Meta-Analytic Guidelines for Evaluating Single-Item Reliabilities of Personality Instruments

Assessment
2014, Vol. 21(3) 272–285
© The Author(s) 2013
Reprints and permissions:
sagepub.com/journalsPermissions.nav
DOI: 10.1177/1073191113498267
asm.sagepub.com

Matthias Spörrle¹ and Magdalena Bekk²

Abstract

Personality is an important predictor of various outcomes in many social science disciplines. However, when personality traits are not the principal focus of research, for example, in global comparative surveys, it is often not possible to assess them extensively. In this article, we first provide an overview of the advantages and challenges of single-item measures of personality, a rationale for their construction, and a summary of alternative ways of assessing their reliability. Second, using seven diverse samples ($N_{\text{total}} = 4,263$) we develop the SIMP-G, the German adaptation of the Single-Item Measures of Personality, an instrument assessing the Big Five with one item per trait, and evaluate its validity and reliability. Third, we integrate previous research and our data into a first meta-analysis of single-item reliabilities of personality measures, and provide researchers with guidelines and recommendations for the evaluation of single-item reliabilities.

Keywords

single-item measures of personality, SIMP, test reliability, test validity, Big Five personality model, meta-analysis, foreign language translation

Even though personality traits have been found to be important determinants of various attitudes and behaviors across all nations and across all social science disciplines, research resources do not always allow for the inclusion of lengthy personality measures. For instance, large global comparative surveys need to use an internationally validated, yet short personality inventory to, for instance, statistically control for personality factors. To partially address this demand, our research provides the scientific community with the first psychometrically sound German-language version of an established English single-item measure of the Big Five personality dimensions (i.e., Single-Item Measures of Personality or the SIMP; Woods & Hampson, 2005).

After addressing the advantages of single-item instruments in general we briefly summarize the current status of the SIMP; here, we also discuss the concerns raised toward single-item measurement of personality. Finally, we present our empirical validation of the German version of the SIMP (i.e., the SIMP-G). Because there are no recommendations on the evaluation of single-item reliabilities and because guidelines used to interpret Cronbach's alpha are not transferable to single-item reliabilities (Osburn, 2000), we used an empirical approach in order to derive reference values for single-item reliabilities: We report a meta-analysis which provides some first guidelines for the evaluation of single-item measures in the context of personality assessment. Single-item measures are sometimes criticized because of

their (supposedly low) reliability and the impossibility to estimate their reliability by means of some standard procedures (e.g., Cronbach's alpha). Based on existing solutions to overcome these challenges our analyses provide the first quantitative review of single-item reliabilities of personality instruments and offer first guidelines for the evaluation of single-item reliabilities.

Advantages of Single-Item Measures

Because single-item measures provide researchers with some advantages they are frequently used in diverse fields of research, for example, to assess job satisfaction (Nagy, 2002), organizational justice (Jordan & Turner, 2008), positivity/negativity of attitudes toward objects (Larsen, Norris, McGraw, Hawkey, & Cacioppo, 2009), or self-esteem (Robins, Hendin, & Trzesniewski, 2001). Their biggest advantage is, of course, their brevity: They take up less space and time (for responding as well as data coding) and

¹University of Applied Management, Erding, Germany

²University of Cologne, Cologne, Germany

Corresponding Author:

Magdalena Bekk, Cologne Graduate School in Management, Economics, and Social Sciences, University of Cologne, Albertus-Magnus-Platz, 50932 Cologne, Germany.

Email: bekk@wiso.uni-koeln.de

leave room for assessing additional constructs (Drolet & Morrison, 2001). A second benefit of single items, which was stated by Wanous, Reichers, and Hudy (1997) is that they do not give the respondent the impression that the questions are repeated. Consequently, respondents are more willing to give their time and answer the questions. Because of these reasons, single-item measures, for example to assess personality traits or life satisfaction, are also frequently used in large global studies such as, for instance, the World Values Survey (Bruni & Stanca, 2006), the European Social Survey (Georgellis, Tsitsianis, & Yin, 2009), or the Socio-Economic Panel (Gerstorf et al., 2008).

Regarding their psychometric properties Wanous et al. (1997) empirically demonstrated that single-item and multiple-item measures of job satisfaction correlated strongly, thus indicating that single-item measures are not generally unacceptable (for similar results when examining different constructs, see Bergkvist & Rossiter, 2007; Finkel, Burnette, & Scissors, 2007). Correspondingly, Burisch (1984a, 1984b, 1997) found out that short measures can be as valid as long measures and that this is especially true for personality measures. He argues that it will become more and more difficult with a growing number of items in a measure to find additional items which are suitable and will increase the validity of the measure, until, at one point, an additional item will actually reduce the validity (Burisch, 1984b). He therefore suggests that many measures might be improved by shortening them.

