Survey Research Methods (2022)
Vol. 16, No. 3, pp. 371-385
doi:10.18148/srm/2022.v16i3.7940
European Survey Research Association

An Item Response Theory Analysis and Psychometric Properties of the Czech Version of the Satisfaction with Life Scale

Radka Hanzlová
Department of Sociology
Faculty of Arts
Charles University
Prague, Czech Republic

The original version of the Satisfaction with Life Scale (SWLS) has been successfully tested in many studies, mainly using Classical Test Theory (CTT). Using Item Response Theory (IRT), this study examines psychometric properties of the original SWLS comprising five items, and its abbreviated versions and possible usage of a shorter response scale in a Czech data sample. IRT analysis was used to test the psychometric properties of the original SWLS, comprising five items rated on a 7-point Likert scale, and its results were compared the abbreviated version with four (SWLS-4) and three items (SWLS-3) and was also applied to evaluate the adequacy of the response scale's length. For the analysis, data from a representative sample of the Czech population older than 18 years were used ($N = 1,000$). The results showed that all the three tested versions of the SWLS reached excellent psychometric properties, including unidimensionality, with slight differences between them. Further testing of the response scale's adequacy indicates that shortening the response categories from seven to five will be appropriate. The results of this Czech study show that abbreviated versions of the SWLS can be used interchangeably without impacting its psychometric qualities. The results also confirm that a shorter response scale is appropriate for the application. The findings from measurement invariance indicate that the SWLS allows meaningful comparisons in life satisfaction across gender and age.

Keywords: IRT; Psychometrics; SWLS; Czech Republic; Measurement Invariance

1 Introduction

Subjective well-being is a widely used term, especially in sociology and psychology, which was encountered in the late 1950s when it began to be used as a quality of life indicator (Keyes, Shmotkin, & Ryff, 2002). Subjective well-being comprises three measurable components: positive affect (PA[1]), negative affect (NA), and life satisfaction (LS) (Andrews & Withey, 1976). Particularly, this study focuses on the life satisfaction dimension, which can be defined as a cognitive and global evaluation of the quality of one's life as a whole (Pavot & Diener, 1993). While the information on people's life satisfaction is important for them and society in general, life satisfaction has various social, economic, health, and psychological implications and it is an important basis for policy interventions (Dolan & Metcalfe, 2012; Dolan & White, 2007).

In cross-national research, life satisfaction is usually measured with one simple question. For example, in European Social Survey (ESS), the question is: "All things considered, how satisfied are you with your life as a whole nowadays?" Answers were rated on an 11-point scale from 0 (extremely dissatisfied) to 10 (extremely satisfied). Measurement with only one indicator is very problematic regarding methodology, especially due to low reliability because responses can be significantly influenced by an external factor, such as the previous question, current mood of the respondents, weather, etc. (Huppert et al., 2009; Schwarz, 1987; Schwarz & Clore, 1983; Schwarz & Strack, 1991).

The most dominant and verified multi-item instrument for measuring people's life satisfaction as a whole is the Satisfaction with Life Scale (SWLS Diener, Emmons, Larsen, & Griffin, 1985), which comprises five items rated on a 7-point Likert scale (Table 1). These five items were developed from 48 self-report items related to global life satisfaction alongside positive and negative affects. Factor analysis of all items extracted three factors, such as the three components of subjective well-being. Since the original aim was only

Contact information: Radka Hanzlová, Department of Sociology, Faculty of Arts, Charles University, nám. Jana Palacha 2, 116 38 Prague 1, Czech Republic. E-mail: rahanzlova@gmail.com.

[1] See Appendix A for the meanings of all abbreviations used in the article

to measure life satisfaction, affect items and items from the life satisfaction factor with low factor loadings were eliminated, and ten items were left. Because some of the items had the same meaning, half were eventually dropped, resulting in the final version with five items (Diener et al., 1985). Several other instruments for measuring life satisfaction have been developed to date, however, all are based on the original SWLS with five items. An example of these instruments would be the Temporal Satisfaction With Life Scale (TSWLS Pavot, Diener, & Suh, 1998), or the Riverside Life Satisfaction Scale (RLSS Margolis, Schwitzgebel, Ozer, & Lyubomirsky, 2019).

The SWLS has remained extensively tested (Diener et al., 1985; Emerson, Guhn, & Gadermann, 2017; Pavot, Diener, Colvin, & Sandvik, 1991; Vassar, 2007). Most studies using Classical Test Theory (CTT) have confirmed the good psychometric properties of the SWLS, including validity, internal consistency, and test-retest reliability (Diener et al., 1985; Lucas, Diener, & Suh, 1996; Pavot et al., 1991). However, the results regarding its measurement invariance (Emerson et al., 2017), and especially dimensionality, vary (Clench-Aas, Nes, Dalgard, & Aarø, 2011). Most studies have confirmed a unidimensional structure with one factor, but some studies suggest a two-factor structure comprising "present" (items 1 to 3) and "past" (items 4 and 5) factors (Clench-Aas et al., 2011). Nevertheless, only a few studies have tested the modifications of the SWLS (Kjell & Diener, 2020; Vittersø, Biswas-Diener, & Diener, 2005). Generally, shortening the measurement instruments is desirable, especially in online surveys, because the overall length of the questionnaire significantly affects the response rate (Sandy, Gosling, Schwartz, & Koelkebeck, 2017).

One of the most suitable methods for evaluating and developing scales is Item Response Theory (IRT), which, however, has not been widely applied to the SWLS so far (Nima, Cloninger, Persson, Sikström, & Garcia, 2020; Oishi, 2006; Vittersø et al., 2005). The biggest advantage of IRT is that it focuses on functioning individual items within the scale and can improve the scale's psychometric qualities overall (Ayala, 2009; DeMars, 2010). The application of IRT can also reveal if the number of points on the response scale to individual items is not redundant (O'Connor, Crawford, & Holder, 2015).

The SWLS was tested in many countries worldwide, but not in the Czech Republic, except for one old study testing its psychometric properties on a small sample of university students (Lewis, Shevlin, Smékal, & Dorahy, 1999; Navrátil & Lewis, 2006). These attributes (small sample size and homogenous sample) are problematic due to the relevant validation and generalisability of the results that characterise most studies (Clench-Aas et al., 2011).

This study mainly tests the psychometric properties of the Czech version of the traditional SWLS, comprising five items, and compares the results with their abbreviated versions, with four (SWLS-4) and three items (SWLS-3), using IRT on a representative sample of the Czech population older than 18 years ($N = 1{,}000$). Measurement invariance of the original SWLS across gender and age group is also included in this study.

2 Method

2.1 Sample

The data were collected through personal interviews (combination of methods PAPI 64% and CAPI 36%) in October 2019 by the Public Opinion Research Centre. The data are a representative sample of the Czech population over the age of 18, selected by the quota method according to gender, age, education, region, and size of the residence. Respondents who did not state quota variables or had three or more missing values on the SWLS were excluded from the analysis (a total of six respondents). The final sample comprised 1,000 respondents, of which 483 were men (48%, mean age 48.0 years) and 517 women (52%, mean age 48.2 years). The distribution of the research sample by age groups and education levels can be found in Table 2 (additional tables are given in the Appendix B).

