Computers in Human Behavior 71 (2017) 172–180

Contents lists available at ScienceDirect

Computers in Human Behavior

journal homepage: www.elsevier.com/locate/comphumbeh

Full length article

Survey method matters: Online/offline questionnaires and face-to-face or telephone interviews differ

XiaoChi Zhang [a, *], Lars Kuchinke [b], Marcella L. Woud [a], Julia Velten [a], Jürgen Margraf [a]

[a] Mental Health Research & Treatment Center of Ruhr-Universität Bochum, Germany
[b] Experimental Psychology, Ruhr-Universität Bochum, Germany

ARTICLE INFO

Article history:
Received 21 December 2015
Received in revised form
11 May 2016
Accepted 2 February 2017
Available online 2 February 2017

Keywords:
Survey method
Mode effect
ANCOVA
Measurement invariance

ABSTRACT

Self-report inventories enable efficient assessment of mental attributes in large representative surveys. However, an inventory can be administered in several ways whose equivalence is largely untested. In the present study, we administered thirteen psychological questionnaires assessing positive and negative aspects of mental health. The questionnaires were administered by four different data collection methods: face-to-face interview, telephone interview, online questionnaire, and offline questionnaire. We found that twelve of the questionnaires differed in survey methods. Although, some studies showed that social desirability tends to be highest for telephone survey and lowest for web survey. Furthermore, the effects of social desirability should be the same for the online and offline samples. However, there were no statistically significant differences between the face-to-face and telephone samples for the anxiety scale, the stress scale, and the tradition scale. We also found that for eight scales, the online sample was statistically different from the offline sample in the respondent answers. Moreover, the survey method effects were only moderated by age. Finally, measurement invariance across the four survey methods was tested for each self-report measure. There was full strong measurement invariance established for nine of thirteen scales and partial strong measurement invariance for the remaining four scales across the four survey methods. These findings indicated that measurement invariance was affected by different survey methods.

1. Introduction

Self-report measures are widely used to study and assess personality characteristics and various aspects of health and behavior. More recently, however, traditional paper pencil surveys have been challenged by computer supported surveys. Since the rapid expanding of the internet, online surveys became more and more popular (Griffiths, Lewis, Ortiz de Gortari, & Kuss, 2014). There are a number of advantages for this approach: simplified work for the interviewers, fast data processing, and low costs (Beebe, Mika, Harrison, Anderson, & Fulkerson, 1997; Rosenfeld, Booth-Kewley, & Edwards, 1993). Not surprisingly, however research found that different survey methods can lead to different responses although the same questions were asked (Kiesler & Sproull, 1986). This is called "mode effect", and a number of such effects have been identified. Social desirability is one of the most studied mode effects. The results of these studies, however, have been inconsistent. To illustrate, many studies examined data quality and the effects of social desirability when using different survey methods. In some studies, computer surveys yielded similar results as paper and pencil surveys, e.g., on attitude questionnaires (Booth-Kewley, Edwards, & Rosenfeld, 1992) or for personally sensitive questions (Knapp & Kirk, 2003). In other studies, however, different results were found when using different survey methods, e.g., on satisfaction-dissatisfaction questions (Dillman et al., 2008) or on questions about consumption frequency and preferences related to wine (Szolnoki & Hoffmann, 2013). Furthermore, response biases for telephone interviews and internet questionnaires caused by social desirability have been reported (Chang & Krosnick, 2009). Here, more social desirability was manifested for telephone compared to Internet surveys, respectively. Some studies also showed that biases related to social desirability tended to be highest for telephone surveys and lowest for web surveys (Holbrook, Green, & Krosnick, 2003; Kreuter, Presser, &

* Corresponding author.
E-mail addresses: xiaochi.zhang@rub.de (X. Zhang), lars.kuchinke@rub.de (L. Kuchinke), Marcella.woud@rub.de (M.L. Woud), Julia.velten@rub.de (J. Velten), juergen.margraf@ruhr-uni-bochum.de (J. Margraf).

http://dx.doi.org/10.1016/j.chb.2017.02.006
0747-5632/

Tourangeau, 2008). More recently, however, a meta-analysis concluded that social desirability was the same in offline, online and paper surveys (Dodou & de Winter 2014). Hence, this shows that the scientific state concerning the effects of social desirability is still inconsistent, and more research is needed to advance our understanding of its effects and underlying mechanisms.

A possible explanation of these inconsistencies could be the lack of large representative population samples with sufficient power to detect relevant effects. Moreover, in-depth investigations of measurement invariance across different assessment modes are sparse. Some studies examined the measurement invariance when using web surveys compared to paper and pencil methods (Davidov & Depner, 2011; Fang, Wen, & Prybutok, 2014). Human value scales were found scalar invariant between online and paper-pencil surveys in Davidov and Depner's study. But, in Fang's study, paper-pencil survey was found nonequivalent to social media surveys on personal and global innovativeness scales. To the best of our knowledge, there is no research yet examining the measurement invariance for psychological questionnaires across common survey methods within representative samples. When comparing groups, it is assumed that the used measures target the same construct in all groups. If this assumption does not hold, however, the comparisons across the groups can neither be evaluated meaningfully nor interpreted adequately. Therefore, the establishment of measurement invariance is a prerequisite when applying self-report measures (Milfont & Fischer, 2010). Hence, its investigation is an important target when using self-report measures.

Within this context there is another issue to consider. That is, it may make a difference whether the self-report scales target more or less general, innocuous personality characteristics or more sensitive constructs such as positive or negative aspects of mental health. The latter concepts are often related to issues that many people consider socially sensitive, e.g., social support, represented by the number of friends one has, or personal (un-) happiness (Fydrich, Sommer, Tydecks, & Brähler, 2009; Kessler et al., 2015; Maercker et al., 2015). Following this, our study addressed these particular domains.

The present study had two main foci, namely examining the role of social desirability for and the existence of measurement invariance in various data collection methods assessing positive and negative aspects of mental health. Therefore, four survey methods in four German representative samples were applied: face-to-face interviewing, online questionnaires, offline questionnaires, and telephone interviewing. All four survey methods included thirteen different measures assessing positive and negative mental health. In order to ensure sufficient statistical power and generalizability of the results, we studied large representative population samples (N > 2000 for each sample). There were three research aims. The first is related to the role of social desirability. Social desirability was operationalized as the difference in responses for different kinds of self-report measures for all four survey methods. There were two research questions: Will the largest difference in responses for the different kind of measures occur between online and telephone samples (see Holbrook et al., 2003), or between offline and telephone samples (see Dodou & de Winter 2014). Will the online sample deliver the same responses for different kind of self-report measures as the offline sample? This would be in line with results of the meta-analysis by Dodou and de Winter (2014). The second aim involved an exploratory question and concerned the moderating role of age, gender, and education level for the observed effect of social desirability. The third aim concerned the measurement invariance. Here, we tested the configural invariance, weak invariance, and strong invariance across the four survey methods.

