high, happy end of the happiness continuum.

2 Method

2.1 Participants and Procedure

One thousand thirty-seven undergraduates (346 males and 691 females), aged 17–47 (M = 19.7, SD = 3.8) completed the SHS. They were recruited online through a departmental subject pool, which at our institution constitutes a diverse and representative university sample. Remuneration consisted of a small academic credit that participants could apply to a psychology course. Participants completed the SHS online at their convenience in a location of their choosing. The data were all collected in a single "sweep", and we did not use multi-stage sampling despite the large N. The reliability and validity of online SHS administrations are similar to those for paper-based administrations (Howell et al. 2010).

2.2 Measure

The SHS consists of the following four items, which are rated on 7-point Likert-type scales. Item 1: "In general, I consider myself: (1) not a very happy person, (7) a very happy person". Item 2: "Compared to most of my peers, I consider myself: (1) less happy, (7) more happy". Item 3: "Some people are generally very happy. They enjoy life regardless of what is going on, getting the most out of everything. To what extent does this characterization describe you? (1) not at all, (7) a great deal". And Item 4: "Some people are generally not very happy. Although they are not depressed, they never seem as happy as they might be. To what extent does this characterization describe you? (1) not at all, (7) a great deal".

2.3 Analytic Methods

The implementation of Samejima's (1969) graded response model (GRM) in the ltm package in R (Rizopoulos 2006) was used for the IRT analyses of the polytomous SHS items. For polytomous data, the number of threshold (or difficulty) parameters for each item is the number of response options minus one. The SHS items have a seven-point response scale, which results in the following six GRM response dichotomies: (1) Option 1 versus Options 2, 3, 4, and 5; (2) Options 1 and 2 versus Options 3, 4, and 5; and (3) Options 1, 2, and 3 versus Options 4 and 5; (4) Options 1, 2, 3 and 4 versus Options 5 and 6; (5) Options 1, 2, 3, 4 and 5 versus Option 6; and (6) Options 1, 2, 3, 4, 5 and 6 versus Option 7. The threshold parameter for each of these response dichotomies represents the location on the latent trait continuum where there is a 50 % probability of endorsing the higher response option(s).

3 Results

The means, standard deviations, and correlations between the four SHS items are provided in Table 1. In accordance with the scoring requirements for this measure (Lyubomirksy

Table 1 Item means, standard deviations, and Pearson correlations

	Mean	SD	Item 1	Item 2	Item 3
Item 1	5.39	1.26			
Item 2	4.97	1.38	0.79		
Item 3	4.76	1.46	0.70	0.71	
Item 4	5.06	1.59	0.53	0.51	0.58

Table 2 Response option frequencies

	Option 1	Option 2	Option 3	Option 4	Option 5	Option 6	Option 7
Item 1	14	19	40	146	253	383	182
Item 2	22	25	99	203	285	273	130
Item 3	27	54	124	184	298	243	107
Item 4	26	49	124	158	162	317	201

and Lepper 1999), the responses for Item 4 were reflected before all of the analyses. Cronbach's alpha for the four items was 0.87. The item response option frequencies are provided in Table 2.

The first eigenvalue in the matrix of polychoric item inter-correlations, 2.86, was 8.5 times larger than the second eigenvalue, 0.34, indicating a single dominant dimension in the item pool (Morizot et al. 2007). Parallel analysis and Velicer's minimum average partial test both indicated just one factor. The Goodness of Fit Coefficient (which is often provided in structural equation modeling analyses) for a one-component model was 0.99.

The discrimination and threshold parameters from the GRM analyses appear in Table 3, and the item and test information functions appear in Figs. 1 and 2. The functions are slightly wavy because they are based on so few items. The item and test information functions all tail off at the upper end of the latent trait continuum, most notably above $z = 1.5$. The discrimination parameter for Item 4, 1.64, was much lower than the values for the other items (which were all 2.92 or above). The information function for Item 4 indicates that this item functions poorly at all ranges of the latent trait continuum.

The item response option curves are provided in Fig. 3. Such curves can reveal response options that are redundant or that are rarely or improperly by respondents. The curves for all four items indicate that there are too many response options. Response option 2 was rarely used, for all items, by participants in the relevant (low) region of the latent trait, and there was generally insufficient differentiation between the usages of response options 2, 3, and 4. The problem was particularly severe for Item 4.

