PSIHOLOGIJA, 2015, Vol. 48(4), 431–449

UDC 303.64
159.9.072
DOI: 10.2298/PSI1504431M

The impact of frequency rating scale formats on the measurement of latent variables in web surveys – An experimental investigation using a measure of affectivity as an example

Natalja Menold[1] & Christoph J. Kemper[2]

[1]*GESIS—Leibniz Institute for the Social Sciences, Mannheim, Germany*
[2]*University of Luxembourg, Luxembourg*

The effects of verbal and/or numerical labeling and number of categories on the measurement of latent variables in web surveys are addressed. Data were collected online in a quota sample of the German adult population (N = 741). A randomized 2x2x2 experimental design was applied, with variation of the number of categories, as well as of verbal and numerical labeling, using an abbreviated version of the Positive and Negative Affect Schedule (PANAS). Experimental manipulation of the rating scale formats resulted in an effect on measurement model testing and reliability, as well as on factorial and convergent validity. In addition, measurement invariance between several rating scale formats was limited. With the five category end verbalized and fully labeled seven category formats, acceptable results for all measurement quality metrics could be obtained.

Keywords: rating scales, labeling, number of categories, reliability, validity, measurement equivalence, PANAS

Psychological concepts are often measured by a number of observed variables (responses to questionnaire items) that aim to represent certain latent dimensions. Data are collected with the help of rating scales that allow respondents to grade their judgments on a continuum, for example, an agree-disagree or a never-always continuum. Researchers are interested in obtaining reliable and valid measurements of the corresponding latent constructs. However, not only the items' content but also the rating scale formats can impact on the measurements' reliability and validity (e.g., Krosnick & Fabrigar, 1997).

The number of categories and category labels are basic, visual, task-related cues for the respondents. The number of response categories and labels for extreme categories represent the frequency and range of the continuum visualized by a rating scale, whereas verbal or numerical labels of intermediate categories clarify the meaning of sub-ranges (Parducci, 1983). The verbal labeling of each

Corresponding author: natalja.menold@gesis.org

category seems to reduce interpretation variability across respondents (Krosnick & Fabrigar, 1997). Correspondingly, some authors found that fully verbalized rating scales increase reliability (e.g., Alwin & Krosnick, 1991; Menold, Kaczmireck, Lenzner, & Neusar, 2014; Weng, 2004), compared to the formats in which only the end categories are verbally labeled (with and without numerical labels). However, fully verbalized rating scales require respondents to increase their cognitive effort when attending to each verbal label. Difficulties in the mapping stage may increase the chance of superficial information processing ("satisficing"), which, in turn, reduces the measurement quality (Krosnick & Fabrigar, 1997). In line with these considerations, some studies did not find any positive effects of fully verbalized format on reliability or validity, compared with end categories only verbalized rating scales (e.g., Andrews, 1984; Wakita, Ueshima, & Noguchi, 2012; for a meta-analysis see Churchill & Peter, 1984).

Compared with verbal labels, numerical labels are assumed to be more burdensome for respondents because it is not natural to use numbers for self-descriptions (Krosnick & Fabrigar, 1997). Christian, Parson, and Dillman (2009) showed that a format using both numerical and verbal labels required a longer administration time than a format using verbal labels only. However, Churchill and Peter (1984) did not find any effect of numerical labels on reliability.

Concerning the number of categories, five to seven categories are associated with the highest reliability or validity (for overviews, see Krosnick & Fabrigar, 1997; Maitland, 2009). Nonetheless, researchers continue to discuss differences between five and seven categories (e.g., Revilla, Saris, & Krosnick, 2014), because the respondents' cognitive burden is expected to be higher in the case of seven categories (cf. Toepoel & Dillman, 2010). However, the majority of previous studies considered only either the number of categories (e.g., Mullins, Polson, Lanch, & Kehoe, 2007) or the type of labeling (e.g., Alwin & Krosnick, 1991; Krosnick & Berent, 1993) when addressing measurement quality. Some researchers have even mixed labeling and number of categories when comparing rating scales. For instance, Revilla et al. (2014) compared, in one Split-Ballot Multitrait-Multimethod (MTMM) study, a five category fully labeled agree-disagree format with the seven end category verbalized only format and concluded that five categories perform better than seven categories with respect to measurement quality. However, this result can also be explained by using full verbalization for the five-category case. The data from this study do not provide conclusive proof about the effect of the number of categories or of verbalization, since it is not known how the five category end verbalized or seven category fully verbalized formats performed.

Studies that have systematically and simultaneously addressed both aspects – the verbal labeling and the number of categories – are rather scarce. Of these, Weng (2004) analyzed the effects of the number of categories (three to nine) and verbal labeling on Cronbach's α (Cronbach, 1951) and retest reliability. He found a better retest reliability for seven fully verbally labeled rating scales but no effect on Cronbach's α. However, applying Cronbach's α as a reliability metric requires that items are at least essentially tau-equivalent measures of the latent

construct, i.e., have equal factor loadings (e.g., Lord & Novick, 1968). This assumption was not tested by Weng (2004). Wedell, Parducci, and Lane (1990) showed that providing seven fully labeled categories is optimal with respect to discriminating power and the reduction of context effects, which represents a kind of priming. However, the authors only compared scales with three and seven categories. In addition, there is a lack of studies that simultaneously varied verbal and numerical labels as well as the number of categories.

With respect to web surveys, only a small number of studies on labeling of rating scales is available. Menold et al. (2014) found a higher cross-sectional reliability for fully verbalized five-category rating scales, compared to rating scales with verbally labeled end categories and numeric labels for each category. However, they did not use a web survey per se but applied an experimental laboratory setting with a web survey interface. Tourangeau, Couper, and Conrad (2007) investigated the effects of color shading of response options in rating scales in web surveys and found that a full verbalization of labels eliminated the context effects of color, which were given in rating scales without verbal labels for each category or in rating scales with numeric labels and verbal labels for the end categories only. Toepoel and Dillman (2010) confirmed this result. In another experiment, Toepoel and Dillman (2010) addressed the effect of uneven spacing of rating-scale categories on the responses. The effect of spacing was not observable in the case of fully labeled rating scales, in which either verbal or numeric labels were used; however, it did appear in the case of end category only verbally labeled rating scales. Nevertheless, measurement quality, i.e., reliability or validity, has rarely been addressed in the context of web surveys.

Whilst there are numerous studies that address the effect of rating scale formats (besides web surveys) on reliability (retest, internal consistency) and convergent or divergent validity, there is less evidence on how rating scale formats affect the results of measurement model tests and factorial validity, if latent variables are measured with the help of multiple observed variables. Split-Ballot MTMM studies, which have been gaining increasing popularity in the social sciences (e.g., Saris & Gallhofer, 2007; Revilla et al., 2014), do not focus on latent variable measurement; additionally, randomized fully crossed between-group experimental design is not realized with this approach. The latter is very relevant because it allows for causal explanations about the effects of rating scale formats on measurement quality.