2. Methods

Participants were recruited within the Bochum Optimism and Mental Health Studies (BOOM) program, which aimed to identify protective factors related to positive mental health in different countries. Four representative German samples were tested in 2012, each one using a different data collection method: face-to-face interview, online questionnaire, telephone interview, or offline-panel (Forsa.Omninet). Each sampling had its own procedure:

The face-to-face sample ($N = 1870$) and the online sample ($N = 2039$) were both conducted via the market research company GfK, and included the same weighting factors, i.e., age, gender, state, city size, size of household and occupation of head of household. The face-to-face sample used the Computer Assisted Multimedia Questioning (CAM) method and the online sample used the Computer Assisted Web Interviewing (CAWI) method.

The Offline sample (Forsa.Omninet) ($N = 2021$) was collected by a German market research company named Forsa Ltd. The respondents answered the questions on their home PC or on their TV screen, which are linked to Forsa's own proprietary environment using a device called “set-top-box", implying that the internet was not needed for this data collection method. The Forsa.Omninet sample currently consists of 10.000 representatively selected households in Germany. The data was weighted by age, gender, federal state, and education.

The telephone sample ($N = 2007$) was conducted by another German market research company called USUMA. The sampling frame, which is called “ADM-Telefonstichproben-System”, is based on the amount of available telephone numbers in Germany as updated by the government agency in charge of the German telephone network. It covers all possible telephone numbers in Germany, independent of whether they are used or not. The data was weighted by age, gender, and household size.

All these specification of weighting factors are based on the most recent data provided by the federal statistical office in Germany.

2.1. Positive mental health scales

2.1.1. Sense of coherence

This scale is a shortened form (Schumacher, Gunzelmann, & Brähler, 2000) of the 29-item-version from Antonovsky (Antonovsky, 1987) and consists of 9 items assessing comprehensibility, manageability, meaningfulness. Each item (e.g. ‘Do you have the feeling that you are in an unfamiliar situation and don't know what to do?’) has a 7-point Likert scale. This short version was validated by Schumacher in a representative German sample. Cronbach's α in our four samples varied from 0.78 to 0.89.

2.1.2. Resilience

This scale is a shortened form (Schumacher, Leppert, & Gunzelmann, 2004) of the 25-item-version from Wagnild and Young (Wagnild & Young, 1993). It consists of 11 items assessing positive resilient personality characteristics on a 7-point Likert scale from 1 (‘I disagree’) to 7 (‘I agree’). The German version has been validated by Schumacher et al. Cronbach's α in our four samples varied from 0.88 to 0.93.

2.1.3. Satisfaction with life

This scale (Diener, Emmons, Larsen, & Griffin, 1985) consists of 5 items focusing on global life satisfaction. A 7-point Likert scale from 1 (‘strongly disagree’) to 7 (‘strongly agree’) indicates the agreement with each item. Cronbach's α in our four samples varied from 0.84 to 0.92.

2.1.4. Positive mental health

This 9-item questionnaire (Lukat, Margraf, Lutz, van der Veld, & Becker, 2016) comprises statements like: 'Much of what I do brings me joy'. These items can be answered on a 4-point Likert scale ranging from 1 ('I disagree') to 4 ('I agree'). An earlier version of the scale was used successfully in our earlier Dresden Predictor Study where it showed good reliability. Cronbach's α in our four samples varied from 0.89 to 0.92.

2.1.5. Social support

This scale includes 14 items that measure perceived emotional and instrumental support and social integration (Fydrich et al., 2009). It uses a 5-point Likert scale ranging from 1 ('not true') to 5 ('true') in one sum score. Cronbach's α in our four samples varied from 0.90 to 0.95.

2.1.6. Subjective happiness

This scale (Lyubomirsky & Lepper, 1999) is one of the most commonly used measures of happiness. It consists of four items. Responses are made on a 7-point Likert scale whose anchor words change according to the question. Cronbach's α in our four samples varied from 0.70 to 0.85.

2.1.7. Self-efficacy

The general self-efficacy scale (GSE; Schwarzer & Jerusalem, 1995) consists of 10 items designed to assess the person's perceived ability to manage circumstances effectively. We conducted a pilot study that obtained good psychometric properties for a shorter 5-item solution (Cronbach's alpha = 0.85), which we used in the present sample. Items can be answered on a 4-point Likert scale ranging from 1 ('I disagree') to 4 ('I agree'). Cronbach's α in our four samples varied from 0.80 to 0.86.

2.2. Negative mental health scales

2.2.1. Depressive, anxious and stressed state

We used 21 selected items from the Depression Anxiety and Stress Scale (DASS-42; Lovibond & Lovibond, 1995) to assess levels of the person's depression, anxiety and stress (seven items per subscale). Each item is rated on a 4-point Likert scale. Across our four samples, Cronbach's α of depressive state varies from 0.85 to 0.92, of anxious state varied from 0.78 to 0.87, and of stressed state varies from 0.86 to 0.90.

2.2.2. Pessimism

The Life Orientation Test (LOT-R; Glaesmer, Hoyer, Klotsche, & Herzberg, 2008; Scheier, Carver, & Bridges, 1994) consists of 10 items of which three items assess pessimism, three items assess optimism and the remaining four items are filler items. Responses are made on a 5-point Likert scale ranging from 0 ('I strongly agree') to 4 ('I strongly disagree'). According to Scheier et al. (1994), optimism and pessimism can be viewed as opposite poles of the same dimension. By adding all six scores, a total pessimism score can be calculated. Cronbach's α in our four samples varied from 0.61 to 0.79.

2.3. Additional scales

2.3.1. Tradition

This is a subscale with 4 items from the Schwartz Portrait Value questionnaire (PVQ; Schwartz, 1992), which measures the value orientations. Respondents are presented with a portrait of a person and are asked to indicate how similar the respondent is to the person portrayed. Answers range from 'very similar' to 'very dissimilar', coded from 1 to 6. Cronbach's α in our four samples varied from 0.58 to 0.71.