The GRM analyses were therefore conducted once again without the problematic Item 4. The threshold and discrimination parameters for the three-item scale appear in Table 3. The item information functions appear in Fig. 4 and the test information function appears in Fig. 1. The results indicate that assessment of the latent variable is not affected by the elimination of Item 4. The item and test information functions were highly similar to those for the four-item scale (Figs. 1, 2). The test information functions for the three- and four-item scales were then converted into reliability functions and the results are plotted in Fig. 5. There were essentially no differences between the reliabilities for the two scales across the full length of the latent trait continuum. The response option curves for the three-item scale are very similar to the curves that appear in Fig. 3.

Table 3 Discrimination and threshold parameters for 4-item and 3-item SHS Scales

	Discr. index	Threshold indices					
	a	b1	b2	b3	b4	b5	b6
4 Items							
Item 1	3.95	−2.47	−2.09	−1.68	−0.94	−0.12	1.00
Item 2	3.48	−2.34	−1.95	−1.27	−0.49	0.33	1.25
Item 3	2.92	−2.32	−1.73	−1.06	−0.39	0.49	1.45
Item 4	1.64	−2.96	−2.19	−1.32	−0.63	−0.03	1.23
3 Items							
Item 1	3.98	−2.51	−2.10	−1.66	−0.91	−0.10	1.01
Item 2	3.75	−2.33	−1.93	−1.24	−0.46	0.34	1.25
Item 3	2.70	−2.38	−1.75	−1.06	−0.38	0.52	1.49

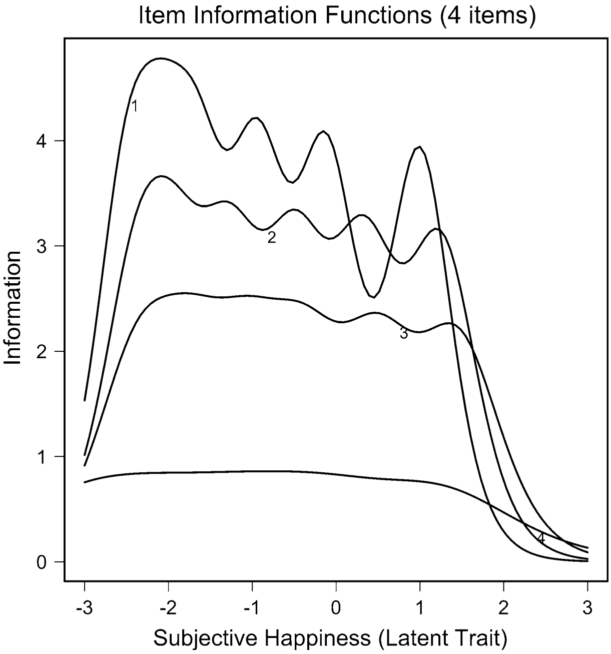

Fig. 1 Item information functions for the Subjective Happiness Scale (4 items)

IRT analyses were also conducted using the nonparametric procedures in the TestGraf program (Ramsay 1993; see also Santor and Ramsay 1998). The output is primarily graphic, and item parameters are not provided by this IRT method. The procedures are nevertheless useful for confirming findings from other methods, such as those from Samejima's GRM. The nonparametric IRT analyses of the present data produced plots that were similar to those for the GRM that are provided in this manuscript. We also ran separate analyses (both GRM and non parametric) for males and females and found no differences in the findings and no evidence of gender bias.