Measurement model and factorial validity deal with certain assumptions about measuring latent variables. One type of measurement model is referred to as the congeneric measurement model (CTT) (e.g., Raykov & Marcoulides, 2011). CTT is based on two assumptions. The first assumption is that there should be significant and substantially strong relationships between the observed and the latent variables. The second assumption is that the observed variables measure only one latent variable (univocality or unidimensionality assumption)[1].

1 We do not address or discuss the case of general or hierarchical models, for which there are more complex relationships between the observed and latent variables (e.g., Raykov & Marcoulides, 2011).

Methods for reliability assessment rely on measurement model assumptions; therefore, these should be tested prior to any assessment of reliability (Lord & Novick, 1968; Raykov & Marcoulides, 2011). Sometimes, it is assumed that a construct is being measured with a set of unidimensional factors (latent variables) so that, besides the CTT assumptions for each dimension, additional assumptions about the number of dimensions and relationships between them are met. For example, in the case of personality measurement and the application of the Big Five model, five unidimensional factors are expected to be supported by empirical investigations, e.g., by means of factor analyses. Supporting assumptions of CTT (in the case of unidimensional measures) and other dimensionality assumptions with empirical data analysis, using a Factor Analysis, is referred to as factorial validity (Byrne, 2011; Cronbach & Meehl, 1955). Obtaining factorial validity is crucial, because the results provide evidence for the theoretical assumptions about the structure of a scientific concept. Furthermore, it serves as a basis for scoring, summarizing, or modeling variable values to represent measures for one or more latent variables. Menold and Tausch (2015) provided evidence that rating scale formats (5 vs. 7 categories and verbal labeling) may affect the results of testing CTT assumptions for paper-and-pencil surveys. The authors also discuss measurement equivalence between various rating scales. Measurement equivalence is related to the comparability of loadings and intercepts of items between groups of persons when measuring a latent dimension or construct (e.g., Byrne, 2011). If rating scales affect factor loadings or intercepts of single items, then the comparability of measures or results may be limited, across groups, when different rating scales are used. Such comparisons are needed when researchers intentionally or unintentionally change rating scale formats, e.g., in different waves of one survey, or when data from different surveys are compared. A lack of measurement equivalence with respect to the factorial structure and factor loadings (configural and metric invariance, respectively) makes comparisons between groups impossible. Means can be compared between groups only when the items' intercepts are invariant (scalar invariance). For a more detailed discussion on measurement invariance, see Byrne (2011) or Steinmetz (2015). Because there is a lack of studies that address the measurement invariance of different rating scale formats (cf. Menold & Tausch, 2015), this issue is also addressed in the present article.

In summary, previous research does not provide clear-cut results with respect to the effects of verbal vs. numerical labeling or number of categories on the measurement of latent variables in web surveys, which is addressed in the present study. With reference to the measurement of latent variables, we consider the results when testing CTT assumptions and factorial validity and reliability, as well as convergent and divergent validity (see definitions in, e.g., Cronbach & Meehl, 1955). These results will help researchers to understand the role of rating scale formats in the measurement of latent variables as a potential source of heterogeneity and measurement error, which may also limit the comparability of results obtained with the same items but different rating scale formats.

Method

Sample and Procedure

The sample consisted of N = 741 participants (51.8% female). The mean age of the participants was 48.3 years (SD = 13 years). Regarding education, 40.1% of the participants held a secondary school degree (German Hauptschule); 29.1% were at an intermediate secondary school level (German Realschule); and 30.8% had a high school degree (German Abitur).

Data collection was conducted online (computer-assisted web interview, CAWI) by a commercial survey institute charged with the data collection. Potential participants from an online access pool of the survey institute were invited to participate in a study on the improvement of survey measures. They were sampled according to predefined quotas for age, gender, and education to roughly resemble the German adult population above 17 years of age. After logging in, eligible participants completed the study questionnaire containing socio-demographic variables and personality and attitude measures, including the items used in the research presented here. Participation was rewarded with tokens worth 0.75 €.

PANAS

We assessed affectivity using the "Positive and Negative Affect Schedule" (PANAS). The PANAS is a widely used inventory in psychology and other applied sciences (Leue & Beauducel, 2011). The PANAS was proposed by Watson, Clark, and Tellegen (1988) as a measure of two dimensions of affective space – positive affect (PA) and negative affect (NA). As indicators of these dimensions, the PANAS contains 20 adjectives, 10 positive and 10 negative. Numerous studies have supported the psychometric quality of the original PANAS and various adaptations, e.g., the German version (e.g., Krohne, Egloff, Kohlmann, & Tausch, 1996; Leue & Lange, 2011).

Regarding the factorial validity of the PANAS, a two-factor structure was often reported in the literature, in which PA and NA were found to be correlated (cf. Leue & Lange, 2011). However, some studies could not corroborate the oblique two-dimensional structure of the PANAS – they found more than two factors or the factors were sometimes found to be correlated and sometimes not (for an overview, see Leue & Beauducel, 2011). That was the case for both state and trait measures of the PANAS.

In the present study, an abbreviated version of the PANAS with eight items, four measuring positive affect (PA1: active, PA2: enthusiastic, PA3: interested, PA4: determined) and four measuring negative affect (NA1: nervous, NA2: upset, NA3: distressed, NA4: afraid), was administered. We used a short measure to optimally simulate conditions under which survey-based research in the social sciences is usually conducted – measures have to be short to save assessment time and related costs (Ziegler, Kemper, & Kruyen, 2014). Items from the original PANAS were selected by applying statistical and content-related criteria (e.g., considering identified item clusters in the PANAS; see Egloff Schmukle, Burns, Kohlmann, & Hock, 2003), as suggested in the psychometric literature (e.g., Ziegler et al., 2014). The abbreviated PANAS version demonstrated an acceptable model fit in a confirmatory factor analysis (CFA) (oblique 2-factor model; MPlus MLM estimation, χ^2 = 67.2, df = 19, p < .001, RMSEA = .048, CFI = .963, SRMR = .036; factor loadings PA .53 - .71, NA .45 - .68; factor intercorrelation r = -.26, reliability McDonald's Ω_w (McDonald, 1999) = .72 for PA and .73 for NA; N = 1134) of a dataset reported elsewhere (Kemper, Beierlein, Kovaleva, & Rammstedt, 2013). We deemed this abbreviated version of the PANAS as acceptable for our research purposes, as it is of sufficient psychometric quality, according to the results reported above. In the present study, the trait version of the instructions was applied to measure stable dispositions, i.e., Positive and Negative Affectivity. Participants rated the frequency of the affects (see Appendix).