2.3.2. Social rhythm

This scale (Margraf, Lavallee, Zhang, & Schneider, 2016) includes 10 items and assesses the regularity with which participants engage in basic daily activities during the working days and on the weekends. Respondents are asked to assess the regularity of their waking hours, bedtimes, etc. Answers range from 1 'very regularly' to 6 'very irregularly'. Due to a technical error, no social rhythm data were collected by the offline-panel method. Cronbach's α in our remaining three samples varied from 0.61 to 0.79.

Our four samples had three common socio-demographic variables: age, gender, and education (see Table 1 for percentages, means and standard deviations).

2.4. Analysis

After the relationships between methods and the socio-demographic characteristics, which were collected in all four samples (e.g., gender, age or education), were calculated, method was found to be associated with gender, age and education. Hence, a parallelized random sample with $N = 969$ participants was drawn from each representative survey, with the same characteristics in gender, age and education. A series of ANCOVAs controlled for survey method, gender, education, age, two-way interactions between gender and survey method, between education and survey method, and between age and survey method were conducted to test whether the effect of survey method on the questionnaires outcomes was moderated by these variables. Partial eta^2 as effect size will be calculated. With our large sample size, even a very small effect could be statistically significant. Hence, we will not interpret effect sizes that are under the level of a small effect.

As the last step, a multi group analysis will be carried out to examine whether the scales were measurement invariant with four different methods. Therefore, single confirmatory factor analyses (CFA) will be conducted for each scale, to test its proposed factor structure. In case of different model propositions, the model with better fit-indices will be preferred. In case of model misspecifications, it will be tried to identify the cause of error by means of modification indices. For the model estimation we will use the Maximum likelihood estimator, which is robust when using large sample sizes and having more than five response categories (Beauducel & Herzberg, 2006). For the other scales that have five responses or less, a Weighted Least Squares Mean and Variance adjusted (WLSMV; Flora & Curran, 2004) estimator has been recommended and thus will be used.

The measurement invariance testing will include a series of model comparisons. The baseline model (model 1) with no equality constraints will test whether the patterns of the factor structures are the same across the four samples. Configural invariance exists if model 1 has a good fit and if the item loadings are significant in all samples. Model 2 is conducted with factor loadings that are constrained to be equal across the four samples. If model 2 fits the data and the fit is not substantially worse than the fit of the baseline model (model 1), weak/metric invariance is established. In model 3, the intercepts/thresholds will be constrained in addition to loadings among the four samples. Strong/scalar invariance exists if model 3 fits the data and the fit is not substantially worse than the fit of model 2. For model 2 and model 3, if full measurement invariance is not established, partial weak/strong invariance will be examined (Byrne, Shavelson, & Muthén, 1989).

Since the $\chi 2$ difference test is highly sensitive in large samples (Oishi, 2007), additional fit indices will be examined to further assess the model's fit. The root mean square of approximation (RMSEA) will be interpreted as follows: values in the range of

Table 1
Descriptive Statistics of Socio-Demographic Variables and measures.

	Face-to-face	Online	Offline	Telephone
	N = 1870	N = 2039	N = 2021	N = 2007
Gender				
Female (in %)	51.3	46.4	51.2	51.3
Education (in %)				
Not completed elementary school	6.1	1.4	2.4	4.4
Completed elementary school	34.4	8.2	39.7	15.4
Completed middle school	40.1	32.3	30.1	37.4
Graduated from high school	10.9	28.1	14.9	20.8
Completed some higher education	8.6	29.9	13	22.1
Age				
Mean (SD)	49.38 (17.73)	42.20 (14.95)	49.23 (17.19)	49.79 (18.24)

0.00–0.05 indicate close fit, those between 0.05 and 0.08 indicate fair fit, those between 0.08 and 0.10 indicate mediocre fit (Browne & Cudeck, 1993; Steiger, 1990), and values above 0.10 indicate unacceptable fit (MacCallum, Widaman, Preacher, & Hong, 2001). The comparative fit index (CFI; Bentler, 1990) indicates a good fit if values are greater than 0.90. The standardized root mean square residual (SRMR) will also be reported when using Maximum-likelihood-estimator. Here, values smaller than 0.09 indicate a good fit, since equality constraints will mostly lead to decreases in fit indices. The rule of ΔCFI not greater than 0.01 (Vandenberg & Lance, 2000) is recommended.

Data were screened for missing values and identified cases were not included in the analysis. All analyses were calculated with SPSS 22 and R version 3.0.3 with the Package “lavaan”.

3. Results

3.1. Aim 1: the role of social desirability

Means and standard deviations of the questionnaire outcomes of each sample are summarized in Table 2, for representative surveys and parallelized surveys, per survey method. Compared to representative surveys, the measures' values showed very small changes during the parallelization. This indicates that the potential difference of responses for the self-report measures across the survey methods are unrelated by the disparities in gender, age, and levels of education in the representative surveys. Hence, we focus on the results of the representative surveys. As stated before, social desirability was operationalized as the difference in responses for different kinds of self-report measures for all four survey methods. Cohen's d (Cohen, 1988) was calculated to display the difference between all compared samples (for an overview of all Cohen's d, see Table 3), with >0.2 indicating small effect, > 0.5 indicating medium effect, and >0.8 indicating large effect.

3.1.1. Positive mental health scales

Descriptive statistics showed that participants responded most negatively in the online/offline sample. At the same time, participants responded most positively in the telephone sample. Therefore, the largest differences for the seven positive mental health scales were all between the online/offline and telephone samples (see Table 2). The differences between the online and telephone samples, and between the offline and telephone samples were all statistically significant. However, the greatest difference was found between the online and telephone samples for six out of seven positive mental health scales with Cohen's d varied from 0.44 to 0.81. For the subjective happiness scale, the greatest difference with Cohen's d = 0.46 was found between offline and telephone samples. The differences between the telephone and face-to-face samples and between the face-to-face and online samples were statistically significant for all seven positive mental health scales. However, the difference between face-to-face sample and offline sample was only statistically significant for the sense of coherence scale, the social support scale, and the subjective happiness scale. Finally, the difference between online and offline samples was statistically significant for the resilience scale, the positive mental health scale, the social support scale, and the self-efficacy scale.

3.1.2. Negative mental health scales

Descriptive statistics showed that participants responded most negatively in the online sample. At the same time, participants responded most positively in the telephone sample for the

Table 2
Means and Standard deviations of measures in the representative surveys and in the parallelized surveys.