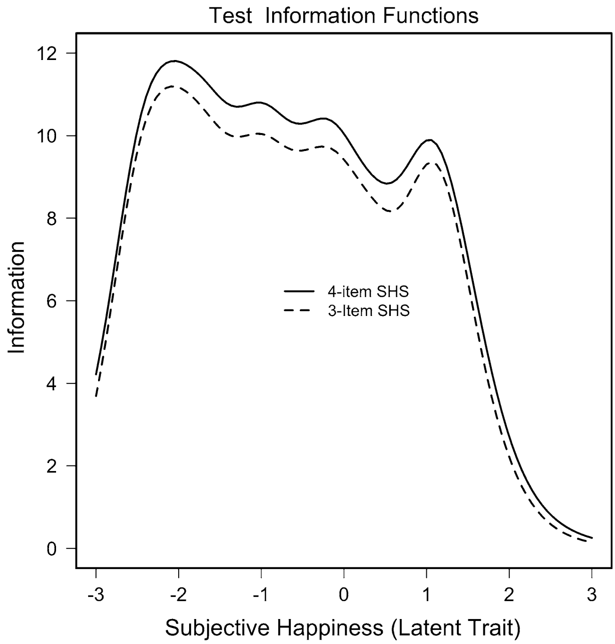

Fig. 2 Test information functions

4 Discussion

Four clear, previously unreported findings emerged from the present IRT analyses of the SHS. The measure displayed weak discrimination power at the high, happy end of the latent trait continuum. The number of response options (seven) is excessive and unnecessary. One of the SHS items is problematic and provides minimal psychometric information relative to the other three items, apparently because of ambiguity in the wording of the item. And a three-item version of the SHS performed just as well as the original, four-item version.

The weak discrimination power at the high end of the latent trait continuum is surprising and it has an unfortunate implication. The SHS is used extensively in positive psychology research where the primary concern is with happy respondents and in which it is common for people of all ages to report high levels of happiness. The weak discrimination power in the upper range means that differences between the scores for persons in this region are not meaningful. High-end SHS scores are almost interchangeable. The primary consequence for researchers is underestimation of the effect sizes for associations between the SHS and other measures or constructs. The attenuation will be most problematic when researchers seek to identify variables that predict scores at the high end of the happiness continuum.

One intuitively meaningful method of increasing discrimination power might be to increase the number of response options. But the present findings indicate that the SHS already has too many response options (seven). Response option 2 was rarely used, and there was generally insufficient differentiation by participants between response options 2, 3, and 4. We suspect that a switch from seven to five response options would be beneficial, as participants apparently do not understand the differences between the current seven

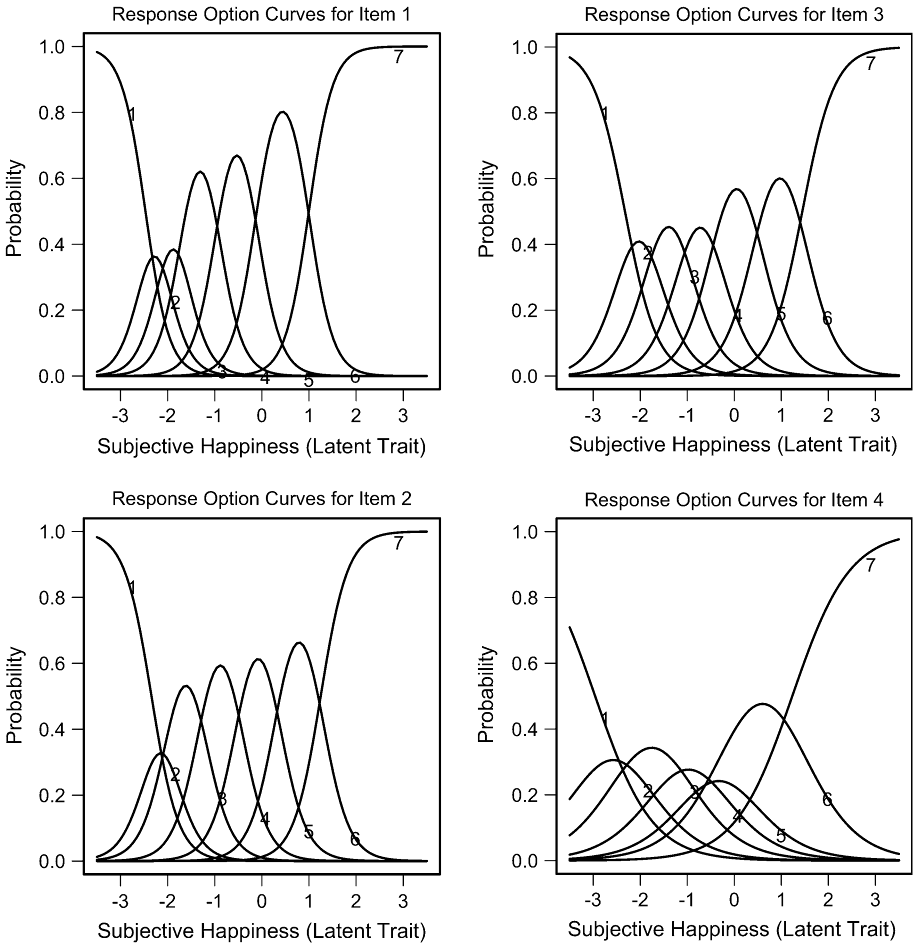

Fig. 3 Response option curves for the Subjective Happiness Scale

response options. Another possibility would be to keep the seven response options but to provide respondents with a written description (label) of the meaning of each number on the response option scale. Only the first and seventh response options currently have such labels, which is a likely reason why respondents are confusing the response options in between. Providing labels might also increase discrimination power at the high end of the latent trait continuum.