Experimental Design

To test the impact of rating scale formats on PANAS ratings, a fully crossed randomized three-factor 2 x 2 x 2 between-subjects design was implemented, with the following factors as independent variables:

- Number of categories: seven versus five categories.
- Verbal labeling: fully verbally labeled categories versus labels only for the end categories.
- Numerical labeling: rating scales with versus without numerical labeling of each category. The numeric labels are not placed in the area of the labels for the rating scale; instead, they are located in the area for the responses (see Appendix). Such a presentation of numbers with the rating scale is very common in psychological assessment.

The following verbal and numerical labels were used in the case of five categories: 1 = never/nearly never; 2 = seldom; 3 = occasionally; 4 = often; and 5 = very often/always. In the case of seven categories, the labels were: 1 = never; 2 = nearly never; 3 = seldom; 4 = occasionally; 5 = often; 6 = very often; and 7 = always.

Applying this design, we obtained independent experimental groups with the following eight rating scale versions: 5ALLN (five category fully verbalized format with numbers for each category); 5ENDN (five category format with numbers for each category and verbally labeled end categories); 5ALL (five category fully verbalized format without numbers); and 5END (five category format with only verbally labeled end categories; without numbers) as well as the same formats with seven categories (7ALLN; 7ENDN; 7ALL; 7END).

The eight experimental groups did not differ in terms of gender ($F_{(7, 741)} = 0.54$, $p > .10$), age ($F_{(7, 741)} = 0.78$, $p > .10$) or education, e.g., measured by years of schooling ($F_{(7, 741)} = 0.68$, $p > .10$). These results were obtained with a MANOVA (Multivariate Analysis of Variance) with subsequent ANOVAs (Univariate Analysis of Variance) and a post-hoc comparison between the experimental groups for each of the demographic variables.

Assessment of the Effects of Rating Scale Formats

To obtain comparability between the different numbers of rating scale categories, the data were separately z-standardized in each experimental group (cf. Krosnick, 2011) using the SPPS 20 software. The standardized values ranged from -3 (representing 1; "never/nearly never") to + 3 (representing 5 vs. 7 or "very often/always" vs. "always"). Next, to avoid numerical problems when using model tests (see below), categories with $n \leq 5$ were summarized with their neighboring categories. This was the case for the right and left extremes of the rating scales ($z = \pm 3$) for a few observed variables in each experimental group.

The experimental manipulation was expected to have an effect on the results of the measurement model test for each dimension of the PANAS, factorial validity, measurement equivalence, and reliability coefficients, as well as convergent and divergent validity. With respect to the measurement model assumptions, CTT was assumed in each group and for each dimension of PANAS. This assumption was tested by unidimensional CFAs. With respect to factorial validity, we aimed at confirming the two-factorial oblique structure of the PANAS in each experimental condition. For both CTT and factorial validity, a significant relationship between the NA and PA items and their corresponding factors were expected to be supported by the data. Prior to the analysis, we tested the normality of the distributions (with the Mardia test and AMOS 21) and found significant multivariate kurtosis in the groups 5END ($K = 4.97$, $p < .05$) and 7ENDN ($K = 6.75$, $p < .001$). Therefore, the analyses were conducted with an estimator robust to non-normality (MLM, Sattora-Bentler χ^2), using the Mplus6 software. Sattora-Bentler χ^2 was used to be able to conduct the Bartlett (1950) k-factor correction

for small samples. Using ML based estimators is appropriate for variables with more than four possible options because, for this case, a variable can be considered as approximately continuous (e.g., Raykov & Marcoulides, 2011, p. 91).

The results of single CFAs provide evidence as to whether CTT and two-factorial structure can be assumed in one experimental condition. However, the results of single CFAs do not provide a significance test for the effect of the experimental manipulation. This effect was assessed with various Multi-Group CFA (MGCFA) models (cf. Byrne, 2011):–

- The first and least restrictive model was the *configural model* (model 1), in which only the numbers of factors and loading patterns were assumed to be equal between the groups.
- The *structural model* (model 2) assumes the invariance of factor variances and covariances. These parameters are constrained to be equal in model 1, to obtain model 2. A substantial difference between the models 1 and 2 provides information on differences of factor variances and their correlations.
- Within the *metric model* (model 3), the factor loadings were set to be equal among the groups in model 1 (which is also the baseline for model 3). Differences in the factor loadings imply that the common factor is explained differently by the items, i.e., differences in the factor loadings may reflect differences in the construct's meaning between the groups (e.g., Bollen, 1989, Steinmetz, 2015).
- In the next step, differences in the intercepts were analyzed by restricting items' intercepts to be equal across the eight groups in model 1 (model 4, scalar, was obtained). The presence or absence of differences between the models 1 and 4 provides information as to whether groups differed with respect to the intercepts. The differences in the intercepts can be described as additive biases, reflecting artificial differences between the item means predicted by the latent variable (Hayduk, 1989).

In the case of invariance, one can identify the sources of variability, as proposed by Byrne (2011), when establishing partial invariance. This procedure was conducted as follows. We looked for the significant Modification Indices (MIs) (higher than 3.84) related only to the parameters restricted to be equal. In the case of model 2, there were, for instance, factor variances and their covariances. We started with the highest MI and included only one modification in the MGCFA model at each step, meaning that a model was tested after each single modification before proceeding with the next modification. Establishing partial invariance can be used as a way to demonstrate differences between the groups with respect to the parameters that were initially set equal.

The model fit of the CFAs and MGCFAs was assessed according to Beauducel and Wittmann (2005) by means of the χ^2 test, the root-mean-square error of approximation (*RMSEA*), and the comparative fit index (*CFI*). In addition, Standardized-Root-Mean-Residual (*SRMR*) was considered, because it is independent of sample size (Bentler, 1995). A *RMSEA* lower than .06 indicates a close fit (Hu & Bentler, 1999), and a *RMSEA* of lower than .08 indicates an acceptable fit when $n \leq 250$ (Raykov, 1998). The *CFI* should be .95 or higher, while *SRMR* should be lower than .11 to accept a model (Hu & Bentler, 1999).

The model comparison test was conducted as suggested by Byrne (2011). Significant change in a model's goodness-of-fit statistics between the nested restricted (Mr) and less restricted (Mnr) models is associated with significant differences among the groups with respect to the parameters that are restricted to be equal. We used the χ^2 difference test ($\Delta\chi^2$) between the models (Mnr – Mr), as suggested, e.g., by Byrne (2011), and obtained the differences (Mnr – Mr), also in *CFI*, *RMSEA*, and *SRMR*, as suggested by Cheung and Rensvold (2002) and by Chen (2007). A significant change of χ^2 and a change of $\leq$ -.01 in *CFI*, of $\geq$ .015 in *RMSEA* and of $\geq$ .01 in *SRMR* indicate sufficient differences between the models.