	Representative Surveys				Parallelized Surveys			
	Face-to-face	Online	Offline	Telephone	Face-to-Face	Online	Offline	Telephone
	M(SD)	M(SD)	M(SD)	M(SD)	M(SD)	M(SD)	M(SD)	M(SD)
Sense of Coherence	46.78 (9.31)	44.87 (9.34)	45.29 (9.50)	50.04 (8.05)	47.71 (8.86)	44.91 (9.53)	45.45 (9.42)	49.34 (8.20)
Resilience	60.18 (10.38)	58.43 (11.05)	60.12 (10.01)	64.79 (9.05)	61.82 (9.48)	58.35 (11.05)	60.35 (9.98)	64.63 (8.89)
Satisfaction with life	24.22 (6.30)	23.45 (6.52)	23.71 (6.12)	27.24 (5.72)	24.65 (6.24)	23.12 (6.53)	23.7 (6.17)	26.91 (5.7)
Positive mental health	19.67 (4.70)	18.71 (4.99)	19.47 (5.78)	21.97 (4.68)	20.25 (4.44)	18.7 (4.93)	19.61 (5.73)	21.55 (4.84)
Social support	59.92 (9.19)	55.8 (11.21)	58.97 (11.00)	63.65 (8.01)	60.84 (8.97)	55.95 (11.25)	59.35 (10.6)	63.68 (7.66)
Subjective happiness	20.72 (4.27)	19.8 (4.75)	19.61 (4.92)	21.68 (4.14)	21.12 (4.1)	20.01 (4.74)	19.75 (4.87)	21.43 (4.19)
Self efficacy	15.27 (2.46)	14.82 (2.57)	15.1 (2.38)	15.93 (2.43)	15.62 (2.39)	14.86 (2.53)	15.05 (2.4)	15.78 (2.47)
Depression	2.79 (3.65)	4.44 (4.74)	3.92 (4.20)	2.37 (3.46)	2.45 (3.39)	4.21 (4.52)	3.69 (4.18)	2.54 (3.63)
Anxiety	1.89 (2.86)	3.34 (3.90)	2.64 (2.99)	1.98 (3.15)	1.61 (2.69)	3.19 (3.65)	2.4 (2.77)	2 (3.08)
Stress	4.49 (3.90)	6.35 (4.77)	5.72 (3.91)	4.81 (4.58)	4.37 (3.81)	6.01 (4.68)	5.62 (3.87)	5.22 (4.75)
Pessimism	8.63 (3.82)	9.14 (4.08)	8.61 (4.32)	7.07 (3.84)	8.18 (3.8)	9.19 (4.1)	8.45 (4.33)	7.4 (3.78)
Tradition	13.18 (3.90)	14.79 (3.74)	14.76 (3.88)	13.44 (4.03)	13.63 (3.82)	14.54 (3.79)	15.22 (3.75)	13.57 (4.02)
Social rhythm	28.97 (8.31)	28.46 (8.85)	/	28.12 (9.41)	29.33 (8.53)	28.77 (9.19)	/	28.55 (9.66)

Table 3
Cohen's d of pairwise mode comparisons for each scale.

	Representative sample						Parallelized sample					
	1 vs 2	1 vs 3	1 vs 4	2 vs 3	2 vs 4	3 vs 4	1 vs 2	1 vs 3	1 vs 4	2 vs 3	2 vs 4	3 vs 4
Sense of Coherence	0.20***	0.16***	0.37***	0.04	0.59***	0.54***	0.30***	0.25***	0.19***	0.06	0.50***	0.44***
Resilience	0.16***	0.01	0.47***	0.16***	0.63***	0.49***	0.34***	0.15**	0.31***	0.19***	0.63***	0.43***
Satisfaction with life	0.12**	0.08	0.50***	0.04	0.62***	0.60***	0.24***	0.15**	0.38***	0.09	0.62***	0.54***
Positive mental health	0.20***	0.04	0.40***	0.14***	0.67***	0.48***	0.33***	0.13*	0.28***	0.17**	0.58***	0.37***
Social support	0.40***	0.09*	0.43***	0.29***	0.81***	0.49***	0.48***	0.15**	0.34***	0.32***	0.93***	0.47***
Subjective happiness	0.20***	0.24***	0.23***	0.03	0.42***	0.46***	0.25***	0.30***	0.08	0.05	0.32***	0.37***
Self efficacy	0.18***	0.07	0.27***	0.11**	0.44***	0.35***	0.31***	0.24***	0.07	0.08	0.37***	0.30***
Depression	0.39***	0.29***	0.12**	0.12**	0.50***	0.40***	0.44***	0.33***	0.03	0.12	0.41***	0.29***
Anxiety	0.42***	0.26***	0.03	0.20***	0.38***	0.21***	0.49***	0.29***	0.14*	0.24***	0.35***	0.14*
Stress	0.43***	0.31***	0.08	0.14***	0.33***	0.21***	0.38***	0.33***	0.20***	0.09	0.17**	0.09
Pessimism	0.13***	0.00	0.41***	0.13***	0.52***	0.38***	0.26***	0.07	0.21***	0.18**	0.45***	0.26***
Tradition	0.42***	0.41***	0.07	0.01	0.35***	0.33***	0.24***	0.42***	0.02	0.18**	0.25***	0.42***
Social rhythm	0.06	/	0.10*	/	0.04	/	0.06		0.09		0.02	

Note: *p < 0.05, two-tailed. **p < 0.01, two-tailed. ***p < 0.001, two-tailed.
1 = Face-to-face sample; 2 = Online sample; 3 = Offline sample; 4 = Telephone sample.

depression scale and the pessimism scale. For the anxiety and stress scales, the most positive values occurred in the face-to-face sample (see Table 2). The differences between the online sample and the telephone sample, and between the offline sample and the telephone sample were all statistically significant. Here, the largest differences were found between the online sample and the telephone sample for the depression scale with Cohen's d = 0.50 and for the pessimism scale with Cohen's d = 0.52. Furthermore, the largest differences was found between the online and telephone samples for the anxiety scale with Cohen's d = 0.42 and for the stress scale with Cohen's d = 0.43. The difference between the telephone and face-to-face samples was statistically significant for the depression scale and the pessimism scale, but not for the anxiety and stress scales. The difference between the face-to-face and online samples was statistically significant for all four negative mental health scales. The difference between the face-to-face and offline samples was statistically significant for the depression, anxiety and stress scales, but not for the pessimism scale. The difference between the online and offline samples was statistically significant for all four negative mental health scales.