2.2 Measures

The SWLS developed by Diener et al. (1985) originally comprises five items rated on a 7-point Likert scale from 1 (strongly disagree) to 7 (strongly agree) (Table 1). For the analysis, the Czech translation was used.[2] The original wording is:

> Below are five statements that you may agree or disagree with. Using the 1–7 scale below, indicate your agreement with each item by placing the appropriate number on the line preceding that item. Please be open and honest in your responding.
>
> **Item 1** In most ways my life is close to my ideal.
>
> **Item 2** The conditions of my life are excellent.
>
> **Item 3** I am satisfied with my life.
>
> **Item 4** So far, I have gotten the important things I want in life.
>
> **Item 5** If I could live my life over, I would change almost nothing.

[2] Available from http://labs.psychology.illinois.edu/~ediener/SWLS.html

Table 1
Frequency distributions of the SWLS items (N = 1,000)

	Points of Likert scale						
	1 Strongly disagree	2 Disagree	3 Disagree slightly	4 Neither agree nor disagree	5 Agree slightly	6 Agree	7 Strongly agree
Item 1: In most ways, my life is close to my ideal.							
Frequency	62	123	237	183	294	88	13
Percent	6	12	24	18	30	9	1
Cumulating	6	18	42	60	90	99	100
Item 2. The conditions of my life are excellent.							
Frequency	47	121	237	209	294	79	13
Percent	5	12	24	21	29	8	1
Cumulating	5	17	41	62	91	99	100
Item 3: I am satisfied with my life.							
Frequency	25	49	139	170	415	168	34
Percent	3	5	14	17	41	17	3
Cumulating	3	8	22	39	80	97	100
Item 4: So far I have gotten the important things I want in life							
Frequency	62	106	208	210	298	102	14
Percent	6	12	21	21	30	10	1
Cumulating	6	17	38	59	89	99	100
Item 5: If I could live my life over, I would change almost nothing							
Frequency	89	132	215	189	243	107	25
Percent	9	13	21	19	24	11	3
Cumulating	9	22	43	62	86	97	100

Table 2
Distribution of sample by gender, age groups, and education levels (N = 1,000)

	Frequency	Percent
Gender		
Male	483	48
Female	517	52
Age groups		
18–24	94	9
25–34	156	16
35–44	182	18
45–54	195	20
55–64	144	14
65+	229	23
Education levels		
Primary education	110	11
Secondary education without GCSE	347	35
Secondary education with GCSE	340	34
University/Higher education	203	20

2.3 Statistical methods

Data preparation and all preliminary analyses including dimensionality testing were performed using the statistical software SPSS 27 and Mplus 7.2. One missing value for item 4 was imputed by the Expectation-Maximization (EM) method available in SPSS.

IRT analysis was conducted in free statistical software R using package mirt. On data was implemented Samejima's Graded Response Model (GRM) designed for polytomous items (Samejima, 1969). By GRM, for each item, the one discrimination parameter and the number of difficulty or threshold parameters, depending on the response scale's length (the number of response categories minus one), were extracted.

In this study, three versions of the SWLS were tested:

- Original five-item version (SWLS-5)
- Abbreviated four-item version without five item (SWLS-4)
- Abbreviated three-item version without items four and five (SWLS-3)

The original five-item version (SWLS-5) was then compared with the original 7-point response Likert scale and the shorter 5-point response Likert scale. The recoding on a 5-point scale was performed as follows: the first two options at both ends of the scale were merged. This means that categories (1) strongly disagree + (2) disagree and (6) agree + (7) strongly agree will be treated as one category (additional tables are given in Appendix B).

3 Results

3.1 Descriptives

Table 3 shows the descriptive statistics and measures of distribution and normality (skewness and kurtosis) for the total scale and each SWLS item. The average response category for items 1 through 5 of the SWLS-5 on a scale of 1–7 were 3.84, 3.87, 4.54, 3.94, and 3.79. The mean for the SWLS-5 with a 7-point response scale was 19.98, and for a 5-point response scale, it was 15.16.

3.2 Dimensionality

Reliability was tested using Cronbach's alpha and McDonald's omega. The results are identical for both methods and were estimated to be 0.90 for SWLS-5, 0.89 for SWLS-4, and 0.86 for SWLS-3. This supports the results of other studies (Pavot & Diener, 1993). Exploratory factor analysis (EFA) of all three versions confirmed a one-factor solution, with the explained variance being 71% for SWLS-5, 74% for SWLS-4, and 78% for SWLS-3. CFAs with ML estimation were then used to compare the versions with each other. All three versions performed very well. RMSEA was 0.074 for SWLS-5, respectively 0.083 for SWLS-4 and 0.000 for SWLS-3, and CFI was 0.990 for SWLS-5, 0.994 for SWLS-4 and $1,000$ for SWLS-3.

For SWLS-5, a comparison of a one-factor with a two-factor solution was performed. The two-factor model comprising "present" (items 1 to 3) and "past" (items 4 and 5) factors showed a slightly better fit than the one-factor model (RMSEA = 0.064 vs. 0.074; CFI = 0.994 vs. 0.990). This result corresponds to the results of a study by Clench-Aas et al. (2011). For further analysis, a one-factor model was chosen because of the testing of the abbreviated versions of SWLS.

3.3 IRT analysis

The discrimination and threshold parameters from the GRM analysis for all versions appear in Table 4, the item information functions appear in Figure 1 (SWLS-5), 3 (SWLS-4), and 5 (SWLS-3), the test information function with standard error for each version separately in Figure 2 (SWLS-5), 4 (SWLS-4), and 6 (SWLS-3), and finally the test information function for all three versions together in Figure 7. Both item and test information functions show good functioning, especially between −2.0 and 2.5 of the latent trait continuum. On the lower and upper ends of the latent trait continuum, there is a visible diminishing of the amount of information.