The results with respect to the comparison of model 1 with the models 2, 3, and 4 provide evidence concerning the effect of experimental manipulation on the latent factor

correlations, factor loadings, and intercepts. Factor correlations and factor loadings relate to the factorial validity of the PANAS, while factor loadings and intercepts are relevant with respect to measurement invariance and comparability of data obtained with different rating scales.

Reliability measures were assessed with methods based on CTT tests. Unlike Cronbach's α, these methods do not rely on assumptions with respect to the parallelism or equivalence of the test parts or with respect to the comparability of items' variances (e.g., Raykov & Marcoulides, 2011). First, we obtained composite reliability (ρ) with 95% Confidence Interval (CI), as described by Raykov and Marcoulides (2011). This measure is very similar to McDonalds' Ω. Composite reliability uses factor loadings and error terms from a CFA as parameters. The values of ρ were obtained for each PA and NA dimension separately in each single rating scale group, based on the acceptable CFA results when testing CTT assumptions. To reach an acceptable model fit, single correlated error terms (suggested by MIs) had to be included in some groups (see note in Table 3). Error term covariances can be justified as being substantial, i.e., there is an additional substantial latent factor, which may explain error covariances. Alternatively, they may not be substantial but, rather, represent a biased impact from known or unknown sources. Since we do not have any evidence to consider non-systematic correlated error terms included in the models as substantial effects (because they involve single rating scale groups and, within these, different items), we can assume that they are due to non-systematic method effects that would be caused by the experimental variation of rating scales. Correlated errors were suggested to be included in the equation when assessing ρ, as a part of the error variance (Raykov, 2012). In addition, Guttman's Lamda 4 (λ_4) (Guttman, 1945) was obtained, since this is a method which is also based on CTT but does not require large samples, unlike the CFA-based reliability assessment. Guttman's λ_4 is a measure for the two test halves that identifies the splits with the largest reliability.

In the last step, the convergent and divergent validity was estimated on the basis of the most tenable MGCFA model (1 to 4, see above). In this model, we included correlations between the latent NA and PA variables with exogenous (summarized, not modeled as latent) variables for Extraversion (E) and Neuroticism (N), as measured with the BFI-10 inventory (Rammstedt & John, 2007). Extraversion and Neuroticism are the two subscales of the Big Five personality inventory that were found to correlate with NA and PA in a number of previous studies. In these studies, the trait Negative Affect correlated substantially with measures of N but was generally unrelated or correlated negatively with measures of E (e.g., Krohne et al., 1996; Watson & Clark, 1997). Conversely, the trait Positive Affect was strongly and positively related to E but not, or negatively, to N. To compare the correlations of PA and NA with E and N between the groups, we obtained the 95% CIs with the standard Mplus command.

Results

CTT and Factorial Validity Tests with single CFAs

One-factor CTT models for each PANAS dimension – PA and NA – yielded reasonable model fit (non-significant χ^2, $RMSEA < .05$, $CFI > .95$, $SRMR < .06$) in the rating scale groups 5ALLN, 5END and 7ENDN. In the 7ALLN and 7ALL groups, the model fit for NA was also reasonable, whereas, for PA, RMSEA was close to .10 (7ALL) or higher (7ALLN), while other goodness-of-fit statistics were acceptable (non-significant χ^2, $CFI > .95$, $SRMR < .06$). Therefore, we also considered the model fit as reasonable in the 7ALL and the 7ALLN groups. For the 5ENDN, 5ALL, and 7END groups, model fit was poor either for PA (5ENDN, 5ALL) or both PA and NA dimensions (7END). For the models with

acceptable model fit, there was a non-significant factor loading in the 5ALLN (NA2) group. In summary, unidimensional CTT for either PA or NA or both could not be supported by the data in the rating scale groups 5ALLN, 5ENDN, 5ALL, and 7END. Therefore, five category rating scales (except 5END) seem to be problematic when establishing unidimensionality, compared to seven category rating scales (except 7END).

When evaluating the results of the CFAs for the two-factor models, a poor model fit according to χ^2, *RMSEA*, and *CFI* was obtained for the rating scale formats 5ALL and 7ALLN, whereas in all other groups, a reasonable model fit was obtained according to all goodness-of-fit statistics (table 1).

Table 1
CFA results for the two-factor model for each experimental condition

		5ALLN	5ENDN	5ALL	5END	7ALLN	7ENDN	7ALL	7END
standardized loadings									
PA	1	.38***	.41***	.45***	.77***	.78***	.66***	.65***	.58***
	2	.53***	.26*	.43***	.51***	.65***	.70***	.43***	.65***
	3	.66***	.21	.77***	.53***	.72***	.75***	.79***	.74***
	4	.61***	.95***	.65***	.76***	.51***	.72***	.77***	.75***
NA	1	.76***	.74***	.71***	.71***	.62***	.65***	.74***	.59***
	2	.17	.39***	.23*	.47***	.60***	.35**	.56***	.25
	3	.63***	.73***	.36**	.71***	.55***	.42**	.65***	.57***
	4	.87***	.67***	.89***	.77***	.81***	.91***	.75***	.65***
$r_{(PA\ with\ NA)}$		-.65***	-.66***	-.50**	-.61***	-.13	-.38**	-.37***	-.19
goodness-of-fit									
χ^2(df=19)		19.10	18.33	27.96*	19.75	39.93**	13.3	25.23	25.21
CFI		.99	1.00	.91	.99	.87	1.00	.96	.94
SRMR		.06	.06	.08	.05	.08	.05	.06	.06
RMSEA		.03	.01	.08	.04	.12	.00	.07	.06
90% CI of RMSEA		.00-.10	.00-.09	.00-.13	.00-.10	.07-.16	.00-.06	.00-.12	.00-.12
N		86	94	91	90	92	94	94	100

Notes. ***p < .001; **p < .01; *p < .05; not signed: p > .05; K-factor corrected χ^2 is reported.

With respect to the plausibility of the estimated parameters, non-significant factor loadings of the NA2 were found in the 5ALLN and 7END groups, whereas PA3 did not have a significant loading in the 5ENDN group. The factor correlations are not significant in the 7ALLN and 7END groups, which is different from other groups. Therefore, a two-factor model, which assumes that all items have at least a significant relationship with the corresponding factor, cannot be supported by the data in the 5ALLN, 5ENDN, 5ALL, 7ALLN and 7END groups, whereas the factorial validity seem to be acceptable in the 5END, 7ALL and 7ENDN groups.

Factorial Validity and Measurement Invariance Test by Means of MGCFAs

The baseline model for the measurement invariance among the eight groups was model 1 (configural), in which no equality with respect to the factor variances and correlations, factor loadings, or factor intercepts was assumed. This model yielded a tenable model fit according to *CFI, RMSEA* and *SRMR*, as shown in table 2. Therefore, it can be concluded that the configural invariance assumption is supported by the data.