3.1.3. The additional scales

For the tradition scale, participants responded least traditionally in the online sample and most traditionally in the face-to-face sample. The greatest and significant difference with Cohen's d = 0.42 was thus found between the face-to-face and the online samples. The differences were statistically significant between the face-to-face and offline samples, and between the online/offline and telephone samples. However, no statistically significant difference was found between the face-to-face and telephone samples and between the online and offline samples for the tradition scale. For the social rhythm scale, participants responded that they lived most irregularly in the face-to-face sample and most regularly in the telephone sample. Here, only one statistically significant difference with Cohen's d = 0.1 was found between the face-to-face and telephone samples. The social rhythm scale was not assessed in the offline sample. Hence, we could not compare this sample with other samples. Regarding the other possible comparisons, we neither found a statistically significant difference between the face-to-face and online samples, nor between the online and telephone samples.

3.2. Aim 2: moderating role of age, gender, education level

The results of the ANCOVAs with survey method, gender, and educations as between-subject factors and age as covariate revealed three two-way interactions: between gender and survey methods, between education and survey methods, and between age and survey methods are presented in Table 4, for each self-report measure. The commonly used index for effect sizes in analyses of variance is the partial Eta squared, as suggested by Cohen (1988) with >0.01 indicating a small effect, > 0.06 indicating a medium effect and >0.14 indicating a large effect.

3.2.1. Positive mental health scales

Across the positive mental health scales, survey method was found to be a significant between-subjects factor with the effect size varied from 0.011 to 0.028. Gender was found to be either significant, but with an effect sizes under the small effect level, or being non-significant for six of the seven positive mental health scales. Except for the social support scale, gender was found to be a significant between-subjects factor, with an effect size equaling 0.011. Education was either significant with effect sizes under the small effect level or non-significant for four of the seven positive mental health scales. For the other three scales, education was found to be a significant between-subjects factor, with the effect sizes varying from 0.011 to 0.014. Age was found to be either significant with effect sizes under the small effect level or non-significant for six of the seven positive mental health scales. Except for the sense of coherence scale, age was found to be a significant between-subjects factor, with an effect size equaling 0.027. The interaction between gender and survey method was either significant with effect sizes under the small effect level or non-significant for all seven positive mental health scales. The same was true for the interaction between education and survey method. The interaction between age and survey method was either significant with effect sizes under the small effect level or non-significant for three of the seven positive mental health scales. For the sense of coherence scale, the resilience scale, the positive mental health scale, and the subjective happiness scale, the interaction between age and survey method was found to be a significant between-subjects factor, with effect sizes varying from 0.012 to 0.017.

3.2.2. Negative mental health scales

Across the negative mental health scales, survey method was found to be a significant between-subjects factor with the effect sizes varied from 0.012 to 0.028. Gender was found to be either significant with effect sizes under the small effect level or non-significant for all four negative mental health scales. Education was found to be significant for the stress scale, but with an effect

Table 4
Effect size (partial eta^2) from ANCOVA controlled for survey method, gender, education, age, interactions between gender and survey method, between education and survey method, and between age and survey method in the representative surveys.

	Survey method	Gender	Education	Age	Gender*Survey method	Education*Survey method	Age*Survey method
Sense of Coherence	0.028***	0.001**	0.014***	0.027***	0.001*	0.005***	0.017***
Resilience	0.022***	0.001**	0.011***	0.006***	0.001*	0.006***	0.013***
Satisfaction with life	0.013***	0	0.012***	0.002***	0.001*	0.004**	0.005***
Positive mental health	0.019***	0	0.007***	0.004***	0.001*	0.004***	0.012***
Social support	0.022***	0.011***	0.006***	0	0.002**	0.003*	0.006***
Subjective happiness	0.023***	0.001**	0.005***	0.01***	0.003***	0.003**	0.014***
Self efficacy	0.011***	0.002***	0.003***	0.009***	0	0.009***	0.006***
Depression	0.028***	0.001*	0.011***	0.01***	0	0.003**	0.017***
Anxiety	0.024***	0.001*	0.014***	0.001**	0	0.003**	0.013***
Stress	0.012***	0.005***	0.005***	0.02***	0	0.003**	0.007***
Pessimism	0.019***	0	0.024***	0.005***	0	0.003*	0.01***
Tradition	0.006***	0	0.018***	0.034***	0.002**	0.003*	0.007***
Social rhythm	0.001	0	0.002*	0.002**	0	0.004**	0

Note: *p < 0.05. **p < 0.01. ***p < 0.001.

size under the small effect level. For the other three negative mental health scales, education was found to be a statistically significant between-subjects factor, with effect sizes varying from 0.011 to 0.024. Finally, age was found to be significant with effect sizes under the small effect level for all four negative mental health scales. The interaction between gender and survey method was not significant for all four negative mental health scales. However, the interaction between education and survey method was significant with effect sizes under the small effect level for all four negative mental health scales. The interaction between age and survey method was also significant with effect sizes under the small effect level for the stress scale and the pessimism scale. Finally, the interaction between age and survey method was found to be a significant between-subjects factor, with an effect size of 0.017 for the depression scale, and 0.013 for the anxiety scale.

3.2.3. The additional scales

Survey method was found to be either significant with effect sizes under the small effect level or non-significant for the two additional scales. No statistically significant gender differences were found for the two additional scales. Education was found to be a statistically significant between-subjects factor with an effect size of.018 for the tradition scale. It was also significant with an effect size under the small effect level for the social rhythm scale. Age was found to be a statistically significant between-subjects factor with an effect size of 0.034 for the tradition scale. Furthermore, it was significant with an effect size under the small effect level for the social rhythm scale. All three interactions, i.e., between gender and survey method, between education and survey method, and between age and survey method, were found to be either significant with effect sizes under the small effect level or non-significant for the two additional scales.

3.3. Aim 3: measurement invariance per self-report measure

3.3.1. Sense of coherence

The one-dimensional model of this scale did not show a reasonable global fit, especially for the offline sample. Modification indices suggested that there were correlated error residuals between item 1 and item 4, item 2 and item 3, item 2 and item 4, and item 3 and item 4. This was true for all four samples. After allowing these error terms correlate, the model fit was improved. All factor loadings in all four samples were at least 0.5. Configural invariance was established and full weak invariance as well. After we set one intercept free, partial strong invariance was established. Furthermore, full strong invariance was established among the face-to-face, online and offline samples.