Single-Item Personality Measures

Looking at single-item personality measures (i.e., one item per factor) two approaches have been used: One approach uses unipolar single items (Five-Item Personality Inventory [FIPI], Gosling, Rentfrow, & Swann, 2003) thereby accepting the risk of acquiescence effects. For example, respondents will tend to affirm an item such as the agreeableness in the FIPI: "I am agreeable, kind (that is, trusting, generous, sympathetic, cooperative, *NOT* aggressive, or cold)." The second approach tries to avoid this risk by using bipolar single items (Denissen, Geenen, Selfhout, & Aken, 2008; Langford, 2003; Rammstedt, Koch, Borg, & Reitz, 2004; Woods & Hampson, 2005). However, even these instruments might suffer from acquiescence when they use desirable and undesirable poles to represent personality factors, for example, for agreeableness *cordial* as one descriptor and *harsh* as opposite descriptor (Rammstedt et al., 2004; see also Denissen et al., 2008). In addition to acquiescence effects both approaches are based on a reduction of longer measures and, thus, capture only one of the personality facets per dimension.

As a solution, Woods and Hampson's (2005) bipolar SIMP places desirable as well as undesirable characterizations at both poles of each personality factor (for an application of this instrument in a Swedish sample, see Bäckman &

Carlstedt, 2010). This approach of placing several, positively as well as negatively connoted items on both ends of each scale aims at addressing two central shortcomings of alternative short personality measures: First, it takes several facets of one personality dimension into account. Second, it reduces the effect of biased responding based on social desirability because both sides of an item include positive as well as negative aspects. Additionally, modifiers, for example, *generally* or *tends to*, are included to further mitigate desirability tendencies. Finally, there is another methodological advantage of this instrument: The SIMP is the only single-item personality measure which is not based on psychometric reductions of longer inventories. Consequently, its items were from the beginning designed to represent the overall personality factor rather than a specific facet.

A criticism that is held up against the SIMP is that one single item might not be able to capture multifaceted characteristics such as the Big Five personality traits. However, this argument might not fully apply to the SIMP: First, as mentioned above, the single-items of the SIMP have from the beginning been constructed to capture the underlying meaning of each overall personality dimension rather than a single specific facet. Due to the item format of presenting several central characteristics of each personality dimension on each pole of the items respondents provide their answers to some extent by calculating a "mental average" for each Big Five trait over all the facets mentioned in each single item. Thus, the multi-item approach of computing an arithmetic mean is replaced by having the participants form a mental average; nevertheless, both approaches assess each Big Five dimension by using an aggregate of several facets.¹

Moreover, even strong proponents of a facet approach consider it appropriate to summarize all facets of one dimension within one expression. A prominent example for this is given in the manual of the NEO Personality Inventory (NEO-PI-R; Ostendorf & Angleitner, 2004): The NEO-PI-R assesses the Big Five dimensions with a total of 240 items assigned to six facets per dimension. Nevertheless, in its short feedback form each of the Big Five dimensions is described in one item with verbal descriptions for the high, intermediate, and low levels of this dimension. This indicates that it is possible to grasp the essential meaning of each Big Five dimension in one item. These five single-item descriptions were also successfully used as single personality items in prior research (Bernard, Walsh, & Mills, 2005).

Research Contribution

We aim at making the SIMP available for the examination of German-speaking samples. German is the second most important language after English in the European Union (TNS Opinion & Social, 2005), with approximately 101 million speakers worldwide (European Commission, 2010). Using this internationally accepted instrument will make it easier

Table 1. Summary of Test Administration and Measures for the Six Samples.

Sample	Sample Characteristics	Personality Measures	Criterion Measures	Timing of Assessment	Assessment Method
Sample 1	$n = 331$; approximately 52% female; $M_{\text{age}} = 28.86$ years; $SD_{\text{age}} = 11.43$, ranging from 17 to 75 years	SIMP-G TIPI	Self-esteem, Portrait Values Questionnaire, PANAS Trait	Cross-sectional	Paper-pencil
Sample 2	$n = 404$ (students), approximately 44% female; $M_{\text{age}} = 23.40$ years; $SD_{\text{age}} = 0.90$, ranging from 22 to 27 years	SIMP-G BFI-10		Cross-sectional	Paper-pencil
Sample 3	$n = 1,559$ ([self-] employees); approximately 55% female; $M_{\text{age}} = 35.38$ years; $SD_{\text{age}} = 10.55$, ranging from 15 to 88 years	SIMP-G NEO-FFI	Life Satisfaction, Emotional Intelligence	Cross-sectional	Online
Sample 4	$n = 372$; approximately 58% female; $M_{\text{age}} = 38.72$ years; $SD_{\text{age}} = 9.06$, ranging from 21 to 66 years	SIMP-G		Cross-sectional	Online
Sample 5	$n = 1,032$; approximately 15% female; $M_{\text{age}} = 43.58$ years; $SD_{\text{age}} = 12.18$, ranging from 18 to 79 years	SIMP-G		Cross-sectional	Online
Sample 6	$n = 122$; approximately 51% female; $M_{\text{age}} = 33.52$ years; $SD_{\text{age}} = 12.04$, ranging from 18 to 69 years	SIMP-G MMs	Altruistic Values	Longitudinal, over 3-month period, assessment once per month	Paper-pencil
Sample 7	$n = 443$; approximately 51% female; $M_{\text{age}} = 35.62$ years; $SD_{\text{age}} = 10.59$, ranging from 16 to 67 years	SIMP-G BFI-10		Cross-sectional	Online