Table 2
Results of model test for MGCFAs and model difference tests

Model	Goodness-of-fit					Model difference test			
	χ^2 *(df)*	CFI	SRMR	RMSEA	90% CI of RMSEA	$\Delta\chi^2(\Delta df)$	*ΔCFI*	*ΔSRMR*	*ΔRMSEA*
1 configural (baseline)	200.441** (152)	.960	.062	.059	.033–.08	-	-	-	-
2 structural	250.289*** (173)	.937	.093	.069	.049–.088	53.144*** (21)	-.023	.031	.010
3 metric	299.245*** (194)	.914	.096	.077	.059–.093	122.042*** (42)	-.046	.034	.018
4 scalar	464.733*** (208)	.789	.127	.115	0.101–0.130	279.340*** (56)	-.171	.065	.056

Notes. ***p< .001; **p<.01 *p<.05; Model specifications and comparison, see text; MLM corrected $\Delta\chi^2$ is reported.

For model 2 (structural invariance), equal latent factor variances and equal correlations between the factors were included in model 1. Although this led to a small change in *χ2, AIC*, *CFI*, and *SRMR* (table 2), this alteration was significant. When establishing partial structural invariances according to the MIs, it became evident that the latent factor correlations in the 7ALLN and 7END groups differed from those in the other groups (as was also observed with the single CFAs).

When the factor loadings were restricted to be equal in the model 1 (model 3, metric), a relatively poor goodness-of fit with respect to χ^2 and *CFI* was obtained. This model differed significantly from the configural model 1 (table 2, model difference test), according to the change in all goodness-of-fit statistics. To establish partial invariance, numerous differences in the factor loadings among the groups were modeled[2].

In the next model, equal intercepts were modeled, using model 1 as the baseline. The resulting model 4 (scalar invariance) achieved a very poor model fit with respect to all goodness-of-fit statistics. This change of the goodness-of-fit was significant. Allowing for the differences in numerous intercepts among

2 The detailed results for all partial invariance models can be obtained on request from the correspondence author.

the groups, as suggested by the MIs, significantly improved the goodness-of-fit of the resulting partial equivalency model.

The results of the measurement equivalence tests show that there were significant differences between the eight groups with respect to all parameters we tested: factor correlations, factor loadings, and items' intercepts. The results of single CFAs and the model comparison test did, therefore, not really support the factorial validity assumptions in most of the five category groups (except the 5END rating scale) and in the two groups with seven categories (7ALLN and 7END). It seems that the oblique factor structure cannot be assumed in the 7ALLN or 7END groups, because independent PA and NA factors were observed here, as shown in table 1.

Reliability

Reliability coefficients (Composite Reliability ρ and Guttman's λ_4) are presented in table 3. The coefficient ρ for PA is below .70 (which is often used as a benchmark for relatively homogeneous measures) in most groups with five categories (5ALLN, 5ENDN, and 5ALL), compared with all seven category and the 5END groups. When considering CIs, one can conclude that reliability is higher in the 7ENDN and 7ALL groups than in the 5ALLN and 5ENDN groups. For NA, ρ is lower than .70 in the 7END group while, in other groups, it is approximately .70 or even higher. The value of ρ for NA is the highest in 7ALL and 5END groups, which also was significantly higher than in the 7END group, in terms of the corresponding CIs.

With the Guttman's λ_4, low values were observed in the 5ALL group for both PA and NA items and in the 7END group for NA items. Although there were some differences between these two reliability coefficients, relatively comparable and reasonable reliability values were found in the 5END, 7ENDN, and 7ALL groups when using both reliability estimation methods.

Table 3
Reliability coefficients in the eight experimental groups

	5ALLN	5ENDN	5ALL	5END	7ALLN	7ENDN	7ALL	7END
Composite Reliability ρ (95% CI)								
PA	.64	.56	.58	.74	.70	.80	.76	.74
	.51–.77	.42–.70	.38–.79	.63–.84	.58–.83	.72–.87	.68–.84	.64–.84
NA	.68	.72	.65	.76	.75	.68	.77	.61
	.59–.77	.63–.81	.53–.76	.70–.84	.67–.83	.58–.77	.70–.84	.50–.71
Guttman's λ_4								
PA	.66	.69	.59	.78	.65	.80	.77	.78
NA	.73	.74	.63	.74	.72	.73	.80	.56

Note. Composite reliability was calculated on the basis of a one-dimensional CTT model for each factor, PA and NA, separately (see text). One correlated error term was included (PA: 5ENDN, 5ALL, 7ALLN, 7END; NA: 7END).

Validity

To assess convergent and divergent validity, we analyzed the differences between the groups with respect to the correlations of the latent PA and NA variables with summarized values of Extraversion (E) and Neuroticism (N). For the convergent validity, positive correlations were expected between PA and E and between NA and N. For the divergent validity, negative or zero correlations were expected between PA and N and between NA and E.

The validity was assessed by means of MGCFAs. First, correlations of N and E were included in the configural model (MGCFA model 1, table 2), because this model was found to be the most tenable one in the previous analysis. However, including both N and E in one model was not sufficient with respect to the model fit ($\chi^2_{(df=254)} = 603.962$; $p < .001$; *RMSEA* = .12; *CFI* = .82; *SRMR* = .702). Therefore, we estimated two different models: the first included correlations of PA and NA with E and the second included such correlations with N. Both models are associated with an acceptable model fit (*RMSEA* = .06; *CFI* = .95; *SRMR* = .06).

The correlations of PA and NA with E and N and their CIs, obtained as parameters from the MGCFA models, are presented in figures 1 and 2. When inspecting the correlations of PA with E (figure 1), one can observe that these are, as expected, positive in almost all rating scale groups. One exception is the 7END group, where the correlation is different from that in other groups due to differences in the CIs. It also tends towards zero and is not significant (p = 0.578). Looking at those correlations and their CIs more closely reveals that their values are higher in almost all five category rating scale groups (except 5ALL) than in the case of seven category groups.

With respect to the convergent validity of NA (displayed by the values of correlations with N in figure 2), it is evident that the 7ENDN and 7END groups displayed significantly lower values than those in other groups (where the correlations are also non-significant).

With respect to the divergent validity of PA obtained by means of correlations with N (figure 1), we observed that they were negative, or tended towards zero, in all rating scale groups apart from the 7ENDN group. The correlation between the PA and N was positive, but not significant in the 7ENDN group. Similar relationships were observed for the correlation of NA with E (divergent validity of NA, figure 2).

