

THE INFLUENCE OF THE RESPONSE FORMAT IN A PERSONALITY QUESTIONNAIRE: AN ANALYSIS OF A DICHOTOMOUS, A LIKERT-TYPE, AND A VISUAL ANALOGUE SCALE

SVEN HILBERT
HELMUT KÜCHENHOFF
NINA SARUBIN
LUDWIG-MAXIMILIANS-UNIVERSITY

TRISTAN TOYO NAKAGAWA
UNIVERSITY POMPEU FABRA

MARKUS BÜHNER
LUDWIG-MAXIMILIANS-UNIVERSITY

In the present study, 866 participants completed a questionnaire on the personality facet “dutifulness.” A dichotomous (DS), a 5-point Likert-type (LTS), and a 100-mm visual analogue scale (VAS) were analyzed regarding their effects on psychometric properties in a repeated measures design. Concerning estimates of reliability, it was shown that Cronbach’s alpha increased with the number of response alternatives of the scales, while McDonald’s omega and validity coefficients remained steady. It is argued that, given the τ -congeneric measurement model of the present study, omega provides the more adequate estimate of internal consistency. Generalized linear regression analyses indicated that the participants judged the intervals between varying levels of dutifulness on the VAS differently from the other two scales. It is concluded that the choice of response format should not be exclusively based on desired psychometric properties but rather on practical considerations.

Key words: Response format; Reliability; Validity; Questionnaire; Personality assessment.

Correspondence concerning this article should be addressed to Sven Hilbert, Department of Psychology, Psychological Methods and Assessment, Ludwig-Maximilians-University, Leopoldstraße 13, 80802 München, Germany. Email: sven.hilbert@psy.lmu.de

In psychology, questionnaires are the most frequently used measure to obtain information about inter- and intrapersonal differences. They are commonly answered with response scales, which are defined by various types of response formats. Among several aspects, response formats are an important element to be considered during the construction of psychological questionnaires. Therefore, response formats in general have been subject to extensive research ever since the beginning of the widespread use of questionnaires in psychology (see Cox III, 1980; Peter, 1979).

In particular, the optimal number of response categories has been thoroughly investigated (e.g., Dawes, 2008; Maydeu-Olivares, Kramp, García-Forero, Gallardo-Pujol, & Coffman, 2009; Preston & Colman, 2000) as it yields important implications for the psychometric properties of a

questionnaire: reliability (e.g., Birkett, 1986; Oaster, 1989; Weng, 2004) as well as validity (e.g., Hancock & Klockars, 1991; Matell & Jacoby, 1971) have been found to be strongly affected by the number of alternatives given on a response scale. Also, response scales have to be suitable for the given subject of investigation as well as for the target group: a scale differentiating too little or too much for a given topic or target-group can lead to inadequate response tendencies (Briggs & Closs, 1999; Faulbaum, Prüfer, & Rexroth, 2009). Response scales comprising a fixed number of response categories are typically defined as Likert-type scales (LTSs; Likert, 1932).

PSYCHOMETRIC PROPERTIES OF RESPONSE FORMATS WITH FIXED INTERVALS

Oaster (1989) used the Texas Social Behavior Inventory (Helmreich & Stapp, 1974) questionnaire with three to nine response categories to examine the effect of the number of alternatives per choice point on the temporal stability (retest reliability) of the scales. Oaster's results showed an increase in retest reliability along with an increasing number of response alternatives. The finding is in line with previous research on the intraclass reliability of LTSs (e.g., Cicchetti, Shoinralter, & Tyrer, 1985; Finn, 1972) but has been challenged by a recent study on surveys (Drake, Hargraves, Lloyd, Gallagher, & Cleary, 2014).

Weng (2004) found a steady increase in internal consistency from two to nine response categories, in line with various previous studies (e.g., Birkett, 1986). This was also suggested by a Monte Carlo simulation study by Lozano, García-Cueto, and Muñiz (2008). However, the increase in reliability of scales with growing response possibilities need not be infinite: Birkett reported a 6-point scale to result in the highest internal consistency when comparing it to a dichotomous scale (DS) and a 14-point scale.

Preston and Colman (2000) investigated the influence of the scale on convergent and criterion-related validity. While criterion-related validity did not differ significantly between the scales, convergent validity tended to be larger with an increasing number of response possibilities, reaching its peak at seven points. In order to investigate the effect of response formats on criterion-related validity, Hancock and Klockars (1991) compared formats with 5-point and 9-point scales, observing higher coefficients for the latter.

Notably, Matell and Jacoby (1971) did not find a significant relationship between the number of response categories and intraclass reliability or criterion-related validity. However, in their study, they compared response scales between two and 19 points, resulting in 18 different groups, each comprising only 20 subjects. The absence of a significant difference in reliability or validity between the groups is therefore likely to be caused by a lack of statistical power.

Probably the — psychometrically — most exhaustive analysis of the relationship between response formats and psychometric properties of scales was conducted by Maydeu-Olivares et al. (2009), who compared the consequences of using different response formats within the frameworks of classical test theory (CTT), item factor analysis (IFA), and item response theory (IRT). For their investigations, they used questionnaires with two, three, and five response categories. In the CTT framework, coefficient alpha (Cronbach, 1951), the most widely applied estimate of internal consistency (Graham, 2006), was used to estimate the internal consistency and showed increasing estimates with more response categories. In the other two frameworks, coefficient omega (McDonald, 1970) was used and showed no significant variation associated with the number of response categories, in both the IFA and the IRT frameworks. In addition,

discriminant and convergent validities remained widely unaffected by the number of response alternatives in all frameworks.

Taken together, the optimal number of response categories — with regards to psychometrical properties of the scale — seems to lie around seven. The sketchy picture of results may well be a product of the innumerable influences on the psychometric properties of questionnaires and differences in measurement error, which tend to be higher for response formats with many response possibilities (Goggin & Stoker, 2014; Shulman & Boster, 2014).

PSYCHOMETRIC PROPERTIES OF RESPONSE SCALES WITHOUT FIXED INTERVALS

In addition to LTSs with varying numbers of categories, visual analogue scales (VASs) are used to quantify responses in questionnaires. This response format has, however, received considerably less attention, which is not surprising since LTSs are the most commonly used response format in questionnaires (Bühner, 2011) and VASs are mostly used in clinical assessment (Wewers & Lowe, 1990) and online surveys (Couper, Tourangeau, Conrad, & Singer, 2006). In one of the few studies including LTSs and VASs, Flynn, van Schaik, and van Wersch (2004) compared a 65-mm VAS and a 7-point LTS for a questionnaire on psychological coping. Their results showed satisfying reliability coefficients for both scales, while the construct validity was higher for the VAS. Interestingly, the 7-point scale resulted in significantly higher mean levels of functional coping compared to the VAS, possibly due to a lack of experience of the participants with this response format, as the authors speculated. Grant et al. (1999) found the VAS superior to a 5-point LTS regarding retest reliability, as did Joyce, Zutshi, Hrubes, and Mason (1975). Jaeschke, Singer, and Guyatt (1990) found the criterion-related validity of a 7-point LTS and a VAS comparable. However, as noted by Jaeschke et al., the small sample size of only 20 patients limits the generalizability of this result. Notably, dichotomous response formats have, to the knowledge of the authors, never been compared to VASs.

Thus, findings regarding the VAS remain diverse, as do the methodological aspects of the aforementioned studies: psychological coping (Flynn et al., 2004), medical conditions (Grant et al., 1999; Joyce et al., 1975), and perceived quality of life (Jaeschke et al., 1990) are only some of the various measures investigated in the aforementioned studies. Also, several authors used only one item for the VAS (Grant et al., 1999; Joyce et al., 1975), hence providing very limited comparability to multiple-item instruments (Wewers & Lowe, 1990). Most authors stressed the importance of training and correct explanation of the use of VASs: a lack thereof may result in respondents' tendencies to use VASs in the same way as they use LTSs and not make use of the full range of the former (Couper et al., 2006; Ferrando, 2003).