Note. SIMP-G = German adaptation of the Single-Item Measures of Personality; TIPI = Ten-Item Personality Inventory; NEO-FFI = NEO-Five Factor Inventory; BFI-10 = Ten-Item Big Five Inventory; MM = Mini-Markers.

for researchers examining German-speaking samples (e.g., from Austria, Germany, Italy, Luxembourg, or Switzerland) to address an international audience. Furthermore, using measures consistently in various languages (e.g., the SIMP in Great Britain or the United States, the SIMP-G in Germany or Austria) will simplify the collection, aggregation, and comparison of international data.

Method

Participants and Procedure

A total of 4,263 participants (44% women, mean age = 35.99 years, $SD = 11.98$, range 15–88 years) took part in seven studies examining the psychometric properties of the SIMP-G (for details on the seven studies, see Table 1).

Measures

Single-Item Measures of Personality—German Version. The five single items of the SIMP by Woods and Hampson (2005) were independently translated by three translators who were both familiar with the Five-Factor Personality Model and spoke German and English fluently. In line with the translation method used by Muck, Hell, and Gosling (2007) two of these translators next derived a combined translation of the SIMP questionnaire. This translation was then given to another bilingual expert of personality research, who compared the German translation with the original English version to see whether the translation captured the original

meaning of the items (no major changes in the translation were necessary). This procedure resulted in the German version of the SIMP, that is, the SIMP-G (see the appendix for the final version of the SIMP-G; see Table 2 for descriptive statistics and inter-item correlations). To obtain estimates of this instrument's single-item reliability as well as convergent and discriminant validity we used a broad range of alternative Big Five personality measures.

Ten-Item Personality Inventory. We used Muck et al.'s (2007) German version of the Ten-Item Personality Inventory (TIPI). Cronbach's alphas for the five personality dimensions ranged from $\alpha = [.27; .67]$.²

Ten-Item Big Five Inventory. We used the German version of the Ten-Item Big Five Inventory (BFI-10) by Rammstedt and John (2007), using the suggested additional agreeableness item (i.e., 11 items in total), $\alpha = [.19; .73]$.

Mini-Markers. We used Weller and Matiaske's (2008) German version of the 40-item Mini-Markers (MMs), $\alpha = [.81; .90]$.

NEO-Five Factor Inventory. We applied Borkenau and Ostendorf's (2008) German version of the 60 items of the NEO-Five Factor Inventory (NEO-FFI), $\alpha = [.76; .89]$.

To examine the external correlations of the SIMP-G we used several constructs. Unless described otherwise we used forward-backward translations to obtain German versions of the instruments.

Table 2. Descriptive Statistics and Interitem Correlations of the SIMP-G Personality Factors.

	M	SD	SK	KU	(1)	(2)	(3)	(4)	(5)
(1) Extraversion	5.32	2.22	-0.18	-1.02	—	-.28	-.06	.05	-.03
(2) Agreeableness	5.48	2.06	-0.28	-0.77	-.27	—	.11	-.16	.12
(3) Emotional Stability	4.79	2.05	-0.14	-0.84	-.06	.10	—	-.08	.03
(4) Conscientiousness	5.81	2.11	-0.45	-0.68	.05	-.15	-.09	—	-.21
(5) Openness	5.37	2.08	-0.17	-0.80	-.02	.11	.03	-.20	—
Mean	5.35	2.10	-0.24	-0.82					

Note. SK = skewness, KU = kurtosis; $N = 4,263$; SIMP-G = German adaptation of the Single-Item Measures of Personality. Since all skewness and kurtosis values are within or close to the range of -1.00 and $+1.00$ suggested by Muthén and Kaplan (1985) parametric analyses seem justified. However, as there were some minor deviations from normality, we additionally computed Spearman's nonparametric rank correlations which overall were very similar to Pearson's correlation, thus indicating that the parametric indicators of association are not severely influenced by these deviations. For Pearson correlation (above main diagonal): $|r| \geq .03$, $p < .05$ (two-tailed); $|r| \geq .05$, $p < .01$ (two-tailed). For Spearman rank correlation (below main diagonal): $|r| \geq .03$, $p \leq .05$ (two-tailed); $|r| \geq .05$, $p \leq .01$ (two-tailed).