Another clear finding was the poor performance of Item 4: “Some people are generally not very happy. Although they are not depressed, they never seem as happy as they might be. To what extent does this characterization describe you?”. The correlations between this item and the other three items were slightly but uniformly weaker than the inter-correlations between the other three items. More serious is the fact that the IRT analyses revealed Item 4 to be uninformative. The information function for this item was distinctly low across the full range of the latent trait continuum. The item may assess a dimension that is different from the latent trait that is assessed by the other SHS items. We suspect that the

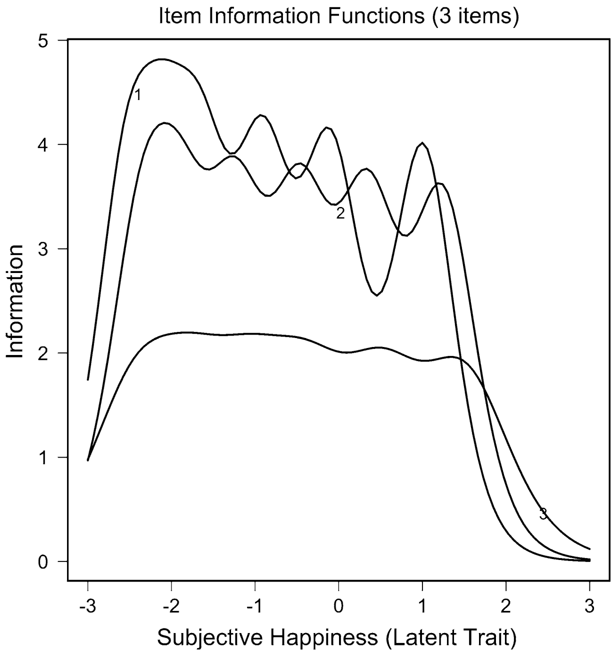

Fig. 4 Item information functions for the Subjective Happiness Scale (3 items)

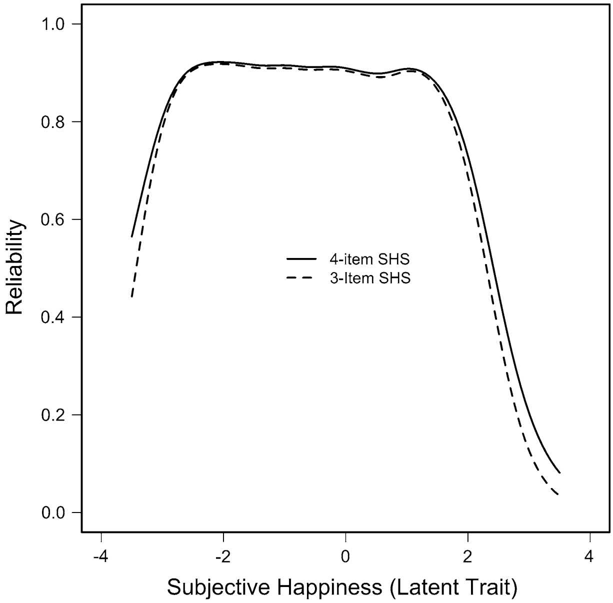

Fig. 5 Reliability functions for the 4-item and 3-item Subjective Happiness Scales

problem is with the wording of the item. The item asks people the extent to which they are simultaneously not depressed and not happy. A literal interpretation of the Item 4 language is that it is asking respondents whether they consider themselves to be in the not depressed, not happy, middle of the continuum. If a respondent selects the response option "not at all" this could indicate that they are either very happy or very depressed. Item four is thus a fundamentally different question from what is asked in the other three SHS items. The consequence is a low and flat psychometric information function for Item 4. We suspect that some, but not all, respondents are confused by Item 4. We also suspect that some respondents may be able to rise above the ambiguity after they read the content of the first three items and then guess at the true purpose of Item 4.