The results for the SWLS-5 (Table 4 and Fig. 1) show that the scale has very good functioning in terms of both item and model fit. However, item 5 indicated a worse fit to the scale compared to the other four items. Item 5 has a value of discrimination parameter 2.39, which is lower than the values for the other items (which were all 2.80 or above), but still very high (should be greater than 1). The information function analysis shows that item 1 has the highest discrimination estimates and provides more information than the remaining items, especially item 5, which line is flatter and located lower indicating that this item provides less information and contributes little to the scale. Generally, the best functioning in providing overall information is between the low ($\theta = -2.00$) and high ($\theta = 2.50$) values of the latent trait. At $\theta = 0.0$, we obtained reliable information at about 3.00 from item 1, at about 2.50 from item 2, at about 2.00 from item 3, at about 2.30 from item 4, and at about 1.70 from item 5. The difficulty parameters for SWLS-5 were between −2.32 and 2.66. Here, item 3 had the lowest difficulty parameter on response 1 (−2.32), and item 4 had the highest estimated difficulty parameter on response 7 (2.66). The results also show that the differences between categories around difficulty parameters are unequal across items. This means, for example, that for item 3, a category of 7 (strongly agree) is 2.21, while for item 4 is 2.66. Moreover, the differences in difficulty varied within each item (i.e. the distance between categories for each item). For example, for item 3, the difference between thresholds b1 and b2 is −0.65, between b2 and b3 is −0.77, between b3 and b4 is −0.57, between b4

Table 3
Descriptive statistics and testing normality of SWLS-5

	7-point response scale				5-point response scale			
	Mean	Std.Dev.	Skewness	Kurtosis	Mean	Std.Dev.	Skewness	Kurtosis
Item 1	3.84	1.42	−0.20	−0.73	2.89	1.29	−0.03	−1.21
Item 2	3.87	1.36	−0.19	−0.62	2.91	1.25	−0.05	−1.13
Item 3	4.54	1.29	−0.65	0.20	3.53	1.73	−0.65	−0.47
Item 4	3.94	1.42	−0.30	−0.61	2.99	1.28	−0.12	−1.13
Item 5	3.79	1.55	−0.09	−0.81	2.85	1.36	0.06	−1.26
SWLS total	19.98[a]	5.95	−0.25	−0.36	15.16[b]	5.33	−0.11	−0.87

[a] Mean on a range of 5–35 [b] Mean on a range of 5–25

and b5 is −1.31 and between b5 and b6 is −1.23 (Table 4). In Figure 2, there is the test information function and the standard error. This means that the SWLS-5 has good reliability and a small standard error in this range. The highest level of test information is located around −1.0 to −0.3 (θ), thus indicating that this score has the smallest standard errors and provides the most information of the scale. However, around below −3.00 of θ and above 2.80 of θ the standard error increases sharply, and the information provided by the scale is negligible.

The SWLS-4 comprises four items (items 1 to 4) without "problematic" item 5. Generally, SWLS-4 outperformed SWLS-5, and the values of discriminant parameters for all items were more acceptable (Table 4). The highest value of the discrimination parameter is still for item 1; however, the value of the discrimination parameter for item 2 increased to almost the same values as for item 1. On the other hand, the discriminant parameter slightly decreased for item 4, which is lower than the other three items (2.59 vs. 2.80 and more). This result is confirmed by the item information functions in Figure 3, where item 4 is located lower and is flatter, especially between values −2.00 and 0.00 of the latent trait. These results indicate that item 4 functions slightly worse and provides less information than the other three items. Generally, as with the SWLS-5, this scale performs the best between the low (θ = −2.00) and high (θ = 2.50) values of the latent trait. At θ = 0.00, we obtained information at about 3.00 from item 1, at about 2.90 from item 2, at about 1.70 from item 3, and at about 2.00 from item 4. The difficulty parameters for SWLS-4 ranged from −2.33 (item 3) to 2.72 (item 4). Similar to SWLS-5, the differences between categories around difficulty parameters are unequal across items, and they vary within each item. The graph (Fig. 4) with test information and standard error shows very similar results to those of SWLS-5 (Fig. 2).

The third tested version is SWLS-3 comprising items 1 to 3. Generally, this version has similar features as SWLS-4 and SWLS-5. The values of the discrimination parameters were 2.70 and higher for all three items (Table 4). The graph with item information functions (Fig. 5) shows the best solution compared to SWLS-5 and SWLS-4. Although item 3 is again located a little lower and flatter, it provides unique information to the scale at the left end of the latent trait continuum.

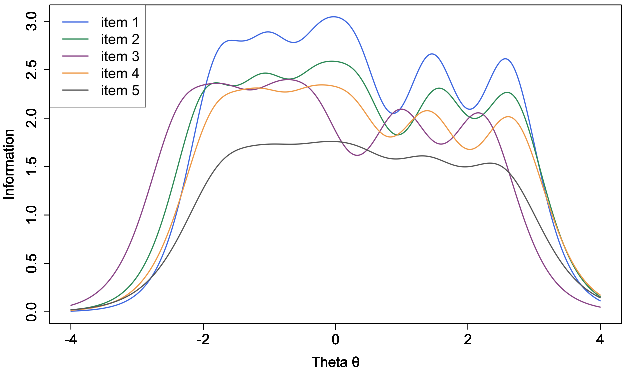

Figure 1. Item information functions for SWLS-5

The test information functions for all three versions appear in Figure 7. The results indicate that the assessment of the latent variable is not significantly affected by the elimination of items 5 and 4. The test information functions are very similar both regarding the shape and the value of the overall information provided. However, the differences between SWLS-5 and SWLS-3 are bigger. From the results of the analysis, I would recommend SWLS-4.

The category characteristic curves for the original SWLS-5 are provided in Figure 8. These curves show how well or badly each response category functions alongside the transition from one category to the next. This may reveal that response categories are poorly used by respondents or are redundant. The curves for all five items indicate that there are too many response categories. The extreme categories, options 1 and 7 were rarely used by respondents for all items (see Table 1). The neutral category (option 4 on a 7-point response scale) is redundantly used by respondents for all items because the distance between the thresholds from the adjacent category 3 to 4 and 4 to 5 is very small. More-

Table 4
Discrimination and threshold parameters for SWLS-5, SWLS-4, and SWLS-3

	Discrimination parameter	Difficulty parameters for each threshold					
	a	b1	b2	b3	b4	b5	b6
SWLS-5							
Item 1	3.19	−1.76	−1.01	−0.23	0.28	1.45	2.60
Item 2	2.95	−1.96	−1.09	−0.27	0.33	1.55	2.65
Item 3	2.82	−2.32	−1.67	−0.90	−0.33	0.98	2.21
Item 4	2.80	−1.81	−1.13	−0.38	0.24	1.40	2.66
Item 5	2.39	−1.68	−0.97	−0.23	0.36	1.37	2.49
SWLS-4							
Item 1	3.21	−1.76	−1.01	−0.23	0.28	1.45	2.60
Item 2	3.18	−1.92	−1.07	−0.27	0.32	1.52	2.59
Item 3	2.80	−2.33	−1.67	−0.90	−0.34	0.99	2.20
Item 4	2.59	−1.86	−1.15	−0.38	0.25	1.43	2.72
SWLS-3							
Item 1	2.98	−1.80	−1.03	−0.23	0.29	1.48	2.65
Item 2	3.59	−1.88	−1.05	−0.26	0.31	1.48	2.52
Item 3	2.70	−2.37	−1.69	−0.91	−0.34	1.00	2.22