Single-Item Self-Esteem Scale. We used the single-item measure by Robins, Hendin, and Trzesniewski (2001). The authors report a test–retest reliability of $r = .75$.

Portraits Value Questionnaire (PVQ). We administered the German 40-item version of the Portraits Value Questionnaire (PVQ; Schmidt, Bamberg, Davidov, Herrmann, & Schwartz, 2007) to assess 10 basic values, $\alpha = [.62; .85]$.

Satisfaction With Life Scale. We used the five-item Satisfaction With Life Scale of Diener, Emmons, Larsen, and Griffin (1985; German version by Schumacher, Klaiberg, & Brähler, 2003), $\alpha = .90$.

Emotional Intelligence (EQ) Scale. Wong and Law's (2002) measure was used to cover four facets of emotional intelligence (German version: Spörrle, Welp, Ringenberg, & Försterling, 2008). The Dimensions were assessed reliably, $\alpha = [.85; .94]$.

Positive and Negative Affect Schedule. We assessed positive ($\alpha = .75$) and negative ($\alpha = .69$) affect as a trait using the 10-item Positive and Negative Affect Schedule (PANAS) short form (Thompson, 2007).

Altruistic Values Scale. We used seven items on altruistic values by Stern, Dietz, Abel, Guagnano, and Kalof (1999), $\alpha = .85$.

Results

Stability-Based (Test–Retest) Single-Item Reliability Indicators

We examined the SIMP-G over three assessment times (with time intervals of one month, $n = 122$) to calculate the stability-based reliabilities. Overall, this resulted in a total of 5 (personality dimensions) $\times$ 3 (time intervals, i.e., t_1t_2 , t_1t_3 ,

t_2t_3) = 15 reliability estimates (r_{tt}). These estimates ranged from $r_{tt} = .63$ (openness t_1t_3) to $r_{tt} = .82$ (extraversion t_2t_3) and were overall satisfactory (mean $r_{tt} = .73$).³ The mean retest reliabilities for the five dimensions were .79 (extraversion), .68 (agreeableness), .71 (emotional stability), .74 (conscientiousness), and .72 (openness). These test–retest reliabilities obtained for the SIMP-G are very similar to the retest reliabilities obtained for the English version of the SIMP; for this the authors reported an overall mean retest reliability of .71 (Woods & Hampson, 2005). The only noticeable difference might be found for the dimension conscientiousness, with the SIMP-G showing a slightly better mean retest reliability (.74) than the English SIMP (.60).

This classical test–retest approach, as Alwin (2007) pointed out, cannot account for changes in the true score over time. To address this shortcoming he recommended two stability estimation approaches, one by Heise (1969) and another one by Wiley and Wiley (1970) which are both based on linear structural quasi-Markov simplex models (and both require three points of assessment at least), but use different constraints in their model assumptions. These test–retest reliability procedures resulted in an overall mean reliability of the SIMP-G of $r_{tt} = .81$ (for the Heise as well as the Wiley and Wiley estimate). The test–retest reliabilities estimated for the five dimensions, both for the approach by Heise (r_{ttH} ; Heise estimates based on the assumption of equal reliabilities, thus resulting in only one reliability value), as well as by Wiley and Wiley (r_{ttW} ; Wiley and Wiley estimates based on the assumption of equal error variances; the average value of the three obtained estimates is presented) were $r_{ttH} = .81$, $r_{ttW} = .82$ (extraversion), $r_{ttH} = .74$, $r_{ttW} = .76$ (agreeableness), $r_{ttH} = .73$, $r_{ttW} = .73$ (emotional stability), $r_{ttH} = .76$, $r_{ttW} = .77$ (conscientiousness), and $r_{ttH} = .92$, $r_{ttW} = .91$ (openness). Comparing these findings with the classical test–retest reliabilities (see above) indicates that accounting for variation in the true score increased the reliability estimate.

Consistency-Based Single-Item Reliability Indicators

Wanous and Hudy (2001) suggested two different approaches to derive a consistency-based reliability estimate of single-item measures: The first is based on the correction for attenuation formula, using the reliability of a reference measure of the same trait and the empirical as well as assumed underlying construct correlation between these measures. Their second approach also requires a reference measure and uses communalities from factor analysis, by using all the items from the single-item and multi-item measure simultaneously in a factor analysis, as reliability estimates. Denissen et al. (2008) suggests using the same factor-analytic approach but recommended to use the squared main factor loading of the single item, as this is not inflated by secondary factor loadings. We examined the consistency-based single-item reliability of the SIMP-G using all these estimation techniques, that is, the correction for attenuation formula (r_{SI}), the communalities (h^2), and the squared main factor loadings (a^2) in five samples using the TIPI ($n = 331$), BFI-10 ($n = 847$; two samples), MMs ($n = 122$), and NEO-FFI ($n = 1,559$) as reference measures for the SIMP-G.