ESTIMATION OF RELIABILITY BY MEANS OF INTERNAL CONSISTENCY

Importantly, the vast majority of studies used alpha as an estimate of reliability (in terms of internal consistency). Alpha, however, can only serve as an accurate estimate of reliability for an at least essentially τ -equivalent model, meaning that all items represent the latent variable to the same degree (i.e., equivalent factor loadings). Otherwise, alpha provides only an estimate of the minimal reliability (e.g., Graham, 2006; Socan, 2000). Most studies, however, investigate τ -congeneric mod-

els (i.e., models with differing factor loadings), in which case reliability is best estimated by omega (Graham, 2006). The difference between these two coefficients should therefore be taken into account when comparing the results of different studies regarding scale reliability.

The choice of the reliability coefficient is also influenced by the possible multidimensionality of the scale (see Peter & Churchill, 1986). Socan (2000) argued that very few scales in the social sciences are truly unidimensional. This poses a problem for the calculation of alpha as well as omega, however less for the latter as multidimensionality may be modeled and taken into account (e.g., Dunn, Baguley, & Brunnsden, 2014).

PERCEIVED INTERVALS AND CENTERS OF SCALES

A fundamentally important aspect of response scales lies within the interpretation of the intervals and centers of a scale: intervals on the VAS can be chosen freely when responding to a question in terms of (dis-)agreement, whereas scales of the Likert-type comprise predefined spacing between the response categories, thereby theoretically representing a coarse measure of a latent metric scale (Clason & Dormody, 1994). On a DS, the interval between the two categories can be interpreted as the probability of choosing one of the two alternatives, usually represented by a logistic curve (Agresti, 2002). While the density distribution of VASs has been subject to investigation and been found comparable to the distribution on LTSs (Ferrando, 2003), a comparison between the intervals on scales with fixed categories and a VAS has, to the knowledge of the authors, not yet been investigated and would reveal important information about the perception of the intervals on different response formats.

RATIONALE

The present study, therefore, uses a repeated measures design (allowing for the analysis of convergent validity and regression analysis between scales) including a DS, a 5-point LTS, and a VAS to compare the response formats regarding their psychometric properties and scale intervals. In order to gain information about the use of a VAS in a personality questionnaire, the eight items measuring the facet “dutifulness” of the dimension “conscientiousness” in the German version of the NEO-PI-R¹ (Berth, Goldschmidt, Ostendorf, & Angleitner, 2006) are used. Criterion-related validity and coefficients of reliability are compared between the different response formats. Convergent validity is estimated through the correlations of the latent variables of the three scales. Criterion validity is estimated through the correlations of the scales with the number of unexcused absent days during the last year of high school. This criterion was chosen because absence from work (Conte & Jacobs, 2003) and school (Lounsbury, Steel, Loveland, & Gibson, 2004; MacCann, Duckworth, & Roberts, 2009) have been linked to conscientious behavior (see Rothman, 2001, for sociodemographic factors). Of all five facets of conscientiousness, dutifulness, in particular, has shown the numerically highest correlation with absent days from work (Judge, Martocchio, & Thoresen, 1997). Also, the perceived intervals and central points of the VAS are compared to the dichotomous and the 5-point scale to obtain information about the comparability of the response scales’ category thresholds and centers.

METHOD

Participants

Eight hundred sixty-six (577 female) university students were tested. The students were between 18 and 52 years of age (quartile 1 = 20; median = 22; quartile 3 = 24; range = 18-52) and had prior experience with personality questionnaires. The three biggest groups of participants were students of natural sciences (about 25%), economics (about 20%), and medicine (about 10%).

Procedure

The participants were tested under comparable conditions in a university laboratory in groups of up to 10 persons. They had to fill out a sheet containing information about age, gender, and field of study and, subsequently, one of the three testing batteries (i.e., DLV, LVD, or VDL). Before each of the three response scales were administered to the participants, a short introduction was given in written form and an example of the following response format was shown.

Measures

An adapted version of the eight questions measuring the facet “dutifulness” of the factor “conscientiousness” of the German version of the NEO-PI-R (Berth et al., 2006) served as questionnaire. It was chosen because it is of neutral content and easy to understand. Some of the original items were rephrased for the sake of clarity.² Three versions of the questionnaire were presented on three separate sheets of paper, each containing the same eight items but a different response scale (i.e., VAS, 5-point LTS, or DS).

The DS comprised simply two boxes marked with *ja* [yes] and *nein* [no] to indicate agreement or disagreement, respectively. The 5-point LTS comprised five boxes, marked with *völlig unzutreffend* [completely incorrect], *unzutreffend* [incorrect], *weder noch* [neither], *zutreffend* [correct], and *völlig zutreffend* [completely correct]. The VAS consisted of a 10-cm horizontal line, the left and right ends labeled with *trifft überhaupt nicht zu* [not at all correct] and *trifft voll und ganz zu* [entirely correct], respectively. The response had to be indicated with a vertical dash along the line. The order of the three sheets was randomized across participants, in order to counterbalance possible effects of the order of response format, resulting in three different test batteries: DLV, LVD, and VDL.³

Analyses

Unless stated otherwise, all analyses were conducted using the statistical software R (R Development Core Team, 2011). The reliability of each of the three scales was analyzed by alpha⁴ and omega in order to compare the results under the assumption of (essentially) τ -equivalent and τ -congeneric models, respectively. The formula for alpha includes the variances of items and scale:

$$\alpha = \frac{k}{k-1} \left(1 - \frac{\sum_{i=1}^k \sigma_{Y_i}^2}{\sigma_X^2} \right)$$

k denotes the number of items, so the first fraction corrects for the number of items in the scale. The $\sigma_{Y_i}^2$ stand for the variances of the items and σ_X^2 denotes the variance of the scale.

As shown in the following formula, omega is calculated using factor loadings and unique variances of the manifest variables in the model:

$$\omega = \frac{\left(\sum_{i=1}^k \lambda_i \right)^2}{\left(\sum_{i=1}^k \lambda_i \right)^2 + \sum_{i=1}^k (1 - h_i^2)}$$

k denotes the number of items (manifest variables) while the λ_i stand for the factor loadings on the general factor and the h_i^2 denote the unique variances of the k items (manifest variables).

Repeated measures analyses of variance (rmANOVAs) were conducted to test for differences in the overall score according to the different response scales as well as effects of test order. The individual means of the scales were compared using post hoc repeated measures t -tests. Criterion-related validity was calculated, using Pearson's r , between the questionnaires' sum scores and the criterion. F -tests were conducted to test for differences between variances.

Linear regression analyses were conducted to analyze the predictive value of variables (and their interaction effects) on the questionnaires' sum scores and the criterion. The relationships between the response patterns of the VAS, the LTS, and the DS were investigated by generalized linear regressions for ordinal (proportional odds regression) and binary (logistic regression) responses using the R package "VGAM" (Yee, 2010).

For the analysis of discriminant validity, the questionnaire was split into two scales, one comprising items not related to work (Items 2, 3, 4, 5, and 6), the other one comprising items related to work (Items 1, 7, 8), in order to create a multitrait-multimethod matrix (Campbell & Fiske, 1959). The resulting linear structural equation model (SEM) mimicked the model which Hsiao, Wu, and Yao (2014) used to construct a multitrait-multimethod matrix for the investigation of construct validity.

Finally, the overall structure of the test batteries was analyzed using the statistical software MPlus (Muthén & Muthén, 1998) to estimate SEMs using mean- and variance-adjusted weighted least squares estimates (WLSMV; Muthén, Du Toit, & Spisic, 1997), which allow including dichotomous variables in the model. When required, the significance levels were adjusted by applying Bonferroni correction to a total alpha-level of .05. All tests were run two-tailed. All estimated parameters differed significantly from 0 unless stated otherwise.

RESULTS

Effects of Test Order and Descriptives

Possible effects of test order (i.e., whether the questionnaire was in the first, second, or third position within the test batteries) were compared for each response format individually. No

significant difference between the mean scores was observed for any of the scales (all $p > .05$). The descriptive statistics for the scales of each response format can be seen in Table 1.

TABLE 1
Descriptive statistics response scales

Scale	<i>N</i>	<i>M</i>	Med	<i>SD</i>	Min	Max	Amount max	Amount min
VAS	866	184.00	604	102.49	184	800	3	0
LTS	866	23.61	24	3.69	9	32	8	0
DS	866	6.43	7	1.29	0	8	207	3

Note. *N* = sample size; Amount max = number of subjects with maximal score; Amount min = number of subjects with minimal score.