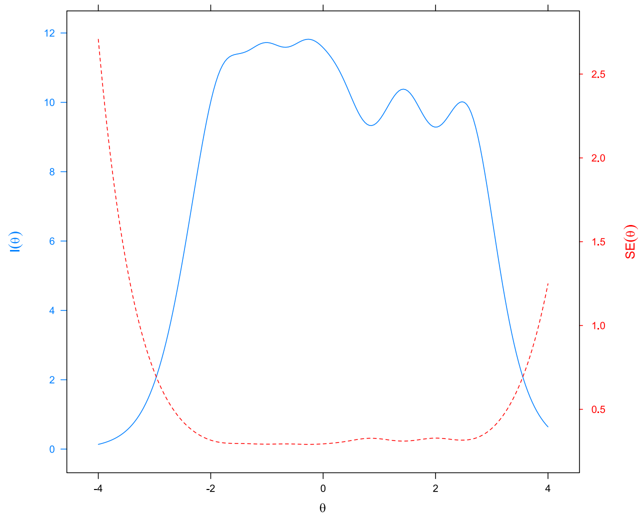

Figure 2. Test information function and standard error for SWLS-5

over, there is generally insufficient differentiation between response options 2, 3, and 4 in the centre and left (lower) part of the latent trait continuum between values 0.00 and −2.00. This problem relates particularly to items 3, 4, and 5. Based on these results, I believe that reducing the range of response categories would be appropriate. In this study, the shortening of the response categories from seven to five, which can be done by merging the categories, was tested (Fig. 9). At the same time, I also think it would be useful to explore a 6-point response scale without the neutral category in further research.

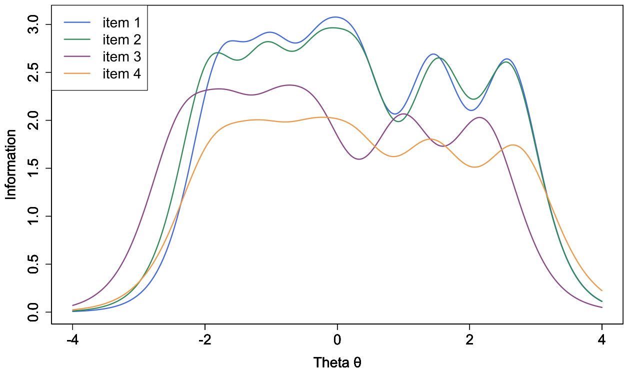

Figure 3. Item information functions for SWLS-4

The comparison of the reliability scores of the scale with a 7-point and a 5-point response scale for all the three versions (Table 5) also suggests that reducing the response categories will not cause a significant reduction since the values are almost the same.

3.4 Testing validity

The validity of the SWLS was tested by correlation with a one-item question on overall life satisfaction. The wording of this question is: “All things considered, how satisfied are you with your life?”. Answers were rated on a 10-point scale from 1 (at least) to 10 (at most). The calculated corre-

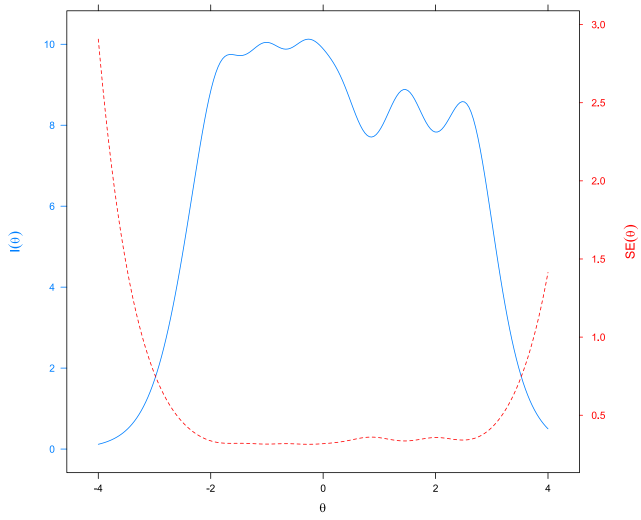

Figure 4. Test information function and standard error for SWLS-4

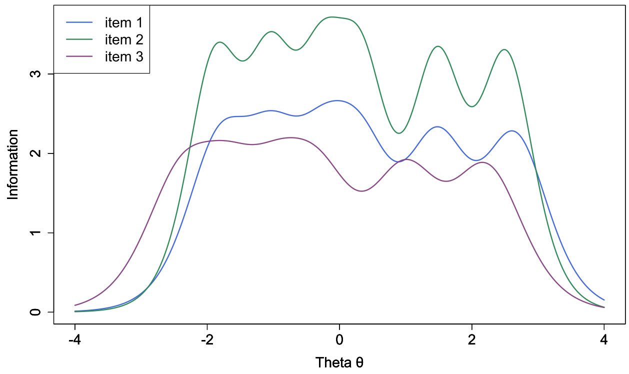

Figure 5. Item information functions for SWLS-3

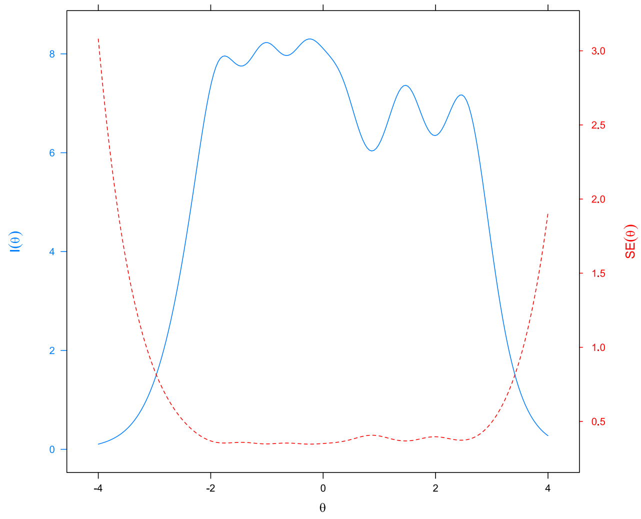

Figure 6. Test information function and standard error for SWLS-3

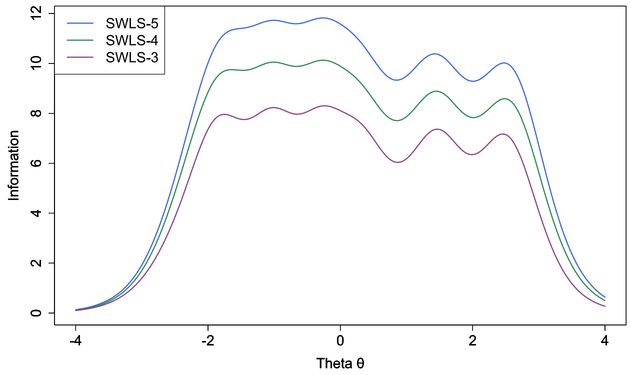

Figure 7. Test information functions for all three versions

Table 5

The comparison of the reliability scores

	7-point response scale	5-point response scale
SWLS-5	0.91	0.90
SWLS-4	0.90	0.88
SWLS-3	0.88	0.86

lations are high for all tested versions, with the highest in the case of SWLS-3 (0.569), then SWLS-4 (0.561), and SWLS-5 (0.557).