Mean reliability estimates across all SIMP-G factors and samples were $r_{SI} = .50$, $h^2 = .57$, and $a^2 = .57$. These values are quite similar to those of the original SIMP version ($h^2 = .50$, Woods & Hampson, 2005). Mean single-item reliabilities across all methods and samples for the five factors were .70 (extraversion), .46 (agreeableness), .54 (emotional stability), .55 (conscientiousness), and .49 (openness).

Meta-Analysis of Single-Item Reliabilities of Personality Instruments

How should these reliability indicators be interpreted? Because no one has yet provided recommendations regarding single-item reliabilities and because guidelines used to interpret Cronbach's alpha are not transferable to single-item reliabilities (Osburn, 2000), we use a meta-analytic approach to provide some first guidelines for the evaluation of single-item measures in the context of personality measures. As data basis we used the findings of Bernard et al. (2005), Denissen et al. (2008), Gosling et al. (2003), Rammstedt et al. (2004), Woods and Hampson (2005) as well as the single-item reliability estimates of the SIMP-G from our samples. Overall, we were able to obtain 240 different reliability indicators of single-item instruments based on the Big Five model (120 from the sources cited above and 120 from our own data sets).

We report weighted reliability means, using the alpha reliability meta-analysis approach based on the varying coefficient model by Bonett (2010), which, unlike alternative models, is not based on unrealistic assumptions

concerning the reliability and the sampling. This approach takes different sample sizes into account and performs well in conditions with heterogeneous samples and effect sizes. We examined both the stability-based as well as the consistency-based single-item reliability by using all estimation approaches mentioned above.

A total of 33 indicators over all dimensions were stability-based and 207 indicators were consistency-based.⁴ Of the 33 stability-based reliability estimates, 11 were based on the retest reliability (r_{tt}), 11 were based on the approach by Heise (r_{ttH}), and 11 were based on the approach by Wiley and Wiley (r_{ttW}), respectively. Of the 207 consistency-based estimates, 81 were based on the correction for attenuation formula (r_{SI}), 60 were based on the communalities (h^2), and 66 on the squared factor loadings (a^2). For each of the Big Five personality dimensions we calculated the mid-50% percentile range of the single-item reliabilities and the 25%, 50%, and 75% percentile borders (values are given in the 1st to 4th numerical row of Table 3 for the stability-based and in the 9th to 12th numerical row for the consistency-based reliabilities). We provide these descriptive values for the five dimensions separately (second to sixth column in Table 3) as well as the mean single-item reliability over all five personality dimensions in the last column of Table 3.

Looking at the last column of Table 3, the stability-based reliability of single-personality items was on average at .76 (mid 50% between .72 and .82). The consistency-based reliability of single-personality items was on average at .54 (mid 50% between .42 and .66). Looking at the percentiles we can see that for the stability-based reliabilities only the 25% smallest values were smaller than .71. This indicates that the majority (75%) of all retest-reliabilities of single-items were larger than the established cutoff value (.70; see Peterson, 1994) for reliabilities of multi-item measures. Thus, we can draw the conclusion that these cutoff values are appropriate to evaluate the retest reliabilities of single items as well.

However, for the consistency-based reliability estimates the 25% smallest values are smaller than .42, half of all values are smaller than .54 and 75% of all values are smaller than .67. Thus, over all personality dimensions and all three estimation approaches, 75% of all values have a consistency-based single-item reliability which is smaller than the recommended cutoff value for a multi-item measure as suggested by Peterson (1994). This demonstrates the need to establish first guidelines for evaluating consistency-based single-item reliability. The values provided in Table 3 might serve as a first orientation not only to interpret the single-item reliability of the SIMP-G but also the reliability of other single-item measures of personality (e.g., the SIMP in other languages).

From the results of our meta-analytic examination we can extract the following reference points for the Big Five personality factors: Over all five personality dimensions,

Table 3. Meta-Analysis of Single-Item Reliabilities.