Comparison of Scale Means and Variances

Before comparing the means and variances, the scales were standardized to a possible range of 0-8 (because each scale comprised eight items) to receive comparable metrics to begin with. An rmANOVA revealed significant differences between the means of the scales, $F(2, 1730) = 957.40, p < .05$. Post hoc paired-sample *t*-tests showed that all of the mean sum scores differed significantly: the VAS showed a higher mean sum score than the LTS, $t(865) = 4.08, p < .01$; Hedges' $g = .08$, and the DS showed a higher mean sum score than both the VAS, $t(865) = 13.84, p < .01$; Hedges' $g = .39$, and the LTS, $t(865) = 16.95, p < .01$; Hedges' $g = .47$. Thus, on average, the respondents described themselves as more dutiful on the DS compared to the other two scales, and more dutiful on the VAS compared to the LTS.

The variances of the sum scores of the three scales were also tested for differences: the first comparison revealed a significant difference between the LTS and the DS, $F(1, 865) = 1.97, p < .01$. Also, the VAS and the DS, $F(1, 865) = 1.59, p < .01$, as well as the VAS and the LTS, $F(1, 865) = 1.23, p < .01$, differed significantly regarding the extent to which the sum scores vary.

Scale Structure and Convergent Validity

A linear SEM was fitted to investigate the structure of the construct dutifulness, as measured by the questionnaire, and the three response formats. First, a basic model was developed (Figure 1, solid lines), consisting of the measurement models of each scale with the according latent factors (dutifulness measured with the dichotomous response scale = D DICH, dutifulness measured with the ordinal LTS = D ORD, dutifulness measured with VAS = D VIS), the respective eight manifest variables and the item-uniqueness factors (Item 1-Item 8). The model yielded poor global model fit, $\chi^2(276) = 16598.14, p < .00$, as well as inappropriate fit index values, RMSEA = .075; 90% CI [.071, .079]; WRMR = 2.093; CFI = .933,⁵ and was, thus, modified. As depicted in Figure 1 (dashed lines), three correlations between item-uniqueness factors were added ($r_{\text{Item1, Item7}} = .67; r_{\text{Item4, Item5}} = .33; r_{\text{Item7, Item8}} = .29$). Variances of the item-uniqueness factors (Item 1-Item 8) were fixed to 1. As nonsignificant negative estimations of variances appeared, the

error variances of the manifest variables D2 and D7 were fixed to 0 and their respective paths were fixed to 1.

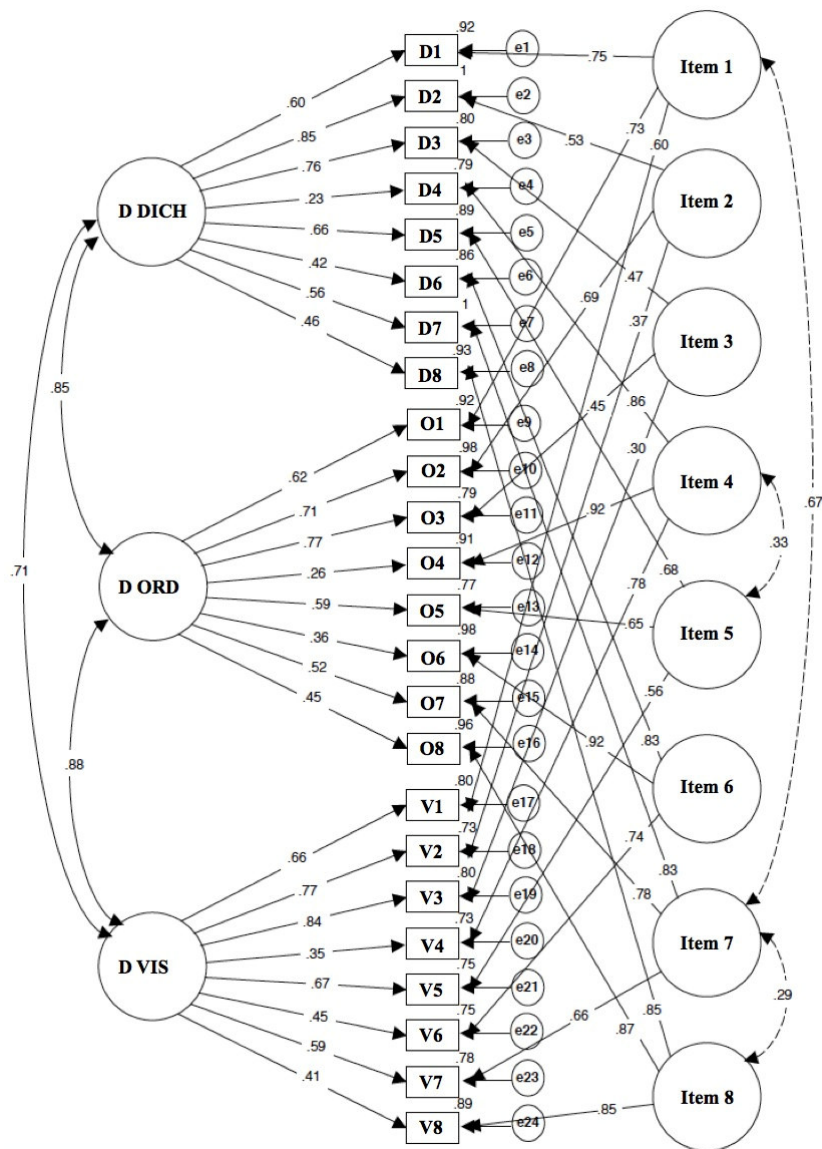


FIGURE 1

Modified structural equation model.

The model comprises three factors for dutifulness mixed measured with the respective response format and eight-item factors denoting the unique variances for each item.

D DICH = dichotomous scale; D ORD = Likert-type scale; D VIS = visual analogue scale.

According to the global model fit the modified model should have been rejected, $\chi^2(224) = 804.39, p < .00$, yet fit index values, RMSEA = .055; 90% CI [.051, .059]; WRMR = 1.52; CFI = .96, corresponded to the cutoff criteria recommended by Hu and Bentler (1999; RMSEA < .06; CFI > .95). Heene, Hilbert, Draxler, Ziegler, and Bühner (2011) showed that the fit indices decrease when factor loadings are lower than the ones used by Hu and Bentler. Yet, given that the

RMSEA and the SRMR only slightly exceeded the values for medium factor loadings and the CFI pointed toward an acceptable model fit and the values of Hu and Bentler as well as Heene et al. were derived for full maximum likelihood procedures, it was decided to regard the model as fitting well enough for the purpose of the study. The WRMR, on the other hand, clearly exceeded the cutoff value ($WRMR < 1$) suggested by Yu (2002). Yet, Yu stated that further research is needed regarding the WRMR and Muthén (2010) even recommended ignoring the WRMR, if all the other fit indices corresponded to their respective cutoff values. The resulting correlations between the response format factors ($r_{D\ VIS, D\ ORD} = .88$; $r_{D\ VIS, D\ DICH} = .71$; $r_{D\ ORD, D\ DICH} = .85$) were used as estimate for the convergent validity and proved to be high but far from perfect, despite the applied correction for attenuation. The correlations all differed significantly from each other (all $p < .05$).

Discriminant Validity

The questionnaire was split into two separate scales, including Items 2, 3, 4, 5, 6 and Items 1, 7, 8, respectively. Subsequently, a SEM was estimated, identical to the previous model but with two scales instead of one. Three latent variables (one for each response format) were estimated for each scale (in addition to the item uniqueness variables). The resulting correlations of the latent variables are listed in a multitrait-multimethod matrix, depicted in Table 2. Numerically, mean correlations were highest for the same scale measured with different response formats, lower for different scales measured with the same response format, and lowest for different scales measured with different response formats.

TABLE 2
 Multitrait-multimethod matrix

	DS 1	LTS 1	VAS 1	DS 2	LTS 2	VAS 2
DS 1	(.754)	.826	.682	.704	.487	.557
LTS 1		(.721)	.835	.487	.616	.620
VAS 1			(.777)	.490	.613	.747
DS 2				(.806)	.909	.807
LTS 2					(.778)	.927
VAS 2						(.748)

Note. DS = dichotomous scale; LTS = Likert-type scale; VAS = visual analogue scale; 1 = Scale 1 (Items 2, 3, 4, 5, 6); 2 = Scale 2 (Items 1, 7, 8). Coefficients in the main diagonal elements (in parentheses) = omega; coefficients in off-diagonal elements = correlations between latent variables, corrected for attenuation; bold-type coefficients = coefficients for different scales measured with the same response format; italic-type coefficients = coefficients for the same scale measured with a different response format.