3.5 Measurement invariance across gender and age groups

Measurement invariance of the SWLS across gender and age groups was tested by multiple-group CFA (MGCFA), the most widely used method for testing invariance (Davidov, Meuleman, Cieciuch, Schmidt, & Billiet, 2014; Kim, Cao, Wang, & Nguyen, 2017). These analyses were conducted with Mplus 7.2. The parameter estimates were obtained using the robust maximum likelihood robust (MLR). To evaluate model fit, chi-square (χ^2), RMSEA (Root Mean Square of Error Approximation), and CFI (Comparative Fit Index) were used. Since the χ^2 is highly sensitive to sample size and almost always significant in large samples (Kline, 2005), it is used for descriptives purposes only. Cut-offs for the goodness of fit were 0.90 for CFI and 0.08 for RMSEA indication acceptable fit and 0.95 for CFI and 0.06 for RMSEA indicating good fit (Hu & Bentler, 1999).

Measurement invariance by MGCFA is based on comparisons models with increasing restrictions. The baseline model is a configural (unconstrained) model that assumes the same factor structure across different groups. The second is the metric model, which is nested in the configural model and requires the factor loadings to be equivalent across groups. The highest is the scalar model that assumes both factor loadings and intercepts to be equal across groups. The scalar

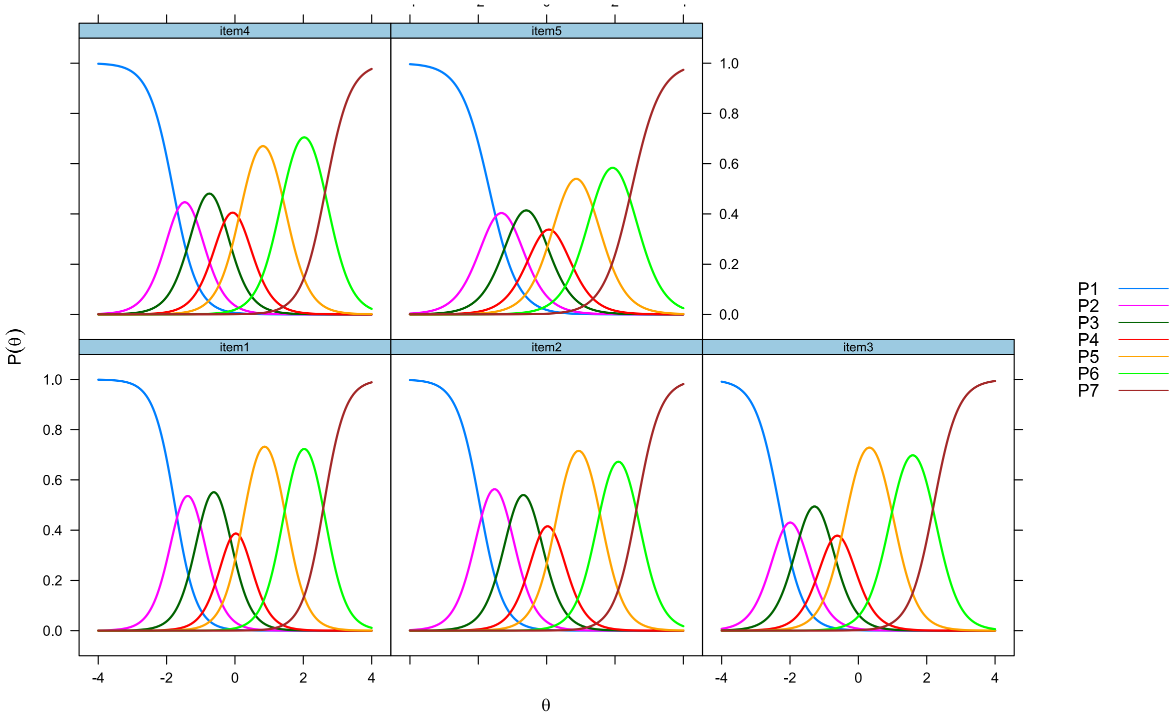

Figure 8. Category response curves for seven-category SWLS-5

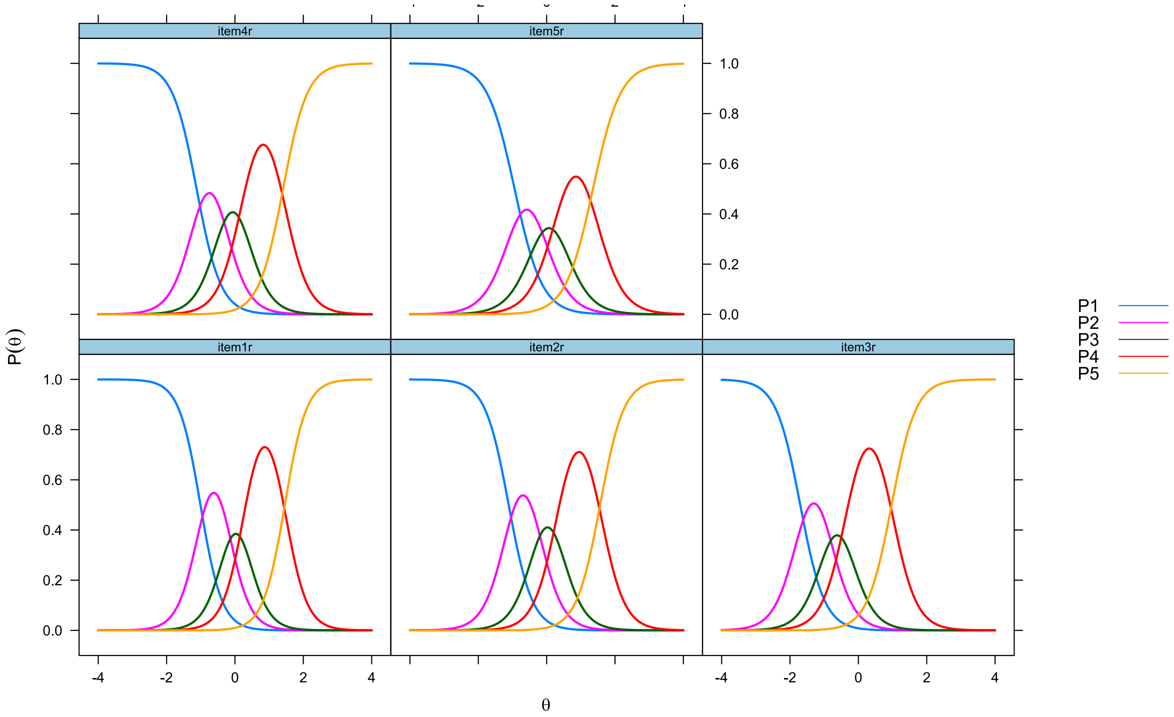

Figure 9. Category response curves for five-category SWLS-5

model, respectively scalar invariance is considered necessary for comparison latent means across groups (Meredith, 1993).