	Extraversion	Agreeableness	Emotional Stability	Conscientiousness	Openness	Mean ^b
Stability-based single item reliability						
Descriptive values ^a						
Mid 50% range	.79-.86	.68-.76	.71-.85	.74-.81	.72-.91	.72-.82
Percentiles						
25%	.77	.65	.71	.70	.68	.71
50%	.81	.75	.73	.76	.79	.77
75%	.86	.77	.85	.81	.91	.84
Means						
Mean ^a	.81	.71	.76	.74	.78	.76
Mean r_{tt}	.74	.62	.70	.65	.62	.66
Mean r_{SM-H}	.83	.75	.79	.78	.87	.81
Mean r_{SM-W}	.84	.77	.79	.79	.87	.82
Consistency-based single-item reliability						
Descriptive values ^a						
Mid 50% range	.64-.77	.35-.52	.45-.63	.51-.70	.35-.55	.42-.66
Percentiles						
25%	.63	.35	.42	.51	.35	.42
50%	.71	.47	.52	.58	.44	.54
75%	.77	.52	.64	.71	.57	.67
Means						
Mean ^a	.70	.45	.53	.59	.45	.54
Mean r_{SI}	.76	.44	.53	.58	.41	.54
Mean h^2	.67	.48	.55	.60	.49	.56
Mean a^2	.66	.44	.53	.58	.44	.53

Note. $K = 240$ (number of reliability estimates; $k = 33$ stability-based, $k = 207$ consistency-based). Meta-analysis data from Bernard, Walsh, and Mills (2005; $k = 5$, single-item NEO PI-R with NEO PI-R); Denissen, Geenen, Selfhout, and Aken (2008; $k = 18$ stability-based over 6 waves with two different openness items; $k = 12$ consistency-based, TIPI-r with BFI); Gosling, Rentfrow, and Swann (2003; $k = 15$, FIPI with BFI-44); Rammstedt, Koch, Borg, and Reitz (2004; $k = 20$, single-item measure with BFI-K, NEO-FFI); Woods and Hampson (2005; $k = 50$, SIMP with BFI, MMs, TDA-35, TDA-100); and SIMP-G single-item reliabilities ($k = 15$ stability-based over 3 waves; $k = 105$ consistency-based, with TIPI, BFI-10, NEO-FFI, and MMs from all three assessment periods) from this study. SIMP-G = German adaptation of the Single-Item Measures of Personality; TIPI = Ten-Item Personality Inventory; NEO-FFI = NEO -Five Factor Inventory; BFI-10 = Ten-Item Big Five Inventory; MM = Mini-Markers; TDA = trait descriptive adjectives.

a. Mean values over all three single-item reliability estimation approaches. b. Mean over all five personality dimensions. r_{tt} = test-retest correlation; r_{SM-H} = quasi-Markov simplex model by Heise (1969; for an overview, see Alwin, 2007); r_{SM-W} = quasi-Markov simplex model by Wiley and Wiley (1970; for an overview, see Alwin, 2007); r_{SI} = single-item reliability based on correction for attenuation formula (under the assumption that the underlying construct correlation $r = .90$, see Wanous & Hudy, 2001); h^2 = communalities (Wanous & Hudy, 2001); a^2 = squared main factor loadings (Denissen et al., 2008). Means are weighted by using the SAS alpha reliability meta-analysis program by Bonett (2010). There were no or, in few cases, only minor (on the second decimal place) differences between the weighted and unweighted means.

the average stability-based single-item reliability lies at .76. For the five dimensions the average stability-based single-item reliability lies at .81 (extraversion), .71 (agreeableness), .76 (emotional stability), .74 (conscientiousness), and .78 (openness). For the consistency-based single-item reliability, the average single-item reliability over all five personality dimensions lies at .54. For the five dimensions the average single-item reliability lies at .70 (extraversion), .45 (agreeableness), .53 (emotional stability), .59 (conscientiousness), and .45 (openness). When comparing the reliabilities of the five personality dimensions across the different estimation procedures extraversion assessment was consistently found to be comparatively reliable whereas agreeableness tended to be the least

reliable one. These overall reference values might not only be useful in evaluating the single-item reliability of personality measures but might also serve as first reference points for the reliability of single-item measures of other constructs, for example, subdimensions of emotional intelligence or specific values.

As explained above the estimation of the consistency-based single-item reliabilities depends on multi-item reference measures. Therefore, the question arises whether the single-item reliability estimate depends on the number of items of the multi-item instrument. Examining the data from our meta-analysis, it is noticeable that the consistency-based single-item reliability estimates (on all three estimation methods) decrease with an increasing number of items

of the reference measure: Mean consistency-based reliability estimates across all SIMP-G factors with a reference measure of no more than 10 items were $r_{SI} = .60$, $h^2 = .70$, and $a^2 = .67$; for multiple-item measures with 11 to 50 items they were $r_{SI} = .48$, $h^2 = .56$, and $a^2 = .57$; and for multiple-item measures with more than 50 items they were $r_{SI} = .36$, $h^2 = .37$, and $a^2 = .38$. Therefore, when evaluating consistency-based reliability estimates of single-item instruments the length of the reference instrument has to be taken into account and should, therefore, be reported.