Estimates of Internal Consistency and Criterion-Related Validity

Alpha and omega are depicted in Table 3 for each response format individually. Alpha increased monotonously with increasing number of response alternatives. Omega, on the other

hand, remained steady between the DS and the LTS, with a slightly higher value for the DS. The VAS, as for alpha, showed the highest estimate of reliability.

TABLE 3
Psychometric properties response scales

	DS	LTS	VAS
Alpha	.492 [.439, .543]	.691 [.655, .719]	.797 [.807, .843]
Omega	.808 [.749, .862]	.768 [.732, .795]	.817 [.783, .839]
Criterion-related validity	-.146; $p < .01$	-.148; $p < .01$	-.149; $p < .01$

Note. DS = dichotomous scale; LTS = Likert-type scale; VAS = visual analogue scale. Criterion-related validity = Pearson's r ; p = probability of committing a Type-I-Error; 95% confidence intervals in parentheses.

Finally, to measure the criterion-related validity, the sum scores of the scales were calculated and correlated with the recorded external criterion “number of unexcused absent days in the last school year before university.” The results indicated that all of the response scales correlated with the external criterion (all $p < .01$). Paired comparisons of the correlation strengths after Fisher transformations into z -values showed no significant differences (all $p > .05$).

Effect of Gender and Age

No significant differences for all response formats (all $p > .05$) were revealed by the t -tests comparing the mean dutifulness of female and male participants. Also, linear regression analyses showed that gender and age do not interact in the prediction of the total scores. Age, however, showed a significant positive relationship with the mean dutifulness scores. As depicted in Table 4, the relationship holds for all response formats.

To investigate the effect of gender, all estimated psychometric properties of the scales were also estimated separately for female and male participants. The estimates differed slightly from the coefficients estimated for the whole sample, as can be seen in Table 5, though the pattern of results remained essentially the same and the correlations between the latent variables — serving as indicators of convergent validity — did not differ significantly from the ones obtained with the whole sample (all $p > .05$). Also, for female as well as for male participants, criterion-related validity coefficients did not differ significantly, as indicated by differences in Fisher-transformed correlation coefficients (all $p > .05$).

To quantify the effect of age on convergent validity, a structural equation model including age as exogenous manifest material, associated with the three latent variables of the scales, was estimated. Age showed significant loadings on all three latent variables (DS = .668; LTS = .645; VAS = .595). Also, the inclusion led to a significant decrease in the correlations between the three latent variables, as indicated by the differences of the Fisher-transformed correlation coefficients (z -values): DS-LTS = .776 ($z = -4.589$, $p < .01$); DS-VAS = .569 ($z = -5.009$, $p < .01$); LTS-VAS = .841 ($z = -3.140$, $p < .01$). Concerning the effect of age on criterion-related validity, linear regression analyses for all three response formats including age and total scale score as predictors, showed no significant predictive value for age on the number of absent days (all $p > .05$).

TABLE 4
Regression scale sum scores on age and gender

	Estimate	SE	<i>t</i>	<i>p</i>
DS				
Intercept	5.393	.3001	17.936	< .001
Age	.048	.013	3.680	< .01
Sex	.301	.574	.524	<i>ns</i>
Age × Sex	-.021	.024	-.876	<i>ns</i>
LTS				
Intercept	20.869	.855	24.403	< .001
Age	.132	.037	3.556	< .01
Sex	1.763	1.632	1.080	<i>ns</i>
Age × Sex	-.112	.069	-1.615	<i>ns</i>
VAS				
Intercept	496.254	23.640	20.992	< .001
Age	4.816	1.030	4.676	< .001
Sex	60.662	45.128	1.344	<i>ns</i>
Age × Sex	-3.703	1.920	-1.928	<i>ns</i>

Note. DS = dichotomous scale; LTS = Likert-type scale; VAS = visual analogue scale; Estimate = unstandardized regression weight; SE = standard error of the regression weight; *t* = *t*-value of the regression weight; *p* = probability of committing a Type-I-Error.

TABLE 5
Psychometric properties for female and male participants separately

	DS	LTS	VAS
Female participants			
Alpha	.499 [.435, .559]	.712 [.675, .747]	.805 [.781, .829]
Omega	.833 [.773, .893]	.796 [.764, .828]	.827 [.792, .862]
Criterion-related validity	-.124; <i>p</i> < .01	-.120; <i>p</i> < .01	-.148; <i>p</i> < .01
Male participants			
Alpha	.488 [.394, .573]	.645 [.581, .704]	.779 [.739, .816]
Omega	.783 [.658, .908]	.722 [.657, .787]	.800 [.735, .864]
Criterion-related validity	-.172; <i>p</i> < .01	-.182; <i>p</i> < .01	-.133; <i>p</i> < .01

Note. DS = dichotomous scale; LTS = Likert-type scale; VAS = visual analogue scale. Criterion-related validity = Pearson's *r*; *p* = probability of committing a Type-I-Error; 95% confidence intervals in parentheses.

Regression from VAS to DS

The logistic regression function links the predicted response value (π) for the DS to the natural parameter (η) of the linear predictor of the VAS via:

$$\pi = \frac{\exp(\eta)}{1 + \exp(\eta)},$$

resulting in the logistic curves observed in Figure 2.

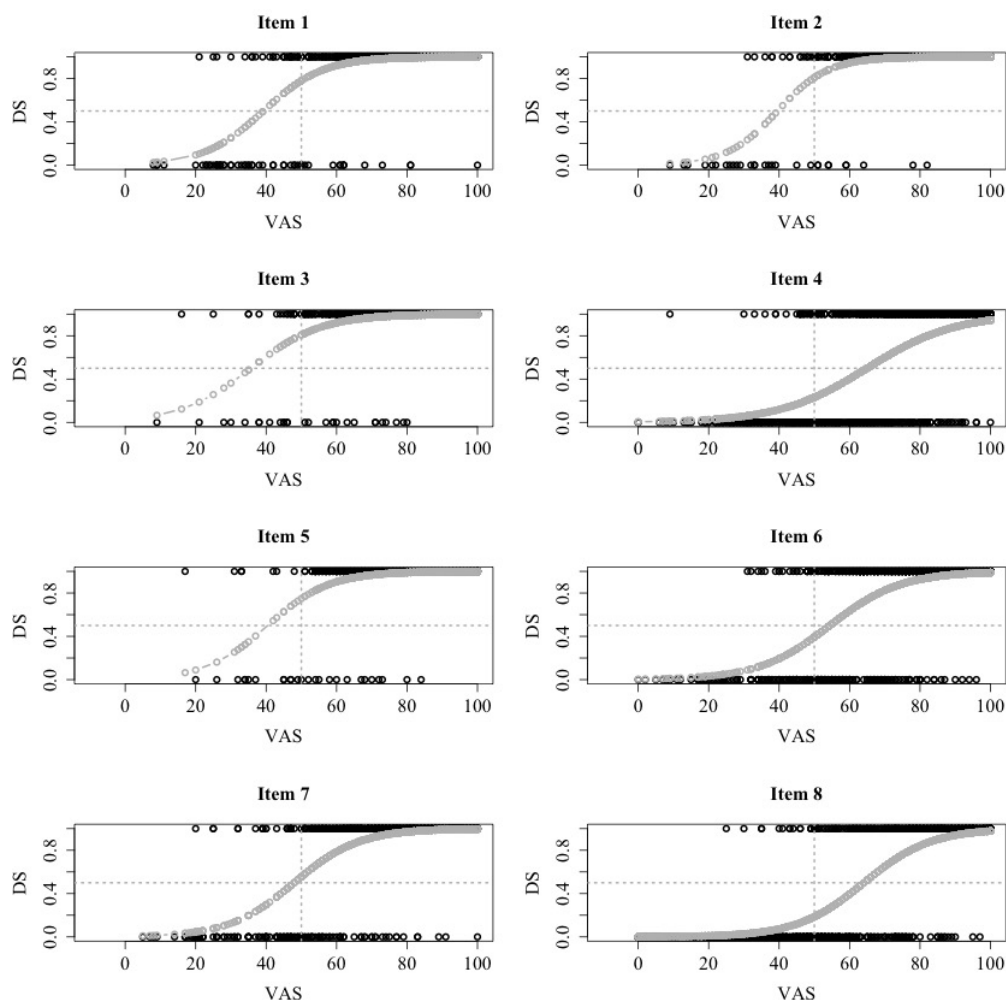


FIGURE 2

Regression dichotomous scale on visual analogue scale.