To assess whether a certain level of invariance is achieved, global fit indices, respectively the change in model fit (Δ) are compared between more and less constrained models. If the Δ is about 0.01 and 0.015 or less in CFA and RMSEA, the more constrained model may be accepted (Chen, 2007). When continuous data for large numbers of groups and large sample size are analysed, Rutkowski and Svetina (2014) suggested the more liberal cut-off values of 0.02 for ΔCFI and 0.03 for ΔRMSEA from configural to metric model, and of 0.01 for both ΔCFI and ΔRMSEA from metric to scalar model. According to Hu and Bentler (1999) and Hirschfeld and Brachel (2014), the best indicator is ΔCFI (less than 0.01).

As shown in Table 6, the highest level, scalar invariance were achieved across gender, thus allowing latent means comparison. In terms of age groups, results revealed both configural and metric invariance. The ΔCFI between the two models was 0.002 and ΔRMSEA 0.016, which is slightly higher but still meets the more liberal criterion. Moreover, the more important is ΔCFI. However, full scalar invariance was not supported, since both ΔCFI and ΔRMSEA was much higher, which indicate a significant decrement (ΔCFI = −0.012, ΔRMSEA = 0.030). Modification indices revealed one item (item 4) to be operating differently across age groups. Therefore, the partial scalar invariance was tested by freely estimating the intercepts of item 4. After releasing the intercepts of item 4 in two groups, the partial scalar invariance was achieved (ΔCFI = −0.003, ΔRMSEA = 0.011). Based on these results, the latent means across age groups can be compared.

4 Discussion

The SWLS is one of the most widely used methods for measuring life satisfaction that is one of the components of subjective well-being (Andrews & Withey, 1976). The original SWLS, comprising five items rated on a 7-point Likert scale has been successfully tested in many countries but not in the Czech Republic. This study examined the psychometric properties of the original Czech version of SWLS in a representative sample of the Czech population ($N = 1,000$) and compared the results with their abbreviated versions with four (SWLS-4) and three items (SWLS-3). A partial goal of this study was also to reveal how the response options function and whether the number of points on the response scale was appropriate. Another goal was also to find out whether SWLS measures comparably between different groups in terms of gender and age.

4.1 Dimensionality

The results for all three tested versions support the unidimensionality of SWLS. EFA extracted a one-factor solution with 71% of the variance explained by this single factor for SWLS-5, 74% for SWLS-4, and 78% for SWLS-3. Each version was also tested by CFA, which confirmed their excellent fit based on values of RMSEA and CFI.

Since some studies support a two-factor solution with the "present" (items 1 to 3) and "past" factor (items 4 and 5), both solutions were compared for SWLS-5. The results showed a slightly better fit for the two-factor model than for the one-factor model (RMSEA = 0.064 vs. 0.074; CFI = 0.994 vs. 0.990). The differences are small, and this finding agrees with several previous studies (Clench-Aas et al., 2011).

4.2 Comparison of the three versions of SWLS

Generally, testing psychometric properties using CTT showed excellent qualities of the original five-item SWLS alongside abbreviated versions with four (SWLS-4) and three items (SWLS-3). Cronbach's alpha and McDonald's omega were 0.90 for SWLS-5, 0.89 for SWLS-4, and 0.86 for SWLS-3. According to these findings, presumably, shortening the scale by eliminating one (item 5) or two (items 4 and 5) items will not affect its psychometric qualities. This assumption was tested using IRT, which is an efficient method for scale development and differential item functioning.

In this study, GRM was applied because the data were ordinal (rated on a 7- or 5-point Likert scale). The results showed that all the five items function well in SWLS-5 since their discrimination parameters are high. Simultaneously, it was revealed that item 5 differs slightly from the other items, as the value of its discriminant parameter is lower, and to the scale contributes the least.

Removing item 5 from SWLS-5 increased the value of the discriminant parameter for item 2 for SWLS-4 and decreased slightly for item 4, which also functions slightly worse and provides less information. These findings (Fig. 3) suggest that item 4 could also be excluded from the scale.

The final abbreviated tested version contained three items (SWLS-3), which is generally the minimum number for the scale. The value of the discrimination parameter decreased for item 1 and conversely increase for items 2 and 3. From the item information functions (Fig. 5) is clear that each item functions sufficiently within the scale.

Observably, from the comparison of all the three versions (Fig. 7), the overall test information provided by SWLS-5, SWLS-4, and SWLS-3 lacks significant difference. This conclusion confirmed the results of the CTT; therefore, abbreviated versions of SWLS with four or three items can be used without affecting its psychometric qualities as well as the reliability and validity of the instrument. According to all the results, I would prefer SWLS-4.

Table 6
Fit indices and difference statistics for measurement invariance models across gender and age

				RMSEA			
					90% C.I.		
	Chi-square	CFI	Δ CFI	Est.	Lower	Upper	Δ RMSEA
Gender							
Configural	28.731	0.990	-	0.061	0.036	0.088	-
Metric	34.287	0.989	0.001	0.054	0.031	0.077	0.007
Scalar	37.591	0.989	0.000	0.047	0.025	0.068	0.007
Age groups							
Configural	35.167	0.997	-	0.032	0.000	0.070	-
Metric	52.229	0.999	0.002	0.016	0.000	0.053	0.016
Scalar	94.469	0.987	0.012	0.046	0.016	0.068	0.030
Partial scalar[a]	76.266	0.996	0.003	0.027	0.000	0.055	0.011

[a] Intercept of item 4 freely estimated in g1 and g6

4.3 Analysis of the response scale

The original SWLS items are rated on a 7-point Likert scale from 1 (strongly disagree) to 7 (strongly agree). This study's results indicate that the number of categories on the response scale is excessive and unnecessary. Option 1 and 7 were rarely used, the neutral response category 4 was redundant, and generally, there was insufficient differentiation by participants between response categories 2, 3, and 4. The problem with the neutral category can be caused by the original Czech translation, which, in my opinion, is inaccurate. However, I suppose that reducing response categories from seven to five would be beneficial since respondents apparently do not sufficiently distinguish between the current seven response categories (Figs. 8, 9). Alternatively, it would be interesting to test a version with six categories of the response scale without a neutral category. The shortening of the response scale is also suggested by O'Connor et al. (2015), who tested the four-item Subjective Happiness Scale (SHS), which is rated on a 7-point scale from 1 (not at all) to 7 (a great deal).

Another possibility would be to maintain the seven response categories but work with labelling the individual response categories and displaying them to the respondents. In this study, all seven points on the response scale had labels, and this scale was presented to the respondents. I would suggest labelling only the first and last response categories.