3.3.2. Resilience

The one-dimensional model of this scale showed a reasonable global fit, however the RMSEAs and the CFI of the offline sample indicated a poor fit. Modification indices suggested that there were correlated error residuals between item 1 and item 2, item 2 and item 7, item 3 and item 6, item 3 and item 7, item 8 and item 9, item 9 and item 10, and item 10 and item 11. This was true for all four samples. After correlating these error terms, the model fit was improved. All factor loadings in all four samples were at least 0.5. Configural invariance was established as well as full weak invariance. Partial strong invariance was established after two intercepts were set free. Furthermore, full strong invariance was established among the online, offline and telephone samples.

3.3.3. Satisfaction with life

The one-dimensional model of this scale showed a good global fit. All factor loadings in all four samples were at least 0.5. Configural invariance was established as well as full weak invariance and full strong invariance. However, there was no factor mean invariance among the four samples.

3.3.4. Positive mental health

The one-dimensional model of this scale showed a reasonable to good global fit. However the RMSEA in the offline sample indicated a poor fit. There were several correlated errors in the offline, but not in the other three samples. Hence, we decided not to correlate any error terms in this model. All factor loadings in all four samples were at least 0.6. Configural invariance was established as well as full weak invariance and full strong invariance among the four samples. But no factor mean invariance was found.

3.3.5. Social support

RMSEAs of this one-dimensional model showed poor fit in three samples except for the telephone sample. Modification indices suggested that the error residuals between item 2 and 3, item 4 and item 13, item 7 and item 14, and item 10 and item 11 should be correlated in all four samples. After correlating these error terms, the model fit was improved. All factor loadings in all four samples were at least 0.4. Configural invariance, full weak invariance and full strong invariance among the four samples were established. But no factor mean invariance was found.

3.3.6. Subjective happiness

The one-dimensional model of this scale showed a reasonable global fit. CFI and SRMR values indicated even an excellent fit. However the RMSEA in the face-to-face and online samples indicated a poor fit. There were several correlated errors in these two

samples, but not in the other samples. Hence, we decided not to correlate any error terms in this model. All factor loadings in all four samples were at least 0.6. Configural invariance was established and full weak invariance, too. Partial strong invariance was established after one intercept was set free. Furthermore, the full strong invariance was established between the offline and telephone samples.

3.3.7. Self-efficacy

The one-dimensional model of this scale with good CFI values showed an excellent fit. The RMSEA indicated a poor fit only in the online and telephone samples. In the other two samples, the RMSEAs were good. Hence, we decided not to correlate any error terms in this model. All factor loadings in all four samples were at least 0.6. Configural invariance was established as well as full weak invariance, full strong invariance, and factor mean invariance among the four samples.

3.3.8. Stress

The one-dimensional model of this scale showed a good fit. All factor loadings in all four samples were at least 0.5. Configural invariance, full weak invariance, full strong invariance, and factor mean invariance were established among the four samples.

3.3.9. Anxiety

The one-dimensional model of this scale showed an excellent fit. All factor loadings in all four samples were at least 0.5. Configural invariance existed as well as full weak invariance and full strong invariance.

3.3.10. Depression

The one-dimensional model of this scale showed an excellent fit. All factor loadings in all four samples were at least 0.6. Configural invariance existed as well as full weak invariance and full strong invariance.

3.3.11. Pessimism

The two factor model without correlated errors showed a good fit. All factor loadings in all four samples were at least 0.6. Configural invariance existed as well as full weak and full strong invariance.

3.3.12. Tradition

RMSEAs of this one-dimensional model showed poor fit in three samples, except for the online sample. Modification indices suggested that the error residuals between item 1 and 4 should be correlated in all four samples. After correlating these error terms, the model fit was improved. All factor loadings in all four samples were at least 0.5. Configural invariance and full weak invariance were established. Partial strong invariance was established after two intercepts were set free. Furthermore, full strong invariance was established between the face-to-face and offline samples, the online and offline samples, and the telephone and offline samples.

3.3.13. Social rhythm

The one-dimensional model of this scale showed a very poor fit. Hence, we tried a two factor construct, which showed a better fit but was still not acceptable. With the help of modification indices, we correlated the errors between item 1 and item 4, item 1 and item 10, item 2 and item 3, item 2 and item 4, item 2 and item 9 item 2 and item 10, item 3 and item 4, item 4 and item 10, and item 9 and item 10. After correlating these error terms, the model fit was improved. All factor loadings in all three samples were at least 0.5. Configural invariance, full weak invariance, full strong invariance, and factor mean invariance existed.

4. Discussion

Our study had three main aims: 1. Examining the role of the mode effect “social desirability” across four survey methods, i.e., face-to-face interviewing, online questionnaires, offline questionnaires, and telephone interviewing; 2. Testing possible moderators for the effect of social desirability for different survey methods; 3. Investigating measurement invariance across the survey methods. We focused on positive and negative mental health scales and used large representative populations' samples in order to have sufficient power and generalizability. Our study yielded four main results. Regarding aim 1, we found that the greatest differences occurred between the online/offline and telephone samples, and this was true for all seven positive mental health scales and two negative mental health scales (i.e., the depression and the pessimism scales). For the anxiety and stress scales, participants responded almost the same in the face-to-face and telephone samples, although we expected the effect of social desirability to be larger for the telephone method than for the face-to-face method. The telephone sample differed statistically significant from the online/offline sample for all twelve scales expect for the social rhythm scale. The telephone sample did not differ statistically significant from the face-to-face sample for the following scales: the anxiety, stress, and pessimism scales. The face-to-face sample differed statistically significant from the online sample for almost all scales, except for the social rhythm scale. At the same time, the face-to-face sample did not differ statistically significant from the offline sample for six scales, i.e., the resilience, the satisfaction with life, the positive mental health, the self-efficacy, and the pessimism scales. Furthermore, there were also differences for self-report measures between the online and offline samples, although, we did not expect to find any difference between these methods. Finally, responses to the social rhythm questionnaire seemed unaffected by the type of survey method. This could be explained as follows: the social rhythm scale asks about regularity with which participants engage in basic daily activities during working days and weekends, e.g. waking up, going to bed, etc. That is, the scale's contents do not target desirable versus undesirable issues. Hence, this could explain the absence of an effect of social desirability. As such, the effect of social desirability occurred only when the scale's contents involved desirable issues. When looking at the comparisons between online and telephone samples and between face-to-face and telephone samples, the positive mental health scales seemed to be more affected than the negative mental health scales. This could be explained by the fact that questions related to positive mental health issues are more personal. To illustrate, it is more difficult to tell another person on the phone or face-to-face that “I have few friends” than to admit that “I'm getting upset”.