Summing up, assumptions with respect to convergent and/or divergent validity could not be confirmed in the 7ENDN and 7END group. For PA, the relationships between the variables were more strongly pronounced for five category scales (except for the fully verbalized form without numbers) than for seven categories.

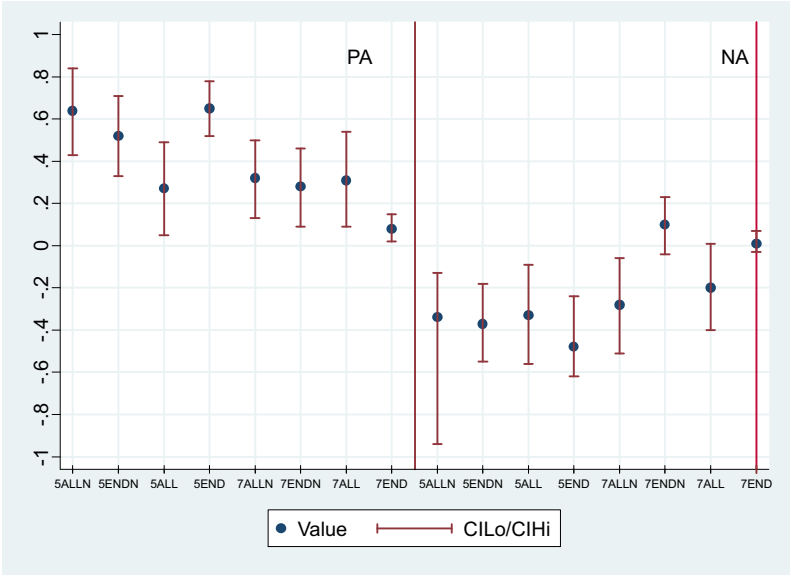

Figure 1. Convergent and discriminant validity: correlations of PA and NA latent factor values with Extraversion (E), with 95% Confidence Interval (CI).

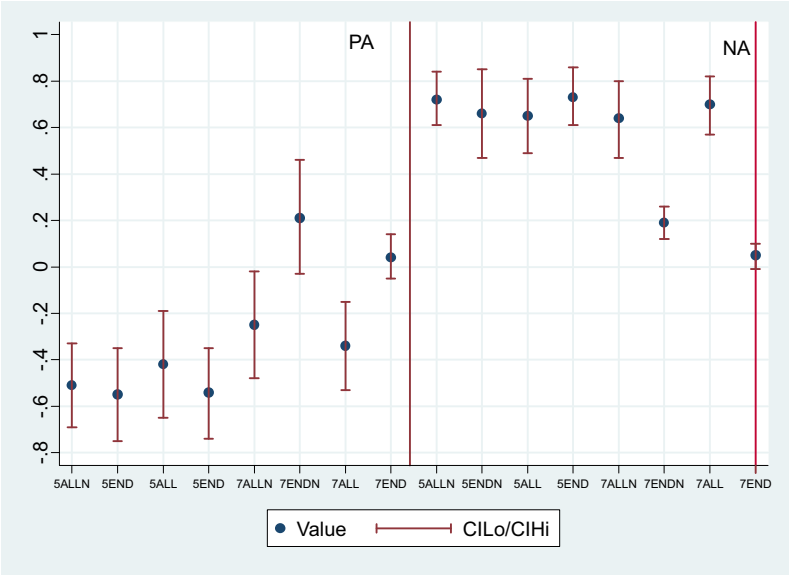

Figure 2. Convergent and discriminant validity: correlations of PA and NA latent factor values with Neuroticism (N), with 95% Confidence Interval (CI).

Discussion

The aim of the present study was to investigate how rating scale formats affect the results of latent variable measurement in web surveys, using a short version of PANAS as an example.

Our first finding is that the results are affected by rating scale groups when testing measurement model assumptions related to unidimensional CTT models. A lack of a substantive relationship with the latent variables, or poor model fit, was observed in the five category groups with either verbal or numeric labels (5ALLN, 5ENDN, 5ALL), and in seven categories with both numeric and verbal labels (7ALLN), or without any labels for intermediate categories (7END). Similar findings were reported by Menold and Tausch (2015), who addressed agreement rating scales in paper-and-pencil surveys with students. These findings are relevant, because the results of measurement model test have implications for measurement-model-based reliability assessment.

Then next finding of the study is that factorial validity depends on rating scale formats. Considering the results of the MGCFA and the single CFAs, experimentally manipulated rating scale formats affect factor loadings and between-factor correlations. For the formats 7ALLN and 7END, the PA and NA latent factors were not correlated, whilst correlations were significant in other rating scale groups. Because the oblique structure of PANAS could also not be confirmed by other investigators, the role of rating scale formats should be considered as a possible explanation for these differences in the PANAS structure, reported in the literature (Leue & Beauducel, 2011). The differences in loadings showed that the respondents understood the rating scale categories differently and even the items, when different rating scale formats are used. Differences in the intercepts showed that the items did not function equally in the experimental groups, leading to the differences in measurement bias (Steinmetz, 2015).

A further finding is that the two formats with five categories (5ENDN and 5ALL), in particular, as well as the 7END format, had lower reliabilities than the other groups. Additionally, the 7END and 7ENDN formats were associated with a low convergent validity.

Our results imply that researchers using multi-item self-report measures should be aware that changing ratings scale formats for pragmatic reasons may have serious consequences for understanding not only the response categories of the ratings scales but also of the items themselves. The differences in psychometric quality may then be the result of these differences in the understanding of both rating scale categories and items.

Evaluating hypotheses with respect to factorial and other aspects of construct validity has important implications for the assumptions of theoretical construct definitions as well as for the development and modification of theories. These should preferably be performed on the results, which are unaffected by a potential bias of rating scale formats. Therefore, it is very important to find rating scale formats that are associated with the smallest possible effects on factorial and other kinds of construct validity.

Our results have further implications for the choice of ratings scale formats. With respect to labeling, the predominant position taken in the literature suggests that an optimal measurement quality can be reached using the ALL format (e.g., Krosnick & Fabrigar, 1997; Toepoel & Dillman, 2010). The results of the present study, which relate to frequency rating scales, show that optimal quality of the ALL format may depend on the number of categories. In the ALL format with seven categories, reasonable results concerning all of the metrics we tested could be obtained, but not in the ALL format with five categories. However, an explanation for this result could be that there was not a unique meaning of the end categories in the five category format, because "never/ nearly never" and "very often/always" were used in one category (one should remember that researchers often use such combinations of labels). Therefore, it would be interesting to continue this research for frequency scales when using end categories with unique meanings. Nevertheless, we found reasonable results with respect to the all measurement quality metrics we evaluated for the 5END format. We suggest, therefore, for measures like those used here selecting the 5END format over the 5ALL format, but the 7ALL format over the 7END format. As far as agreement rating scales used in a different survey mode are concerned, Menold and Tausch (2015) found that the 7ALL format, at least, decreased the heterogeneity of items and increased the composite reliability. Bringing the results from this and other publications (Weng, 2004) and those of the present study together, the results with regard to the 7ALL format can, instead, be considered to be generalizable across different contents of items, types of rating scales, and modes of data collection.