4.4 Measurement invariance

Measurement invariance of the SWLS across gender and age groups tested by MGCFA revealed that is possible to compare latent means across gender and groups since the scalar invariance for gender, respectively partial scalar invariance for age groups was achieved. The latent means for the male and the youngest age group were fixed to zero, and the latent means in the remaining groups were freely estimated. The analysis of the latent means demonstrated that females reported slightly higher life satisfaction than males. In terms of age groups, the youngest people show the highest level of life satisfaction, while the lowest level of life satisfaction has occurred in people in the oldest age group (detailed tables in Appendix B).

5 Strengths and limitations of this study

This study has three major advantages: (1) the large representative research sample for the population of the Czech Republic over the age of 18 selected by the quota method according to gender, age, education, region, and size residence, (2) the high quality of data collection, and (3) generally the application of IRT on SWLS because most of the studies testing this scale used CTT. Especially in the Czech Republic, this is the first study testing the SWLS and, moreover, using IRT, which has remained unused in this area.

The main shortcoming is related to the translation of response categories. The middle category on the response scale (point 4) should be a neutral answer ("Neither agree nor disagree"). However, according to the original Czech translation, it could be understood by the respondents as "I don't know". Subsequently, it would certainly be appropriate to use an accurate translation in testing SWLS.

Another limitation is the method of testing the response scale's length, respectively merging of categories. The better and more appropriate solution should be to test in one research sample via split-ballot test the different response scale and then compare the results.

This study deals only with life satisfaction as one of the three measurable components of subjective well-being. If the goal is to measure subjective well-being as a whole, it is necessary to focus on all its components and test them together using different scales measuring positive and negative affects, and life satisfaction.

6 Conclusion

The overall results indicate that a one-factor latent structure of the original SWLS performed efficiently in the Czech data alongside its abbreviated version with four or three items, therefore could be used interchangeably without affecting its psychometric properties. This study also suggests that reducing the number of response categories from seven to five would be appropriate. The SWLS is a valid instrument for the comparison of latent means across gender and age.

Acknowledgement

This study was supported by the grant SVV 2020–2022 (260556) realized at the Charles University, Faculty of Arts.

References

Andrews, F., & Withey, S. (1976). *Social indicators of well-being: Americans' perceptions of life quality*. doi:10.1007/978-1-4684-2253-5

Ayala, R. (2009). *The theory and practice of item response theory*. Guilford Press.

Chen, F. (2007). Sensitivity of goodness of fit indexes to lack of measurement invariance. *Structural Equation Modeling: A Multidisciplinary Journal*, *14*(3), 464–504. doi:10.1080/10705510701301834

Clench-Aas, J., Nes, R., Dalgard, O., & Aarø, L. (2011). Dimensionality and measurement invariance in the satisfaction with life scale in Norway. *Quality of Life Research*, *20*(8), 1307–1317. doi:10.1007/s11136-011-9859-x

Davidov, E., Meuleman, B., Cieciuch, J., Schmidt, P., & Billiet, J. (2014). Measurement equivalence in cross-national research. *Annual Review of Sociology*, *40*(1), 55–75. doi:10.1146/annurev-soc-071913-043137

DeMars, C. (2010). *Item response theory*. Oxford University Press.

Diener, E., Emmons, R., Larsen, R., & Griffin, S. (1985). The satisfaction with life scale. *Journal of Personality Assessment*, *49*(1), 71–75. doi:10.1207/s15327752jpa4901_13

Dolan, P., & Metcalfe, R. (2012). Measuring subjective well-being: Recommendations on measures for use by national governments. *Journal of Social Policy*, *41*(2), 409–427. doi:10.1017/S0047279411000833

Dolan, P., & White, M. (2007). How can measures of subjective well-being be used to inform public policy? *Perspectives on Psychological Science*, *2*(1), 71–85. doi:10.1111/j.1745-6916.2007.00030.x

Emerson, S., Guhn, M., & Gadermann, A. (2017). Measurement invariance of the satisfaction with life scale: Reviewing three decades of research. *Quality of Life Research*, *26*(9), 2251–2264. doi:10.1007/s11136-017-1552-2

Hirschfeld, G., & Brachel, R. (2014). Multiple-group confirmatory factor analysis in R: A tutorial in measurement invariance with continuous and ordinal indicators. *Practical Assessment, Research & Evaluation*, *19*, 1–12. doi:10.7275/QAZY-2946

Hu, L., & Bentler, P. (1999). Cutoff criteria for fit indexes in covariance structure analysis: Conventional criteria versus new alternatives. *Structural Equation Modeling: A Multidisciplinary Journal*, *6*(1), 1–55. doi:10.1080/10705519909540118

Huppert, F., Marks, N., Clark, A., Siegrist, J., Stutzer, A., Vittersø, J., & Wahrendorf, M. (2009). Measuring well-being across Europe: Description of the ESS well-being module and preliminary findings. *Social Indicators Research*, *91*(3), 301–315. doi:10.1007/s11205-008-9346-0

Keyes, C., Shmotkin, D., & Ryff, C. (2002). Optimizing well-being: The empirical encounter of two traditions. *Journal of Personality and Social Psychology*, *82*(6), 1007–1022. doi:10.1037/0022-3514.82.6.1007

Kim, E., Cao, C., Wang, Y., & Nguyen, D. (2017). Measurement invariance testing with many groups: A comparison of five approaches. *Structural Equation Modeling: A Multidisciplinary Journal*, *24*(4), 524–544. doi:10.1080/10705511.2017.1304822

Kjell, O., & Diener, E. (2020). Abbreviated three-item versions of the satisfaction with life scale and the harmony in life scale yield as strong psychometric properties as the original scales. *Journal of Personality Assessment*, *103*(2), 183–194. doi:10.1080/00223891.2020.1737093

Kline, R. (2005). *Principles and practices of structural equation modeling* (2nd). New York, NY: Guilford Press.

Lewis, C., Shevlin, M., Smékal, V., & Dorahy, M. (1999). Factor structure and reliability of a Czech translation of the satisfaction with life scale among Czech university students. *Studia psychologica*, *41*(3), 239–244.