Black points = data points; Grey points = predicted values for the dichotomous scale (DS) for each data point of the visual analogue scale (VAS).

Table 6 shows the results of the logistic regression with the DS items as response variable and the respective VAS items as covariate. Figure 2 further illustrates the association between the responses for the two scales. The regression coefficients vary considerably between the individual items. The comparability of the scale centers between the formats are indicated by the point of inflection of the logistic curves: it denotes the point at which the estimated odds of choosing “yes” become higher than 50% in the DS. As shown in Figure 2, this point is shifted from the central point (50mm) of the VAS for most items.

TABLE 6
 Regression DS on VAS

Item	Intercept	Slope
1	-4.617	.118
2	-5.466	.138
3	-3.563	.100
4	-5.207	.080
5	-4.608	.114
6	-5.383	.099
7	-5.191	.108
8	-6.725	.105

Note. Intercept = intercept of the linear predictor of the logistic regression; Slope = slope of the linear predictor of the logistic regression.

Regression from VAS to LTS

Table 7 shows the regression coefficients of the proportional odds model with the LTS items as response variable and the respective VAS items as covariate. The estimated thresholds show the point at which the odds of choosing the respective category in the LTS become higher than the previous category. As graphically illustrated in Figure 3, the thresholds vary considerably between the items and are far from equidistant. Models with fixed equidistant thresholds were also fitted and resulted in poorer fit for every item, as tested with likelihood-ratio tests (all $p < .001$).

DISCUSSION

The present study investigated the relationship between a DS, a 5-point LTS, and a VAS. The three response scales showed a monotonous increase in alpha with an increase in response possibilities, while omega did not vary significantly. Also, criterion-related validity showed no differences between the formats. Reflecting convergent validity, the correlations between the scales were high but far from perfect — despite being corrected for attenuation. In a multitrait-multimethod matrix, correlation coefficients indicating discriminant validity showed the same pattern for all response formats. The response patterns between the VAS and the two scales with fixed categories proved to be related — however, notable differences were observed between the individual items of the scale.

While age had a significant effect on convergent validity and the sum score of the scales, gender did not affect either. Also, gender and age did not interact in the prediction of the total dutifulness score. This is in accordance with the results reported by Lehmann, Denissen, Allemand, and Penke (2013).

TABLE 7
Regression LTS on VAS

Item	Threshold 1	Threshold 2	Threshold 3	Threshold 4	Slope
1	–	4.697	7.982	14.816	.161
2	.092	4.197	6.433	11.903	.135
3	.234	3.279	6.172	11.428	.126
4	.058	4.398	7.272	10.984	.114
5	1.230	4.126	7.121	12.234	.139
6	–.485	4.920	7.811	11.897	.129
7	–.088	4.908	8.281	13.261	.144
8	1.138	5.167	8.056	11.490	.125

Note. Threshold = point at which the probability for the choice of the next category becomes higher than for the choice of the previous one; Intercept = intercept of the linear predictor of the logistic regression; Slope = slope of the linear predictor of the logistic regression.

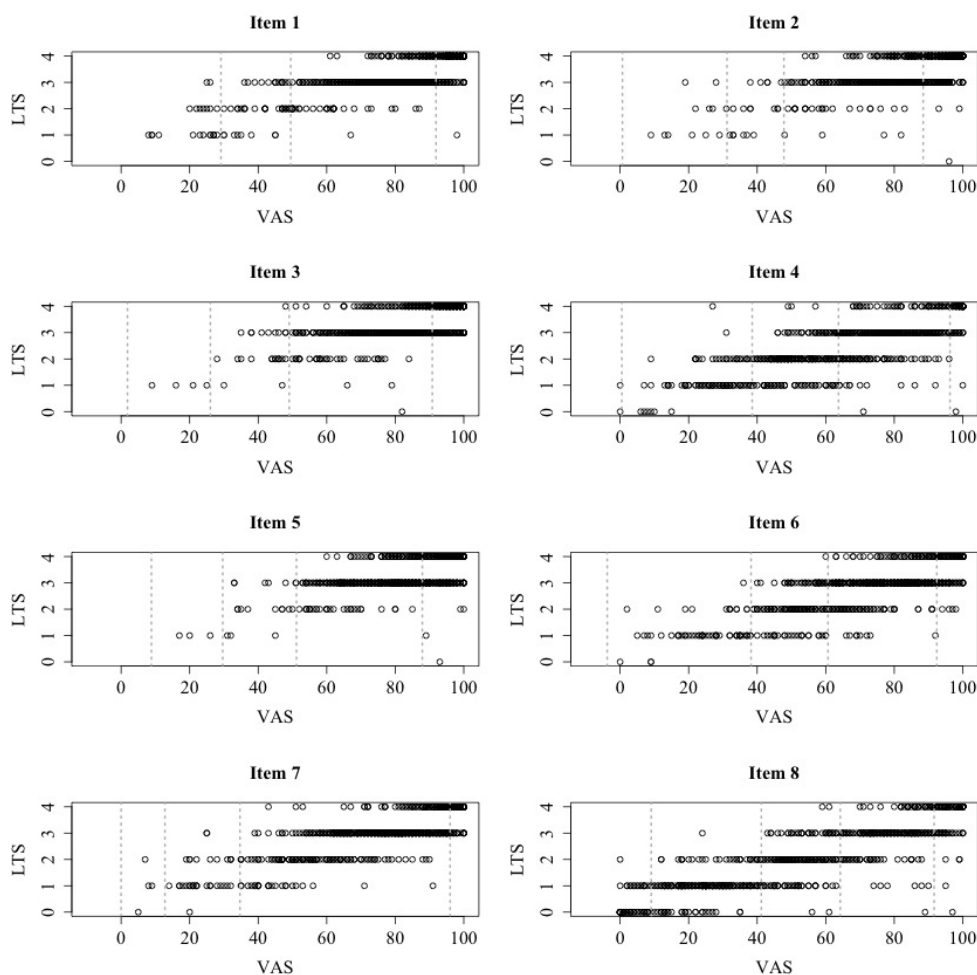


FIGURE 3
Regression Likert-type scale on visual analogue scale.
Black points = data points; Grey lines = estimated thresholds between the categories on the Likert-type scale (LTS); VAS = visual analogue scale.

Reliability and Criterion-Related Validity

The finding that alpha increased as a monotonous function of the level of differentiation of the response scale is in line with most studies concerning the relationship between response formats and reliability (e.g., Cicchetti et al., 1985; Finn, 1972; Oaster, 1989; Weng, 2004). Yet, omega did not vary between response formats, mimicking the findings of Maydeu-Olivares et al. (2009). The finding underlines the importance of the (assumed) measurement model and psychometrical framework: as, under the assumption of a τ -congeneric model, alpha is only an estimate of the minimal reliability (Graham, 2006; Socan, 2000), the discrepancy between the two coefficients clearly demonstrates how the careless use of alpha can be misleading regarding the reliability of a scale. Gignac (2014) illustrated how especially multidimensional scales lead to underestimations of reliability by alpha, which, considering the additional dimensions responsible for the correlations of Items 1, 7, and 8 or the correlation of Items 4 and 5, seems applicable to the current study.