Regarding aim 2, the magnitude of the mode effect was moderated by age. Older participants tended to be more affected by the survey methods than the younger participants. Furthermore, analyses revealed small interaction effects (effect sizes $\geq$ 0.01), namely between age and survey methods, and this was true for four of the seven positive mental health scales (i.e., the sense of coherence scale, the resilience scale, and the positive mental health scale), and for two negative mental health scales (i.e., the depression scale and the anxiety scale). These results further showed that the mode effect was only moderated by age for a specific type of measure. However, no significant effect with an effect size above small effect was found for the interaction between gender and survey methods, or between education level and survey methods. Hence, only age seemed to be a moderator for the effect of social desirability, but that effect was small.

Regarding aim 3, measurement invariance was not always found across the different survey methods. Four of the thirteen mental

health questionnaires did not establish full strong measurement invariance. This result suggests that participants who have the same score on the observed variables would obtain a different score on the latent construct with regard to their sample membership. Results showed that full strong invariance was found for the sense of coherence scale among the face-to-face, online and telephone samples. For the resilience scale, full strong invariance was found among the online, offline, and telephone samples. For the subjective happiness scale, full strong invariance was found only between the offline and telephone samples. This result indicated, that the same constructs of measures were not always hold between online and offline methods. For the tradition scale, full strong invariance was found between the face-to-face and offline, online and offline, and the telephone and offline samples. This indicated that self-report measures using different survey methods can obtain different constructs. However, this means that the scale cannot be compared meaningfully.

5. Conclusions

To the best of our knowledge, we present the first examination of measurement invariance especially for psychological questionnaires across four common survey methods within representative samples. Hence, our study filled an important gap by exploring the equivalence among face-to-face, online, offline, and telephone for positive mental health scales, negative mental health scales, and additional scales. Based on our data, we suggest the following. If different survey methods are used, then the existence of mode effects should be tested. Also in case of using online and offline methods, one should also be careful with the difference in responses. If mode effects are found, survey method should be treated as a control variable in any further analyses. Furthermore, in case age varies, its moderating effect on survey methods should be controlled for, too. Finally, measurement invariance should be tested as well to ensure that measures are meaningfully comparable. For follow up studies, we recommend to investigate the different impression for different survey methods by participants and try to understand what actually causes the differences across the examined constructs, especially between online and offline methods. Here, it would be relevant to examine whether something else than social desirability should be considerate for the mode effects, i.e. response set or administration modality.

References

Antonovsky, A. (1987). *Unraveling the mystery of health: How people manage stress and stay well*. San Francisco: Jossey-Bass.

Beauducel, A., & Herzberg, P. Y. (2006). On the performance of maximum likelihood versus means and variance adjusted weighted least squares estimation in CFA. *Structural Equation Modeling: A Multidisciplinary Journal, 13*(2), 186–203. http://dx.doi.org/10.1207/s15328007sem1302_2.

Beebe, T. J., Mika, T., Harrison, P. A., Anderson, R., & Fulkerson, J. A. (1997). Computerized school surveys. *Social Science Computer Review, 15*(2), 159–169. http://dx.doi.org/10.1177/089443939701500204.

Bentler, P. M. (1990). Comparative fit indexes in structural models. *Psychological Bulletin, 107*(2), 238–246. http://dx.doi.org/10.1037/00332909.107.2.-238.

Booth-Kewley, S., Edwards, J. E., & Rosenfeld, P. (1992). Impression management, social desirability, and computer administration of attitude questionnaires: Does the computer make a difference? *Journal of Applied Psychology, 77*(4), 562–566. http://dx.doi.org/10.1037/0021-9010.77.4.562.

Browne, M. W., & Cudeck, R. (1993). Alternative ways of assessing model fit. In K. A. Bollen, & J. S. Long (Eds.), *Testing structural equation models* (pp. 136–162). Beverly Hills. CA: Sage.

Byrne, B. M., Shavelson, R. J., & Muthén, B. (1989). Testing for the equivalence of factor covariance and mean structures: The issue of partial measurement invariance. *Psychological Bulletin, 105*(3), 456–466. http://dx.doi.org/10.1037/0033-2909.105.3.456.

Chang, L., & Krosnick, J. A. (2009). National surveys via RDD telephone interviewing versus the internet: Comparing sample representativeness and response quality. *Public Opinion Quarterly, 73*(4), 641–678. http://dx.doi.org/10.1093/poq/nfp075.

Cohen, J. (1988). *Statistical power analysis for the behavioral sciences* (2nd ed.). Hillsdale, NJ: Lawrence Earlbaum Associates.

Davidov, E., & Depner, F. (2011). Testing for measurement equivalence of human values across online and paper-and-pencil surveys. *Quality and Quantity, 45*(2), 375–390. http://dx.doi.org/10.1007/s11135-009-9297-9.

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

Dillman, D. A., Phelps, G., Tortora, R., Swift, K., Kohrell, J., Berck, J., et al. (2008). Response rate and measurement in mixed mode surveys using mail, telephone, interactive voice response, and internet. *Social Science Research, 38*(1), 1–48.

Dodou, D., & de Winter, J. C. F. (2014). Social desirability is the same in offline, online, and paper surveys: A meta-analysis. *Computers in Human Behavior, 36*, 487–495. http://dx.doi.org/10.1016/j.chb.2014.04.005.

Fang, J., Wen, C., & Prybutok, V. (2014). An assessment of equivalence between paper and social media surveys: The role of social desirability and satisficing. *Computers in Human Behavior, 30*, 335–343. http://dx.doi.org/10.1016/j.chb.2013.09.019.

Flora, D. B., & Curran, P. J. (2004). An empirical evaluation of alternative methods of estimation for confirmatory factor analysis with ordinal data. *Psychological Methods, 9*(4), 466–491. http://dx.doi.org/10.1037/1082-989X.9.4.466.

Fydrich, T., Sommer, G., Tydecks, S., & Brähler, E. (2009). Fragebogen zur sozialen Unterstützung (F- SozU): Normierung der Kurzform (K-14) [Social Support Questionnaire (F-SozU): Standardization of short form (K-14)]. *Zeitschrift Für Medizinische Psychologie* (January), 43–38.

Glaesmer, H., Hoyer, J., Klotsche, J., & Herzberg, P. Y. (2008). Die deutsche Version des Life-Orientation-Tests (LOT-R) zum dispositionellen Optimismus und Pessimismus. *Zeitschrift Fur Gesundheitspsychologie, 16*(1), 26–31. http://dx.doi.org/10.1026/0943-8149.16.1.26.