The position that numerical labels decrease psychometric measurement quality (Krosnick & Fabrigar, 1997) was confirmed in the present study. When we used numeric labels, either measurement model test or other criteria, such as reliability, or factorial or divergent validity were negatively affected. The results indicate that numerical and verbal labels are understood differently by the respondents so that, in particular, combining both numerical and verbal labels in one rating scale may impede factorial validity, as evident in the 5ALLN and 7ALLN formats. Therefore, we suggest that researchers should be careful with the usage of these formats, because they are burdensome for respondents (Christian et al., 2009). We assume, from our results, that numerical and verbal labels may convey incongruent information regarding the meaning of ranges and sub-ranges of the rating scale. Therefore, error variance may be increased when some respondents use the verbal labels to map their responses and others use the numerical labels.

The randomized, between-group design we used is the strongest design in terms of the causal inference of experimental manipulation on the obtained results (e.g., Shadish, Cook, & Campbell, 2001), compared to studies that use a Structural Equation based MTMM design to obtain both reliability and validity of measures in surveys (e.g., Saris & Gallhofer, 2007).

One limitation of the study that requires discussion is the use of an abbreviated version of the PANAS. On the one hand, the effect of rating scale

formats could be more distinctive in the case of short vs. long item multi-item sets. On the other hand, the current results show that psychometric measurement quality of short item sets can be increased by the choice of rating scales. Short item sets allow researchers to collect data for many constructs while keeping questionnaires short, thus fostering survey participation and decreasing survey costs (Ziegler et al., 2014). Nevertheless, it is important to replicate our results using a long version of the PANAS. Further, we used frequency ratings in the current study. Whether the results found here apply to other commonly used ratings (e.g., agreement) requires further investigation. Finally, one should keep in mind that our results apply to the data collected in a web survey. Conducting such a research in this context is very important, since only few studies have addressed the issue of the impact of rating scale formats on psychometric measurement quality in web surveys, to date.

In conclusion, our results imply that researchers should carefully consider rating scale formats when developing and using questionnaires, because certain rating scales may be a serious source of heterogeneity that may impact on psychometric quality when measuring latent variables.

References

Alwin, D. F., & Krosnick, J.A. (1991). The Reliability of Survey Attitude Measurement: The Influence of Question and Response Attributes. *Sociological Methods and Research, 20,* 139-181.

Andrews, F. M. (1984). Construct Validity and Error Components of Survey Measures: A Structural Equation Approach. *Public Opinion Quarterly, 48*, 409-448.

Bartlett, M. S. (1950). Tests of significance in factor analysis. *British Journal of Psychology, Statistical Section, 3,* 77-85.

Beauducel, A., & Wittmann, W. W. (2005). Simulation Study on Fit Indices in Confirmatory Factor Analysis Based on Data with Slightly Distorted Sample Structure. *Structural Equation Modelling, 12*, 41-47. doi: 10.1207/s15328007sem1201_3

Bentler, P. M. (1995). *EQS structural equations program manual.* Encino, CA: Multivariate Software.

Bollen, K. A. (1989). *Structural Equations with Latent Variables*. New York: Wiley.

Byrne, B. (2011). *Structural Equation Modeling with Mplus: Basic Concepts, Applications, and Programming* (Multivariate Applications). London: Taylor & Francis.

Chen, F.F. (2007). Sensitivity of Goodness of Fit Indexes to Lack of Measurement Invariance. *Structural Equation Modeling, 14*(3), 464-504. doi: 10.1080/10705510701301834.

Cheung, G. W., & Rensvold, R. B. (2002). Evaluating Goodness-of-Fit Indexes for Testing Measurement Invariance. *Structural Equation Modeling, 2*, 233-255. doi: 10.1207/S15328007SEM0902_5.

Christian, L. M., Parson, N., & Dillmann, D. A. (2009). Designing Scalar Questions for Web Surveys. S*ociological Methods and Research, 37*, 393-425.

Churchill, G. A. Jr., & Peter, J. P. (1984). Research Design Effects on the Reliability of Rating Scales: A Meta-Analysis. *Journal of Marketing Research, 21*, 360-375.

Cronbach, L. J. (1951). Coefficient Alpha and the Internal Structure of Tests. *Psychometrika*, 16, 297-334.

Cronbach, L. J., & Meehl, P. E. (1955). Construct validity in psychological tests. *Psychological bulletin, 52*(4), 281-302.

Egloff, B., Schmukle, S. C., Burns, L. R., Kohlmann, C. W., & Hock, M. (2003). Facets of dynamic positive affect: differentiating joy, interest, and activation in the positive and negative affect schedule (PANAS). *Journal of Personality and Social Psychology, 85*(3), 528. doi: 10.1037/0022-3514.85.3.528.

Guttman, L. (1945). A basis for analyzing test-retest reliability. *Psychometrika, 10*, 255-281.

Hayduk, L. A. (1989). *Structural Equation Modeling – Essentials and Advances*. Baltimore and London: The John Hopkins University Press.

Hu, L., & Bentler, P. M. (1999). Cut-off Criteria for Fit Indexes in Covariance Structure Analysis: Conventional Criteria Versus New Alternatives. *Structural Equation Modeling, 6,* 1-55.

Kemper, C. J., Beierlein, C., Kovaleva, A., & Rammstedt, B. (2013). Entwicklung und Validierung einer ultrakurzen Operationalisierung des Konstrukts Optimismus-Pessimismus – Die Skala Optimismus-Pessimismus-2 (SOP2). *Diagnostica, 59*(3), *119-129.*

Krohne, H. W., Egloff, B., Kohlmann, C.-W., & Tausch, A. (1996). Untersuchung mit einer deutschen Form der Positive and Negative Affect Schedule (PANAS). *Diagnostica, 42,* 139-156.

Krosnick, J. A. (2011). Experiments for Evaluating Survey Questions. In J. Madans, K. Miller, A. Maitland, & G. Willis. *Question Evaluation Methods. Contribution to the Science of Data Quality* (pp. 215-238). New Jersey: Wiley.

Krosnick, J. A., & Berent, M. K. (1993). Comparison of Party Identification and Policy Preferences: The Impact of Survey Question Format. *American Journal of Political Science, 37,* 941-964.

Krosnick, J. A., & Fabrigar, L. R. (1997). Designing Rating Scales for Effective Measurement in Surveys. In L. Lyberg, P. Biemer, M. Collins, E. de Leeuw, C. Dippo, N. Schwarz, & D. Trewined (Eds). *Survey Measurement and Process Quality* (pp. 141-164). New York: Wiley.