An inspection of the formulas of alpha and omega illustrates this discrepancy: while alpha is calculated based on the variances of items and scales (which varied significantly in the present study, due to the different scales and even after standardizing the sum scores), omega uses the standardized factor loadings and unique variances of the items within the framework of a latent factor model in order to estimate the reliability of a scale (see Revelle & Zinbarg, 2009, for an exhaustive comparison). Since the factor loadings between (opposed to within) the three scales were comparable in the model fitted in the present study, the invariance of omega could be expected. Alpha, as it is based on variances of items and scales, underestimates the reliability (by means of internal consistency) of the scale if the factor loadings of the items differ: the difference in factor loadings indicates that the items measure the latent variable on a different scale and that, thus, their variances do not indicate the variance of the latent variable to the same degree (see Graham, 2006, for an illustration). Because the assumption of equal factor loadings within each scale is clearly not met in the present investigation, (essential) τ -equivalence cannot be assumed and the lower estimates of reliability by alpha compared to omega are not surprising. Finally, it has to be taken into account, that the WLSMV estimation of the factor loadings in the present study treats the LTS and the DS as ordered, therefore estimating a sequential model (Samejima, 1969), analogously to the IRT framework proposed by Maydeu-Olivares et al. (2009). The treatment of LTSs (and DSs) as ordinal has been widely suggested (e.g., Clason & Dormody, 1994; Goldstein & Hersen, 2000) and has been underlined by the nonequidistant category thresholds estimated by the regression analyses in the present investigation. In addition, possible multidimensionality of the scale, as indicated in the present study, can be taken into account for the estimation of internal consistency (e.g., Dunn et al., 2014) in order to obtain a more robust estimate of reliability — even though omega has been found to decrease under conditions of multidimensionality (Socan, 2000).

The finding that the criterion-related validity did not show an increase with the growing number of response possibilities is in accordance with the results of Preston and Colman (2000) as well as Jaeschke et al. (1989): the predictability of the number of missed days during the last year of high school did not differ between the response formats. Yet, dutifulness showed a rather low correlation with the criterion compared to previous findings (Judge et al., 1997). Nevertheless, the obtained results contradict the assumption that there is a definite gain in criterion-related

validity by the use of a 5-point LTS in comparison to a dichotomous format, as proposed by Preston and Colman as well as Hancock and Klockars (1991). However, Preston and Colman used a global rating of quality of service as the criterion for a restaurant-rating scale, while Hancock and Klockars related videotaped performances in a computer game to a performance scale. Obviously, the comparability between these criteria and the unexcused absent days during the last year of high school, which serve as criterion in the present study, appears rather restricted.

Model Structure and Convergent Validity

It was expected that the questionnaires would capture the same latent variable (namely “dutifulness”), regardless of the response scale. Yet, the correlations between the three latent factors proved to be high but far from perfect (especially when age was taken into account), despite the applied correction for attenuation. The observed correlations cast doubt on the convergent validity of the scale (see Bühner, 2011): despite the comparable psychometric quality, the questionnaires do not seem to capture the same information if the response format is changed. According to Schwarz (1999), response formats, in addition to providing a response to an item, play a part in making the item understandable to responders by hinting at the intention of the questionnaire’s developer. Schwarz’s view is strongly supported by the present results: identical items were presumably interpreted differently when the response format changed, at least concerning the response they elicited. Importantly, this finding suggests that, no matter how this questionnaire measures a trait, it does measure other dimensions that cannot be explained by random measurement error — the trait is therefore not measured purely. This issue needs to be stressed even further by the fact that the correlations between the scales differed significantly from each other, thereby indicating that additional dimensions are measured by the response formats. Maydeu-Olivares et al. (2009) reported no differences between the correlations of the latent variables representing the response formats. Their results, however, did not include a visual analogue scale which may additionally interact with the additional dimensions of the scales.

Further support for this assumption comes from the finding that the mean score differed significantly between all of the response formats. The finding is in accordance with the results of Guyatt, Townsend, Berman, and Keller (1987), who reported a significant higher mean score for a VAS compared to a LTS in a questionnaire on bodily functions. A clear ceiling effect is also observable for the DS with a total of 207 participants reaching the maximum score of eight compared to only three participants who scored 0 on this scale. In comparison, only three and eleven participants reached the maximum score in the VAS and the LTS, respectively.

Now, some attention has to be paid to the modifications applied to the basic model. Allowing additional correlations between the latent item factors seems justified since these factors represent the same questions.

Item 1 (“I fulfill my tasks conscientiously”) and Item 7 (“I work carefully”) capture nearly the same content because the words *gewissenhaft* [*conscientiously*] and *sorgfältig* [*carefully*] are often used as synonyms in German (Synonymwörterbuch, 2006). Also, both are related to the context of working, as is Item 8 (“Only serious illness prevents me from working”). Items 4 (“I always stick to the truth”) and 5 (“One can rely on my word”), also coincided regarding their context: both are related to sticking to one’s principles.

Even though the aforementioned correlations of the item factors make sense regarding the content of the items, it is clearly shown how easily the unidimensionality assumption of a scale is violated: despite careful rephrasing of some items for the sake of clarity and despite using a questionnaire comprising only eight items, unidimensionality is not given for the scale used in the present investigation.

Discriminant Validity

The multitrait-multimethod matrix, constructed in the present study, showed the pattern proposed by Campbell and Fiske (1959) for discriminant validity: numerically, the average correlations were highest for the same scale measured with different response formats, lower for different scales with the same response format, and lowest for different scales measured with different formats. The finding does imply discriminant validity for the two scales, and therefore clearly challenges the unidimensionality assumption for the questionnaire. Yet, the obtained pattern underlines the notion that different response formats can be regarded as different measurement methods and that the correlations between the traits can, thus, serve as a measure of convergent validity. Importantly, this multitrait-multimethod approach cannot be interpreted as investigating discriminant validity for the dutifulness scale as a whole, as this scale was divided into separate subscales.

Comparison of Scale Intervals and Centers

The regressions from the DS on the VAS showed strongly varying points of inflection for the resulting logistic functions between the eight items. The point of inflection of the logistic curve denotes the point at which, in the DS, it becomes more likely to agree with the statement displayed in the questionnaire than to disagree (i.e., value 1 becomes more likely than value 0, or “yes” becomes more likely than “no”). This threshold can be seen as the center of the dichotomous response scale and should, if the DS and the VAS elicit comparable responses, correspond to the center of the VAS (i.e., value 50), as is the case for Item 7. Most of the items (Items 1, 2, 3, and 5), however, show a regression function shifted to the left, meaning that respondents are more likely to select “yes” in the DS before they select a value on the VAS that is closer to “agreement” than to “disagreement” (i.e., higher than 50). For items that show a shift to the right (Items 4, 6, and 8), a value of 0 in the DS is predicted for VAS values higher than 50. It is therefore indicated that, on the individual item level, the perception of the interval between agreement and disagreement does not correspond in the two formats.

The regression from the LTS on the VAS also showed a rather sketchy pattern between the items: the estimated thresholds, denoting the point at which the odds for the choice of a category in the LTS exceed those of the previous one, differ strongly between the items regarding their position, distance, and ratio. This indicates that the association between perceived intervals of the VAS and the LTS cannot be generalized but depends on the corresponding item. The different sizes of the intervals between the thresholds, apparent in every individual item, are particularly interesting: respondents seem to judge the intervals between different levels of dutifulness

differently depending on the response format. The result suggests that the interval scale assumption does not hold for at least one of the two scales, since the thresholds should be equidistant if the interval scale assumption holds on both the LTS and the VAS. Most likely, however, this assumption does not hold for either of the scales, as argued by various authors (e.g., Clason & Dormody, 1994; Goldstein & Hersen, 2000; Wewers & Lowe, 1990). Still, the finding that a monotonous relationship between the thresholds of the LTS and the VAS can be assumed is in line with Ferrando (2003), who found the density distribution of the two scales to be comparable.

Taken together, the regression analyses illustrate how the perceived intervals and centers of the assumed latent scale differ strongly between the response formats. Also, the association between the scales cannot be assumed to be equal for the items.

Effect of Age

Age was positively associated with dutifulness on the level of sum scores as well as on the level of latent variables. The finding is in line with previous research on the personality dimension conscientiousness and especially its facets during the first decade of the 2000s (McCrae, Martin, & Costa, 2005), which comprises the vast majority of the samples in this study. It is therefore not surprising that the correlations between the latent variables representing dutifulness decreased after taking age into account.

Age was not associated with the unexcused absent days from school. This shows that the association between absence and dutifulness was not mediated by age and can be attributed to dutifulness as measured by the scales.

Limitations of the Study and Implications for Future Research

The application of the VAS needs further attention in future research. As already mentioned by several authors, this format seems to confuse part of the respondents. While both the DS and the LTS were directly understood by the participants, the VAS frequently provoked questions, despite the written explanation given at the top of the questionnaire. As proposed by Guyatt et al. (1987), it may be helpful to precede the assessment with a 10-15 minute period of teaching the use of the VAS to the respondents. An additional explanation might also shorten the extended completion times observed for the VAS and help to reduce the sometimes increased amount of missing data (Couper et al., 2006).