Convergent and Discriminant Validity

With a mean correlation of $r_{Mean} = .53$, $r = [.29; .76]$, the SIMP-G showed good convergent validity (i.e., correlation with the mean value of the same dimension of another personality instrument) based on 35 (five dimensions and seven samples) correlations with the TIPI ($n = 331$), BFI-10 ($n = 847$; two samples), MMs ($n = 122$; assessed three times), and NEO-FFI ($n = 1,559$). The majority of the correlations (i.e., 18 of 35) were $r \geq .50$ (indicating a large effect) and only one correlation was $r \leq .30$ (which indicates a less than medium-sized effect, Cohen, 1988). For the five factors the mean correlations were $r = .69$, $[.58; .76]$, for extraversion; $r = .44$, $[.34; .56]$, for agreeableness; $r = .51$, $[.43; .58]$, for emotional stability; $r = .51$, $[.29; .64]$, for conscientiousness; and $r = .46$, $[.40; .54]$, for openness. Thus, the comparatively high reliability of extraversion and the comparatively low reliability of agreeableness are reflected in their convergent validity estimates. Furthermore, our results are in line with the results of McCrae, Kurtz, Yamagata, and Terracciano (2011). Just like their findings our results also indicate on average a higher reliability of the stability-based indicators (compared with the consistency-based indicators) with a lower variability (i.e., smaller range). In line with this higher reliability, McCrae et al. showed that the stability-based indicators were more important predictors of validity criteria than the consistency-based indicators.

To test the discriminant validity between the factors we examined the absolute off-diagonal correlations: Within the SIMP-G as well as between the SIMP-G and alternative personality measures we computed the mean absolute correlation of each personality dimension with the remaining four factors. The 10 absolute off-diagonal correlations between the personality factors within the SIMP-G (mean $|r| = .11$, $[.03; .27]$) as well as the 140 absolute off-diagonal correlations between the SIMP-G and the alternative factor-based multi-item Big Five measurements (mean $|r| = .05$, $[.00; .27]$) were overall very low indicating that the content of the item of one personality dimension did not overlap with the content of the items of other dimensions.

Overall, across the seven samples and factors we obtained 35 convergent and 150 discriminant correlations

(i.e., 4 discriminant correlations for each convergent correlation plus 10 within off-diagonal correlations). In all 35 cases (i.e., for each factor of the Big Five in each of the seven samples) the convergent correlation was higher than even the highest absolute discriminant correlation of the respective factor. For the five factors the mean absolute discriminant correlations were $|r| = .08$, $[.00; .27]$, for extraversion; $|r| = .08$, $[.01; .26]$, for agreeableness; $|r| = .06$, $[.00; .17]$, for emotional stability; $|r| = .04$, $[.01; .15]$, for conscientiousness; and $|r| = .04$, $[.00; .17]$, for openness. The majority of discriminant correlations (i.e., 102 of 140) were $r \leq .10$, and no correlation was $r \geq .30$. The overall mean convergent correlation ($r = .53$) was substantially higher than the overall mean absolute discriminant correlation ($|r| = .05$).

External Correlates

Most of the external correlations of the SIMP-G (i.e., 70 out of 95) were in the direction which was expected based on previous research (see Table 4; correlations which were in the expected direction are indicated in bold numbers), thus indicating high consensus with existing evidence. Moreover, 80 out of 95 correlations were consistent in terms of their direction with those obtained on the basis of the multi-item personality measures in our samples (Table 4 also includes the correlation of the multi-item personality measures with the external criteria for comparative purposes—these correlations are given in brackets). Overall, these findings indicate substantial convergence of the SIMP-G with existing evidence as well as with our data when using the reference instruments.

We compared the correlations between the SIMP-G and the external criteria with the correlations between multi-item measures of personality and the same external criteria. The correlations did not differ significantly in their strength in over half of the cases (i.e., in 68 out of 95 cases). Thus, for most criterion correlations we did not find a significant difference between the application of the SIMP-G and the other Big Five measures using (substantially) more items.

When looking descriptively at the absolute numerical correlation values the multi-item measures showed higher correlations with the criteria than the SIMP-G in the majority of cases (i.e., 68 out of 95 cases). The SIMP-G showed higher correlations in 26 out of 95 cases and exactly the same value as the multi-item measure in 1 case out of 95. This analysis, however, does not take negligible differences in the value of the correlations into account. Therefore, we also used Cohen's (1988) effect size norms for correlation coefficients to build three groups: In the first group, which was the largest (i.e., 46 out of 95 cases), the coefficients of both, the SIMP-G and the multi-item measure with the external criterion, lay in the same effect size interval, either

Table 4. Overview of External Correlations of the SIMP-G and Alternative Personality Instruments With Criteria.