Griffiths, M. D., Lewis, A. M., Ortiz de Gortari, A. B., & Kuss, D. J. (2014). Online forums and solicited blogs: Innovative methodologies for online gaming data collection. *Studia Psychologica UKSW, 14*(3), 5–24.

Holbrook, A., Green, M., & Krosnick, J. (2003). Versus face-to-face interviewing of national probability samples with long questionnaires: Comparisons of respondent satisficing and social desirability response bias. *Public Opinion Quarterly, 67*, 79–125. Retrieved from: http://poq.oxfordjournals.org/content/67/1/79.short.

Kessler, R. C., Sampson, N. A., Berglund, P., Gruber, M. J., Al-Hamzawi, A., Andrade, L., et al. (2015). Anxious and non-anxious major depressive disorder in the world health organization world mental health surveys. *Epidemiology and Psychiatric Sciences, 24*(3), 210–226. http://dx.doi.org/10.1017/S2045796015000189.

Kiesler, S., & Sproull, L. S. (1986). Response effects in the electronic survey. *Public Opinion Quarterly, 50*(3), 402–413. http://dx.doi.org/10.1086/268992.

Knapp, H., & Kirk, S. A. (2003). Using pencil and paper, Internet and touch-tone phones for self-administered surveys: Does methodology matter? *Computers in Human Behavior, 19*(1), 117–134. http://dx.doi.org/10.1016/S0747-5632(02)00008-0.

Kreuter, F., Presser, S., & Tourangeau, R. (2008). Social desirability bias in CATI, IVR, and web surveys: The effects of mode and question sensitivity. *Public Opinion Quarterly, 72*(5), 847–865. http://dx.doi.org/10.1093/poq/nfn063.

Lovibond, S. H. (1995). Lovibond, P.f. *Manual for the depression anxiety stress scales*. Sydney: Psychology Foundation.

Lukat, J., Margraf, J., Lutz, R., van der Veld, W. M., & Becker, E. S. (2016). Psychometric properties of the positive mental health scale (PMH-scale). *BMC Psychology, 4*(1), 8. http://dx.doi.org/10.1186/s40359-016-0111-x.

Lyubomirsky, S., & Lepper, H. S. (1999). A measure of subjective happiness: Preliminary reliability and construct validation. *Social Indicators Research, 46*(2), 137–155. http://dx.doi.org/10.1023/A:1006824100041.

MacCallum, R. C., Widaman, K. F., Preacher, K. J., & Hong, S. (2001). Sample size in factor analysis: The role of model error. *Multivariate Behavioral Research, 36*(4), 611–637. http://dx.doi.org/10.1207/S15327906MBR3604_06.

Maercker, A., Chi Zhang, X., Gao, Z., Kochetkov, Y., Lu, S., Sang, Z., et al. (2015). Personal value orientations as mediated predictors of mental health: A three-culture study of Chinese, Russian, and German university students. *International Journal of Clinical and Health Psychology, 15*(1), 8–17. http://dx.doi.org/10.1016/j.ijchp.2014.06.001.

Margraf, J., Lavallee, K., Zhang, X., & Schneider, S. (2016). Social rhythm and mental health: A cross-cultural comparison. *PLoS One, 11*(3), 1–16. http://dx.doi.org/10.1371/journal.pone.0150312.

Milfont, T. L., & Fischer, R. (2010). Testing measurement invariance across groups: Applications in cross. *International Journal of Psychological Research, 3*(1), 111–121. http://dx.doi.org/10.1007/s11135-007-9143-x.

Oishi, S. (2007). The application of structural equation modeling and item re-sponse theory to cross-cultural positive psychology research. In A. Ong, & M. van Dulmen (Eds.), *Handbook of methods in positive psychology* (pp. 126–138). New York: Oxford University Press.

Rosenfeld, P., Booth-Kewley, S., & Edwards, J. E. (1993). Computer-administered surveys in organizational settings: "Alternatives, advantages, and applications". *The American Behavioral Scientist, 36*(4), 485–511. http://dx.doi.org/10.1017/CBO9781107415324.004.

Scheier, M. F., Carver, C. S., & Bridges, M. W. (1994). Distinguishing optimism from neuroticism (and trait anxiety, self-mastery, and self-esteem): A reevaluation of the life orientation test. *Journal of Personality and Social Psychology*. http://dx.doi.org/10.1037/0022-3514.67.6.1063.

Schumacher, J., Gunzelmann, T., & Brähler, E. (2000). Deutsche Normierung der

sense of coherence scale von Antonovsky. *Diagnostica, 46*(4), 208–213. http://dx.doi.org/10.1026//0012-1924.46.4.208.

Schumacher, J., Leppert, K., & Gunzelmann, T. (2004). Die Resilienzskala – ein Fragebogen zur Erfassung der psychischen Widerstandsfähigkeit als Personmerkmal. *Psychiatrie Und Psychotherapie, 53*(1), 16–39. Retrieved from: http://www.mentalhealthpromotion.net/resources/resilienzskala2.pdf.

Schwartz, S. H. (1992). Universals in the content and structure of values: Theoretical advances and empirical tests in 20 countries. In M. P. Zanna (Ed.), *Advances in experimental social psychology* (Vol. 25, pp. 1–65). Amsterdam: Elsevier.

Schwarzer, R., & Jerusalem, M. (1995). Generalized self-efficacy scale. In J. Weinman, S. Wright, & M.Johnston (Eds.), *Measures in health psychology: A user's portfolio. Causal and control beliefs* (pp. 35–37). Windsor, UK: NFER-NELSON.

Steiger, J. H. (1990). Structural model evaluation and Modification: An interval estimation approach. *Multivariate Behavioral Research*. http://dx.doi.org/10.1207/s15327906mbr2502_4.

Szolnoki, G., & Hoffmann, D. (2013). Online, face-to-face and telephone surveys - comparing different sampling methods in wine consumer research. *Wine Economics and Policy, 2*(2), 57–66. http://dx.doi.org/10.1016/j.wep.2013.10.001.

Vandenberg, R. J., & Lance, C. E. (2000). A review and synthesis of the measurement invariance Literature: Suggestions, practices, and recommendations for organizational research. *Organizational Research Methods, 3*(1), 4–70. http://dx.doi.org/10.1177/109442810031002.

Wagnild, G. M., & Young, H. M. (1993). Development and psychometric evaluation of the resilience scale. *Journal of Nursing Measurement, 1*, 165–178. Retrieved from: http://www.resiliencescale.com/wp-content/uploads/-2014/06/Wagnild-Young-psychom-R.pdf.