In addition, the validity-criterion used in the current study might not have grasped the whole spectrum of dutifulness, as it is restricted to unexcused absence in school. Since dutifulness includes aspects such as telling the truth and sticking to one's own principles, it is reasonable to assume that this criterion is not related to all parts of the construct. The correlation of the scales with the external criterion used here may therefore be merely seen as an estimation of the minimal criterion-related validity.

Despite showing a rather large age range, the main portion of the participants was of standard college age: only 3.5% of the sample were older than 30 years. The conclusions that can be drawn from the present investigation are, thus, limited to participants of an age range of about

10 years (i.e., between about 20 and 30 years). This is especially important since conscientiousness has been found to increase across the whole lifespan (e.g., Lehmann et al., 2013; Specht, Egloff, & Schmukle, 2011).

CONCLUSION

Due to its repeated measures design, the present study shed light on several important issues: first of all, psychometric quality in the personality questionnaire does not seem to increase with a more differentiated response scale — neither in terms of reliability (by means of internal consistency) nor in terms of convergent or (minimal) criterion-related validity. Even though alpha showed a significant increase with increasing response possibilities, omega (the more suitable coefficient for the given τ -congeneric model) did not. A significant portion of the correlation between the latent variables of the scales measured by the different response formats can be accounted for by age, which also significantly predicted the mean score in dutifulness.

For practical use, however, it is important to consider that the DS and the LTS were easy to understand for the respondents, in contrast to the VAS: as already noted by several authors (e.g., Couper et al., 2006; Ferrando, 2003; Flynn et al., 2004) several subjects showed that a lack of experience with the VAS led to some confusion and an additional explanation was needed in some cases. The DS should only be used if the subjects are not expected to obtain high levels on the given scale, since clear ceiling effects were observed.

Finally, the responses to the DS and the LTS could be regressed on the VAS but tended to vary considerably between individual items and formats. It is therefore important to consider that intervals between different levels as well as centers of the scales seem to be judged differently depending on the response format. The relationship between the three formats can, at least on the item level, merely be assumed as monotonous in an ordinal manner, however not as linear. The choice of the response format should therefore be subject to intensive consideration with regards to context, time available, and respondents' experience with the scale.

NOTES

1. The NEO-PI-R is a questionnaire for personality assessment, measuring personality through five independent factors, each comprising six facets of personality, and was developed by Costa and McCrae (1992).
2. Two examples are: (1) "Sometimes I am not as dependable and reliable as I should" was rephrased into "I am dependable" and (2) "I fulfill my financial obligations completely and as soon as possible" was changed into "I fulfill my obligations."
3. D = dichotomous scale; L = Likert-type scale; V = visual analogue scale.
4. Coefficient alpha for dichotomous variables is also known as the Kuder-Richardson coefficient.
5. RMSEA = root mean square error of approximation; CI = confidence interval; WRMR = weighted root mean square residual; CFI = comparative fit index.

REFERENCES

- Agresti, A. (2002). *Categorical data analysis*. Hoboken, NJ: John Wiley & Sons.
- Berth, H., Goldschmidt, S., Ostendorf, F., & Angleitner, A. (2006). NEO-PI-R. NEO-Persönlichkeitsinventar nach Costa und McCrae. Revidierte Fassung [NEO-PI-R personality inventory, adapted from Costa & McCrae. Revised version]. *Diagnostica*, 52, 95-99.

- Birkett, N. J. (1986). Selecting the number of response categories for a Likert-type scale. In *Proceedings of the American Statistical Association, Section on Survey Research Methods* (pp. 488-492). Washington, DC: American Statistical Association.
- Briggs, M., & Closs, J. S. (1999). A descriptive study of the use of visual analogue scales and verbal rating scales for the assessment of postoperative pain in orthopedic patients. *Journal of Pain and Symptom Management, 18*, 438-446. doi:10.1016/S0885-3924(99)00092-5
- Bühner, M. (2011). *Einführung in die Test-und Fragebogenkonstruktion* [Introduction to test and questionnaire construction]. München, Germany: Pearson-Education.
- Campbell, D. T., & Fiske, D. W. (1959). Convergent and discriminant validation by the multitrait-multimethod matrix. *Psychological Bulletin, 56*, 81-105. doi:10.1037/h0046016
- Cicchetti, D. V., Shoinralter, D., & Tyrer, P. J. (1985). The effect of number of rating scale categories on levels of interrater reliability: A Monte Carlo investigation. *Applied Psychological Measurement, 9*, 31-36. doi:10.1177/014662168500900103
- Clason, D. L., & Dormody, T. J. (1994). Analyzing data measured by individual Likert-type items. *Journal of Agricultural Education, 35*, 31-35.
- Conte, J. M., & Jacobs, R. R. (2003). Validity evidence linking polychronicity and Big Five personality dimensions to absence, lateness, and supervisory performance ratings. *Human Performance, 16*, 107-129. doi:10.1207/S15327043HUP1602_1
- Costa, P. T., Jr., & McCrae, R. R. (1992). Revised NEO Personality Inventory (NEO-PI-R) and NEO Five-Factor Inventory (NEO-FFI) professional manual. Odessa, FL: Psychological Assessment Resources.
- Couper, M. P., Tourangeau, R., Conrad, F. G., & Singer, E. (2006). Evaluating the effectiveness of visual analog scales: A web experiment. *Social Science Computer Review, 24*, 227-245. doi:10.1177/0894439305281503
- Cox III, E. P. (1980). The optimal number of response alternatives for a scale: A review. *Journal of Marketing Research, 17*, 407-422. doi:10.2307/3150495
- Cronbach, L. J. (1951). Coefficient alpha and the internal structure of tests. *Psychometrika, 16*, 297-334. doi:10.1007/BF02310555
- Dawes, J. (2008). Do data characteristics change according to the number of scale points used? An experiment using 5-point, 7-point and 10-point scales. *International Journal of Market Research, 51*, 61-77.
- Drake, K. M., Hargraves, J. L., Lloyd, S., Gallagher, P. M., & Cleary, P. D. (2014). The effect of response scale, administration mode, and format on responses to the CAHPS clinician and group survey. *Health Services Research, 49*, 1387-1399. doi:10.1111/1475-6773.12160
- Dunn, T. J., Baguley, T., & Brunson, V. (2014). From alpha to omega: A practical solution to the pervasive problem of internal consistency estimation. *British Journal of Psychology, 105*, 399-412. doi:10.1111/bjop.12046
- Faulbaum, F., Prüfer, P., & Rexroth, M. (2009). *Was ist eine gute Frage? Die systematische Evaluation der Fragenqualität* [What is a good question? The systematic evaluation of question quality]. Wiesbaden, Germany: Springer.
- Ferrando, P. J. (2003). A Kernel density analysis of continuous typical-response scales. *Educational and Psychological Measurement, 63*, 809-824. doi:10.1177/0013164403251323
- Finn, R. H. (1972). Effects of some variations in rating scale characteristics on the means and reliabilities of ratings. *Educational and Psychological Measurement, 34*, 885-892. doi:10.1177/001316447203200203
- Flynn, D., van Schaik, P., & van Wersch, A. (2004). A comparison of multi-item Likert and Visual Analogue scales for the assessment of transactionally defined coping function. *European Journal of Psychological Assessment, 20*, 44-58. doi:10.1027/1015-5759.20.1.49
- Gignac, G. E. (2014). On the inappropriateness of using items to calculate total scale score reliability via coefficient alpha for multidimensional scales. *European Journal of Psychological Assessment, 30*, 130-193. doi:10.1027/1015-5759/a000181
- Goggin, S., & Stoker, L. (2014). Optimal scale length and single-item attitude measures: Evidence from simulations and a two-wave experiment. *American Political Science Association Annual Meeting Paper*, Washington, DC. Retrieved from <http://www.ssrn.com/link/APSA-2014.html>
- Goldstein, G., & Hersen, M. (2000). *Handbook of psychological assessment*. Oxford, UK: Elsevier.
- Graham, J. M. (2006). Congeneric and (essentially) tau-equivalent estimates of score reliability what they are and how to use them. *Educational and Psychological Measurement, 66*, 930-944. doi:10.1177/0013164406288165
- Grant, S., Aitchison, T., Henderson, E., Christie, J., Zare, S., McMurray, J., & Dargie, H. (1999). A comparison of the reproducibility and the sensitivity to change of visual analogue scales, Borg scales, and Likert scales in normal subjects during submaximal exercise. *Chest Journal, 116*, 1208-1217. doi:10.1378/chest.116.5.1208
- Guyatt, G. H., Townsend, M., Berman, L. B., & Keller, J. L. (1987). A comparison of Likert and visual analogue scales for measuring change in function. *Journal of Chronic Diseases, 40*, 1129-1133. doi:10.1016/0021-9681(87)90080-4