Leue, A., & Beauducel, A. (2011). The PANAS Structure Revisited: On the Validity of a Bifactor Model in Community and Forensic Samples. *Psychological Assessment 23*(1), 215-225. doi: 10.1037/a0021400

Leue, A., & Lange, S. (2011). Reliability Generalization An Examination of the Positive Affect and Negative Affect Schedule. *Assessment, 18*(4), 487-501. doi: 1073191110374917.

Lord, F. M., & Novick, M. R. (1968). *Statistical theories of mental test scores.* Reading, MA: Addison-Wesley.

Maitland, A. (2009). How Many Scale Points Should I Include for Attitudinal Questions? *Survey Practice 06. AAPOR e-journal, 2*(5).

McDonald, R.P. (1999). *Test theory: A unified treatment*. Mahwah, NJ: Erlbaum.

Menold, N., Kaczmirek, L., Lenzner, T., & Neusar, A. (2014). How Do Respondents Attend to Verbal Labels in Rating Scales?. *Field Methods, 26*(1), 21-39. doi: 10.1177/1525822X13508270.

Menold, N., & Tausch, A. (2015). Measurement of Latent Variables with Different Rating Scales Testing Reliability and Measurement Equivalence by Varying the Verbalization and Number of Categories. *Sociological Methods & Research*, 0049124115583913.

Mullins, M. E., Polson, J. M., Lanch, T., & Kehoe, K. (2007). Respondent Perceptions of Integrity and Personality Measures: Does Response Format Make a Difference? *Applied H.R.M. Research, 11*(2)*,* 107-118.

Parducci, A. (1983). Category Ratings and the Relational Character of Judgment. In H. G. Geissler, H. F. J. M. Bulfart, E. L. H. Leeuwenberg, & V. Sarris (Eds.), *Modern Issues in Perception* (pp. 262-282). Berlin: VEB Deutscher Verlag der Wissenschaften.

Rammstedt, B., & John, O. P. (2007). Measuring Personality in One Minute or Less: A 10-item Short Version of the Big Five Inventory in English and German. *Journal of Research in Personality, 41*, 203-212.

Raykov, T. (1997). Estimation of Composite Reliability for Congeneric Measures. *Applied Psychological Measurement*, *21*(2) 173-184. doi: 10.1177/01466216970212006.

Raykov, T. (1998). On the Use of Confirmatory Factor Analysis in Personality Research. *Personality and Individual Differences, 24,* 291-293.

Raykov, T. (2012). Scale Construction and Development Using Structural Equation Modeling. In: R. H. Hoyle (ed.), *Handbook of Structural Equation Modeling* (pp. 472–492). New York: The Guilford Press.

Raykov, T., & Marcoulides, G. A. (2011). *Introduction to psychometric theory.* New York: Taylor & Francis.

Revilla, M. A., Saris, W. E., & Krosnick, J. A. (2014). Choosing the number of categories in agree–disagree scales. *Sociological Methods & Research 43*(1), 73-97. doi: 10.1177/0049124113509605..

Saris, W. E., & Gallhofer, I. N. (2007). *Design, Evaluation, and Analysis of Questionnaires for Survey Research*. Hoboken, New Jersey: John Wiley & Sons, Inc.

Shadish, W. R., Cook, T. D., & Campbell, D. T. (2001). *Experimental and Quasi-Experimental Designs for Generalized Causal Inference.* Stamford: Cengage Learning.

Steinmetz, H. (2015). Analyzing observed composite differences across groups. *Methodology, 9*, 1-12. doi: 10.1027/1614-2241/a000049.

Toepoel, V., & Dillman D.A. (2010). Words, numbers, and visual heuristics in web surveys. Is there a hierarchy of importance? *Social Science Computer Review 29*(2), 193-207. doi: 10.1177/0894439310370070.

Tourangeau, R., Couper, M. P., Conrad, F. G. (2007). Color, Labels, and Interpretive Heuristics for Response Scales. *Public Opinion Quarterly, 71*(1). 91-112. doi: 10.1093/poq/nfl046

Wakita, T., Ueshima, N., & Noguchi, H. (2012). Psychological Distance Between Categories in the Likert Scale: Comparing Different Numbers of Options. *Educational and Psychological Measurement, 72,* 533-546, doi: 10.1177/0013164411431162.

Watson, D., & Clark, L. A. (1997). Measurement and mismeasurement of mood: Recurrent and emergent issues. *Journal of Personality and Assessment, 68,* 267-296. doi: 10.1207/s15327752jpa6802_4.

Watson, D., Clark, L. A., & Tellegen, A. (1988). Development and validation of brief measures of positive and negative affect: The PANAS scales. *Journal of Personality and Social Psychology, 54*, 1063-1070. doi: 10.1037/0022-3514.54.6.1063

Wedell, D. H., Parducci, A., & Lane, M. (1990). Reducing the Dependence of Clinical Judgment on the Immediate Context: Effects of Number of Categories and Type of Anchors. *Journal of Personality and Social Psychology, 58,* 319-329.

Weng, L. (2004). Impact of the Number of Response Categories and Anchor Labels on Coefficient Alpha and Test-Retest Reliability. *Educational and Psychological Measurement, 64,* 956-972. doi: 10.1177/0013164404268674

Ziegler, M., Kemper, C. J., & Kruyen, P. (2014). Short scales –Five misunderstandings and ways to overcome them. *Journal of Individual Differences, 35*, 185–189. doi: 10.1027/1614-0001/a000148.

Appendix

Presentation of the PANAs items in the survey, group 5 ALLN

Tagtäglich erleben Menschen verschiedene Gefühle. Manche Gefühle empfinden wir nur selten, andere kommen gelegentlich vor, und wieder andere erleben wir oft.

Uns ist es wichtig zu erfahren, wie häufig Sie ein Gefühl im Allgemeinen erleben.

Wie häufig fühlen Sie sich im Allgemeinen…	nie oder fast nie	selten	gele-gentlich	oft	fast immer oder immer
1. Aktiv	□1	□2	□3	□4	□5
2. Nervös	□1	□2	□3	□4	□5
3. Begeistert	□1	□2	□3	□4	□5
4. Verärgert	□1	□2	□3	□4	□5
5. Interessiert	□1	□2	□3	□4	□5
6. Bekümmert	□1	□2	□3	□4	□5
7. Entschlossen	□1	□2	□3	□4	□5
8. Ängstlich	□1	□2	□3	□4	□5

Translation of the instruction:

People normally experience a variety of emotions. We experience some of them only seldom, some of them sometimes and some of them more often.

We would like to know how often, in general, you experience a certain emotion.

How often do you feel, in general, …

Translation for items and rating scale: see method section of the article.