Lucas, R., Diener, E., & Suh, E. (1996). Discriminant validity of well-being measures. *Journal of Personality and Social Psychology*, *71*(3), 616–628. doi:10.1037/0022-3514.71.3.616

Margolis, S., Schwitzgebel, E., Ozer, D., & Lyubomirsky, S. (2019). A new measure of life satisfaction: The Riverside life satisfaction scale. *Journal of Personality Assessment*, *101*(6), 621–630. doi:10.1080/00223891.2018.1464457

Meredith, W. (1993). Measurement invariance, factor analysis and factorial invariance. *Psychometrika*, *58*(4), 525–543. doi:10.1007/BF02294825

Navrátil, M., & Lewis, C. (2006). Temporal stability of the Czech translation of the satisfaction with life scale: Test-retest data over one week. *Psychological Reports*, *98*(3), 918–920. doi:10.2466/pr0.98.3.918-920

Nima, A., Cloninger, K., Persson, B., Sikström, S., & Garcia, D. (2020). Validation of subjective well-being measures using item response theory. *Frontiers in Psychology*, *10*, 3036. doi:10.3389/fpsyg.2019.03036

O'Connor, B., Crawford, M., & Holder, M. (2015). An item response theory analysis of the subjective happiness scale. *Social Indicators Research*, *124*(1), 249–258. doi:10.1007/s11205-014-0773-9

Oishi, S. (2006). The concept of life satisfaction across cultures: An IRT analysis. *Journal of Research in Personality*, *40*(4), 411–423. doi:10.1016/j.jrp.2005.02.002

Pavot, W., & Diener, E. (1993). Review of the satisfaction with life scale. *Psychological Assessment*, *5*(2), 164–172. doi:10.1037/1040-3590.5.2.164

Pavot, W., Diener, E., Colvin, C., & Sandvik, E. (1991). Further validation of the satisfaction with life scale: Evidence for the cross-method convergence of well-being measures. *Journal of Personality Assessment*, *57*(1), 149–161. doi:10.1207/s15327752jpa5701_17

Pavot, W., Diener, E., & Suh, E. (1998). The temporal satisfaction with life scale. *Journal of Personality Assessment*, *70*(2), 340–354. doi:10.1207/s15327752jpa7002_11

Rutkowski, L., & Svetina, D. (2014). Assessing the hypothesis of measurement invariance in the context of large-scale international surveys. *Educational and Psychological Measurement*, *74*(1), 31–57. doi:10.1177/0013164413498257

Samejima, F. (1969). Estimation of latent ability using a response pattern of graded scores. *Psychometrika*, *34*(4), 1–97.

Sandy, C., Gosling, S., Schwartz, S., & Koelkebeck, T. (2017). The development and validation of brief and ultrabrief measures of values. *Journal of Personality Assessment*, *99*(5), 545–555. doi:10.1080/00223891.2016.1231115

Schwarz, N. (1987). Stimmung als Information Untersuchungen zum Einfluß von Stimmungen auf die Bewertung des eigenen Lebens. doi:10.1007/978-3-642-72885-3

Schwarz, N., & Clore, G. (1983). Mood, misattribution, and judgments of well-being: Informative and directive functions of affective states. *Journal of Personality and Social Psychology*, *45*(3), 513–523. doi:10.1037/0022-3514.45.3.513

Schwarz, N., & Strack, F. (1991). Evaluating one's life: A judgment model of subjective wellbeing. In F. Strack, M. Argyle, & N. Schwarz (Eds.), *Subjective well-being: An interdisciplinary perspective* (pp. 27–48). Oxford: Pergamom Press.

Vassar, M. (2007). A note on the score reliability for the satisfaction with life scale: An RG study. *Social Indicators Research*, *86*(1), 47–57. doi:10.1007/s11205-007-9113-7

Vittersø, J., Biswas-Diener, R., & Diener, E. (2005). The divergent meanings of life satisfaction: Item response modeling of the satisfaction with life scale in Greenland and Norway. *Social Indicators Research*, *74*(2), 327–348. doi:10.1007/s11205-004-4644-7

(*Appendix tables follow on next page*)

Appendix A
Abbreviations

CAPI Computer Assisted Personal Interviewing

CFA Confirmatory Factor Analysis

CFI Confirmatory Fit Index

CTT Classical Test Theory

EFA Exploratory Factor Analysis

EM Expectation-Maximization

ESS European Social Survey

GCSE General Certificate of Secondary Education

GRM Graded Response Model

IRT Item Response Theory

LS Life satisfaction

mirt Multidimensional Item Response Theory

NA Negative affect

PA Positive affect

PAPI Pen-and-Paper Personal Interview

RLSS Riverside Life Satisfaction Scale

RMSEA Root Mean Square Error of Approximation

SPSS Statistical Package for Social Sciences

SWLS Satisfaction with Life Scale

TSWLS Temporal Satisfaction with Life Scale

Appendix B
Tables

Table B1
Frequency distributions of the SWLS items ($N = 1,000$) on a 5-point response Likert scale

	Points of Likert scale				
	1 Disagree	2 Disagree slightly	3 Neither agree nor disagree	4 Agree slightly	5 Agree
Item 1					
Frequency	185	237	183	294	101
Percent	19	24	18	29	10
Cumulating	19	42	61	90	100
Item 2					
Frequency	168	237	209	294	92
Percent	18	23	21	29	9
Cumulating	18	42	62	91	100
Item 3					
Frequency	74	139	170	415	202
Percent	7	14	17	42	20
Cumulating	7	21	38	80	100
Item 4					
Frequency	168	208	210	298	116
Percent	17	21	21	30	12
Cumulating	17	38	59	88	100
Item 5					
Frequency	221	215	189	243	132
Percent	22	22	19	24	13
Cumulating	22	44	63	87	100

Item 1: In most ways my life is close to my ideal.; Item 2: The conditions of my life are excellent.; Item 3: I am satisfied with my life.; Item 4: So far I have gotten the important things I want in life.; Item 5: If I could live my life over, I would change almost nothing.

Table B2
Mean distribution of the SWLS items by gender and age (7-point response scale)

	Gender		Age group (years)					
	Males	Females	18–24	25–34	35–44	45–54	55–64	65+
Item 1	3.85	3.83	4.26	3.83	3.82	3.76	3.87	3.75
Item 2	3.86	3.89	4.35	3.95	3.84	3.86	3.80	3.71
Item 3	4.51	4.57	4.96	4.63	4.42	4.45	4.51	4.51
Item 4	3.94	3.93	4.02	3.82	3.93	3.98	3.84	4.01
Item 5	3.78	3.79	4.29	3.89	3.74	3.83	3.55	3.66
N	483	517	94	156	182	195	144	229

Table B3

Mean distribution of the SWLS items by gender and age (5-point response scale)

	Gender		Age group (years)					
	Males	Females	18–24	25–34	35–44	45–54	55–64	65+
Item 1	2.91	2.87	3.27	2.87	2.88	2.83	2.91	2.80
Item 2	2.90	2.91	3.34	2.96	2.88	2.92	2.83	2.75
Item 3	3.51	3.56	3.91	3.60	3.44	3.45	3.51	3.48
Item 4	3.00	2.97	3.05	2.88	2.99	3.05	2.89	3.03
Item 5	2.86	2.84	3.31	2.96	2.81	2.88	2.62	2.73
N	483	517	94	156	182	195	144	229

Table B4

Latent mean comparison across gender and age

	Latent mean	Raw mean[b]
Gender		
Male (M)[a]	0.000	19.94
Female (F)	0.014	20.01
Age groups		
18–24 (g1)[a]	0.000	21.87
25–34 (g2)	-0.433	20.12
35–44 (g3)	-0.517	19.74
45–54 (g4)	-0.501	19.88
55–64 (g5)	-0.548	19.56
65+ (g6)	-0.584	19.63

[a] Male and g1 are reference groups (latent mean fixed to 0) [b] Raw mean on a range of 5–35