- Hancock, G. R., & Klockars, A. J. (1991). The effect of scale manipulations on validity: Targetting frequency rating scales for anticipated performance levels. *Applied Ergonomics*, 22, 147-154. doi:10.1016/0003-6870(91)90153-9
- Heene, M., Hilbert, S., Draxler, C., Ziegler, M., & Bühner, M. (2011). Masking misfit in confirmatory factor analysis by increasing unique variances: A cautionary note on the usefulness of cutoff values of fit indices. *Psychological methods*, 16, 319-336.
- Helmreich, R., & Stapp, J. (1974). Short forms of the Texas Social Behavior Inventory (TSBI), an objective measure of self-esteem. *Bulletin of the Psychonomic Society*, 4, 473-475.
- Hsiao, Y.-Y., Wu, C.-H., & Yao, G. (2014). Convergent and discriminant validity of the WHOQOL-BREF using a multitrait-multimethod approach. *Social Indicators Research*, 116, 971-988. doi:10.1007/s11205-013-0313-z
- Hu, L., & Bentler, P. M. (1999). Cutoff criteria for fit indexes in covariance structure analysis: Conventional criteria versus new alternatives. *Structural Equation Modeling: A Multidisciplinary Journal*, 6, 1-55. doi:10.1080/10705519909540118
- Jaeschke, R., Singer, J., & Guyatt, G. H. (1990). A comparison of seven-point and visual analogue scales: Data from a randomized trial. *Controlled Clinical Trials*, 11, 43-51. doi:10.1016/0197-2456(89)90005-6
- Joyce, C. R. B., Zutshi, D. W., Hrubes, V., & Mason, R. M. (1975). Comparison of fixed interval and visual analogue scales for rating chronic pain. *European Journal of Clinical Pharmacology*, 8, 415-420. doi:10.1007/BF00562315
- Judge, T. A., Martocchio, J. J., & Thoresen, C. J. (1997). Five-factor model of personality and employee absence. *Journal of Applied Psychology*, 82, 745-755. doi:10.1037/0021-9010.82.5.745
- Lehmann, R., Denissen, J. J., Allemand, M., & Penke, L. (2013). Age and gender differences in motivational manifestations of the Big Five from age 16 to 60. *Developmental Psychology*, 49, 365-383. doi:10.1037/a0028277
- Likert, R. (1932). A technique for the measurement of attitudes. *Archives of Psychology*, 140, 44-60.
- Lounsbury, J. W., Steel, R. P., Loveland, J. M., & Gibson, L. W. (2004). An investigation of personality traits in relation to adolescent school absenteeism. *Journal of Youth and Adolescence*, 33, 457-466. doi:10.1023/B:JOYO.0000037637.20329.97
- Lozano, L. M., García-Cueto, E., & Muñiz, J. (2008). Effect of the number of response categories on the reliability and validity of rating scales. *Methodology: European Journal of Research Methods for the Behavioral and Social Sciences*, 4, 73-79. doi:10.1027/1614-2241.4.2.73
- MacCann, C., Duckworth, A. L., & Roberts, R. D. (2009). Empirical identification of the major facets of conscientiousness. *Learning and Individual Differences*, 19, 451-458. doi:10.1016/j.lindif.2009.03.007
- Matell, M. S., & Jacoby, J. (1971). Is there an optimal number of alternatives for Likert scale items? Study I. Reliability and validity. *Educational and Psychological Measurement*, 31, 657-674. doi:10.1177/001316447103100307
- Maydeu-Olivares, A., Kramp, U., García-Forero, C., Gallardo-Pujol, D., & Coffman, D. (2009). The effect of varying the number of response alternatives in rating scales: Experimental evidence from intra-individual effects. *Behavior Research Methods*, 41, 295-308. doi:10.3758/BRM.41.2.295
- McCrae, R. R., Martin, T. A., & Costa, P. T. (2005). Age trends and age norms for the NEO Personality Inventory-3 in adolescents and adults. *Assessment*, 12, 363-373. doi:10.1177/1073191105279724
- McDonald, R. P. (1970). The theoretical foundations of principal factor analysis, canonical factor analysis, and alpha factor analysis. *British Journal of Mathematical and Statistical Psychology*, 23, 1-21. doi:10.1111/j.2044-8317.1970.tb00432.x
- Muthén, B., Du Toit, S. H., & Spisic, D. (1997). Robust inference using weighted least squares and quadratic estimating equations in latent variable modeling with categorical and continuous outcomes. *Psychometrika*, 75, 1-45.
- Muthén, L. K. (2010). *Model fit index WRMR*. Retrieved from <http://www.statmodel.com/discussion/messages/9/5096.html>
- Muthén, L. K., & Muthén, B. O. (1998). *Mplus 5.2*. Los Angeles, CA: Author.
- Oaster, T. R. F. (1989). Number of alternatives per choice point and stability of Likert-type scales. *Perceptual and Motor Skills*, 68, 549-550. doi:10.2466/pms.1989.68.2.549
- Peter, J. P. (1979). Reliability: A review of psychometric basics and recent marketing practices. *Journal of Marketing Research*, 16, 6-17. doi:10.2307/3150868
- Peter, J. P., & Churchill, G. A., Jr. (1986). Relationships among research design choices and psychometric properties of rating scales: A meta-analysis. *Journal of Marketing Research*, 22, 1-10. doi:10.2307/3151771
- Preston, C. C., & Colman, A. M. (2000). Optimal number of response categories in rating scales: Reliability, validity, discriminating power, and respondent preferences. *Acta Psychologica*, 104, 1-15. doi:10.1016/S0001-6918(99)00050-5
- Revelle, W., & Zinbarg, R. E. (2009). Coefficients alpha, beta, omega, and the glb: Comments on Sijsma. *Psychometrika*, 74, 145-154. doi:10.1007/s11336-008-9102-z
- Rothman, S. (2001). School absence and student background factors: A multilevel analysis. *International Education Journal*, 2, 59-68.

- Samejima, F. (1969). *Estimation of latent ability using a response pattern of graded scores* [Psychometric Monograph No. 17]. Richmond, VA: Psychometric Society.
- Shulman, H. C., & Boster, F. J. (2014). Effect of test-taking venue and response format on political knowledge tests. *Communication Methods and Measures*, 8, 177-189.
- Schwarz, N. (1999). Self-reports: How the questions shape the answers. *American Psychologist*, 54, 93-105. doi:10.1037/0003-066X.54.2.93
- Socan, G. (2000). Assessment of reliability when test items are not essentially tau-equivalent. *Development in Survey Methodology*, 15, 23-35.
- Specht, J., Egloff, B., & Schmukle, S. C. (2011). Stability and change of personality across the life course: The impact of age and major life events on mean-level and rank-order stability of the Big Five. *Journal of Personality and Social Psychology*, 101, 862-882. doi:10.1037/a0024950
- Synonymwörterbuch, D.-D. (2006). *Ein Wörterbuch sinnverwandter Wörter* [Book of synonym words]. Mannheim, Germany: Dudenverlag.
- R Development Core Team (2011). *R: A language and environment for statistical computing*. Vienna, Austria: R Foundation for Statistical Computing. Retrieved from <http://www.R-project.org>
- Weng, L.-J. (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
- Wewers, M. E., & Lowe, N. K. (1990). A critical review of visual analogue scales in the measurement of clinical phenomena. *Research in Nursing & Health*, 13, 227-236. doi:10.1002/nur.4770130405
- Yee, T. W. (2010). The VGAM package for categorical data analysis. *Journal of Statistical Software*, 32, 1-34.
- Yu, C. Y. (2002). *Evaluating cutoff criteria of model fit indices for latent variable models with binary and continuous outcomes* (Doctoral dissertation, University of California Los Angeles). Retrieved from <http://www.statmodel.com/download/Yudissertation.pdf>