We also found that a three-item version of the SHS, without the problematic Item 4, performed just as well as the four-item version. This is remarkable for an already brief measure that is reduced in length by 25 %. Item 4 is thus clearly an uninformative filler question that does not improve the measurement of happiness. Dropping Item 4 should not cause an attenuation in effect sizes for SHS associations with other variables, and it might make the measure less confusing and annoying for respondents.

IRT analyses almost invariably expose previously unknown problems with measures that were developed using CTT methods. The present findings should therefore be viewed in this context. The SHS generally performs very well. IRT analyses have revealed more serious problems with popular measures of other constructs. The present IRT analyses indicate that the already brief SHS can be further simplified: An item can be dropped, and the number of response options can be reduced. The bigger challenge for further development of the measure is to find ways of enhancing fidelity, or discrimination power, among persons with elevated levels of happiness. Subtle language in the wording of a new item, or for the labels of the SHS response options, might be all that is needed.

References

Brunel, F. F., Tietje, B. C., & Greenwald, A. G. (2004). Is the implicit association test a valid and valuable measure of implicit consumer social cognition? *Journal of Consumer Psychology, 14*, 385–404. doi:10.1207/s15327663jcp1404_8.

de Ayala, R. J. (2009). *The theory and practice of item response theory*. New York, NY: The Guildford Press.

DeMars, C. (2010). *Item response theory*. New York, NY: Oxford University Press.

DeNeve, K. M., & Cooper, H. (1998). The happy personality: A meta-analysis of 137 personality traits and subjective wellbeing. *Psychological Bulletin, 124*, 197–229. doi:10.1037/0033-2909.124.2.197.

Diener, E. (1984). Subjective wellbeing. *Psychological Bulletin, 95*(3), 542–575. doi:10.1037/0033-2909.95.3.542.

Edelen, M. O., & Reeve, B. B. (2007). Applying item response theory (IRT) modeling to questionnaire development, evaluation, and refinement. *Quality of Life Research, 16*, 5–18. doi:10.1007/s11136-007-9198-0.

Howell, R. T., Rodzon, K. S., Kurai, M., & Sanchez, A. H. (2010). A validation of wellbeing and internet surveys for administration via the internet. *Behavior Research Methods, 42*, 775–784. doi:10.3758/BRM.42.3.775.

Lyubomirksy, S., & Lepper, H. S. (1999). A measure of subjective happiness: Preliminary reliability and construct validation. *Social Indicators Research, 46*, 137–155. doi:10.1023/A:1006824100041.

Lyubomirsky, S., King, L., & Diener, E. (2005a). The benefits of frequent positive affect: Does happiness lead to success? *Psychological Bulletin, 131*, 803–855. doi:10.1037/0033-2909.131.6.803.

Lyubomirsky, S., Sheldon, K. M., & Schkade, D. (2005b). Pursuing happiness: The architecture of sustainable change. *Review of General Psychology, 9*, 111–131. doi:10.1037/1089-2680.9.2.111.

Morizot, J., Ainsworth, A. T., & Reise, S. P. (2007). Toward modern psychometrics: Application of item response theory models in personality research. In R. W. Robins, R. C. Fraley, & R. F. Krueger (Eds.), *Handbook of research methods in personality psychology* (pp. 407–423). New York, NY: Guilford Press.

Myers, D. G., & Diener, E. (1995). Who is happy? *Psychological Science, 6*, 10–19. doi:10.1111/j.1467-9280.1995.tb00298.x.

Ramsay, J. O. (1993). TESTGRAF: A program for the graphical analysis of multiple choice test and questionnaire data. Unpublished manuscript, McGill University, Montreal, Quebec, Canada. (http://www.psych.mcgill.ca/faculty/ramsay/TestGraf.html).

Reise, S. P., Ainsworth, A. T., & Haviland, M. G. (2005). Item response theory. *Current directions in psychological science, 14*(2), 95–101. doi:10.1111/j.0963-7214.2005.00342.x.

Rizopoulos, D. (2006). Ltm: An R package for latent variable modeling and item response theory analyses. *Journal of Statistical Software, 17* 1–25.

Samejima, F. (1969). Estimation of latent ability using a response pattern of graded scores. *Psychometrika Monograph Supplement, 34*(4), 100.

Santor, D. A., & Ramsay, J. O. (1998). Progress in the technology of measurement: applications of item response models. *Psychological Assessment, 10*, 345–359.