Criterion	Extraversion		Agreeableness		Emotional Stability		Conscientiousness		Openness	
	SIMP-G (multi-item measure)	<i>p</i> _{diff}	SIMP-G (multi-item measure)	<i>p</i> _{diff}	SIMP-G (multi-item measure)	<i>p</i> _{diff}	SIMP-G (multi-item measure)	<i>p</i> _{diff}	SIMP-G (multi-item measure)	<i>p</i> _{diff}
Self-esteem ^a , <i>N</i> = 331 (TIPI)	.19*** (.27***)	ns	-.05 (.16***)	ns	.27*** (.60***)	*	.05 (.19***)	ns	.09 (.33***)	*
PVQ-power ^b , <i>N</i> = 331 (TIPI)	.33*** (.26***)	ns	-.42*** (-.21***)	*	.04 (.19***)	ns	.08 (.11)	ns	-.17*** (.01)	ns
PVQ-Achievement ^b , <i>N</i> = 331 (TIPI)	.26*** (.21***)	ns	-.26*** (-.11*)	ns	-.14* (.02)	ns	.22*** (.24***)	ns	-.08 (.04)	ns
PVQ-Hedonism ^b , <i>N</i> = 331 (TIPI)	.24*** (.26***)	ns	-.15*** (.02)	ns	.00 (.20***)	*	-.11* (-.11*)	ns	.12* (.28***)	ns
PVQ-Stimulation ^b , <i>N</i> = 331 (TIPI)	.20*** (.27***)	ns	-.15*** (-.05)	ns	.05 (.20***)	ns	-.28*** (-.20***)	ns	.19*** (.47***)	*
PVQ-Self-Direction ^b , <i>N</i> = 331 (TIPI)	.24*** (.32***)	ns	-.13* (.00)	ns	.09 (.26***)	ns	-.14** (-.03)	ns	.36*** (.61***)	*
PVQ-Universalism ^b , <i>N</i> = 331 (TIPI)	.00 (.01)	ns	.20*** (.18***)	ns	.01 (-.05)	ns	-.06 (-.05)	ns	.19*** (.25***)	ns
PVQ-Benevolence ^b , <i>N</i> = 331 (TIPI)	.08 (.13*)	ns	.21*** (.28***)	ns	-.09 (-.01)	ns	.04 (.07)	ns	.10 (.18***)	ns
PVQ-Tradition ^b , <i>N</i> = 331 (TIPI)	-.22*** (-.26***)	ns	.19*** (.15**)	ns	-.08 (-.18***)	ns	.14*** (.11*)	ns	-.14*** (-.21***)	ns
PVQ-Conformity ^b , <i>N</i> = 331 (TIPI)	-.12* (-.18***)	ns	.18*** (.17**)	ns	-.13* (-.17**)	ns	.23*** (.21***)	ns	-.22*** (-.30***)	ns
PVQ-Security ^b , <i>N</i> = 331 (TIPI)	-.09 (-.12*)	ns	.03 (.13*)	ns	-.12* (-.03)	ns	.37*** (.42***)	ns	-.20*** (-.21***)	ns
Life Satisfaction ^a , <i>N</i> = 1,559 (NEO-FFI)	.22*** (.45***)	*	.02 (.26***)	*	.21*** (.53***)	*	.06*** (.29***)	*	-.02 (.05)	ns
EQ-SEA ^c , <i>N</i> = 1,559 (NEO-FFI)	.13*** (.24***)	*	.00 (.22***)	*	.08*** (.29***)	*	.07*** (.34***)	*	.09*** (.21***)	*
EQ-OEA ^c , <i>N</i> = 1,559 (NEO-FFI)	.15*** (.27***)	*	.07*** (.22***)	*	-.03 (.06*)	ns	.02 (.20***)	*	.14*** (.22***)	ns
EQ-UOE ^c , <i>N</i> = 1,559 (NEO-FFI)	.21*** (.40***)	*	-.02 (.11***)	ns	.12*** (.34***)	*	.09*** (.46***)	*	.13*** (.18***)	ns
EQ-ROE ^c , <i>N</i> = 1,559 (NEO-FFI)	.05* (.24***)	*	.09*** (.19***)	*	.45*** (.49***)	ns	.01 (.26***)	*	.04 (.03)	ns
Positive Affect Trait ^a , <i>N</i> = 404 (BFI-10)	.32*** (.36***)	ns	-.05 (.03)	ns	-.01 (.21***)	*	.16*** (.38***)	*	.03 (.16***)	ns
Negative Affect Trait ^a , <i>N</i> = 404 (BFI-10)	-.20*** (-.25***)	ns	-.03 (-.15**)	ns	-.23*** (-.48***)	*	.02 (-.07)	ns	.08 (.04)	ns
Altruistic Values ^d , <i>N</i> = 122 (MMs)	-.12 (-.11)	ns	.25*** (.19*)	ns	-.18* (-.01)	ns	.04 (.09)	ns	.03 (.12)	ns

Note. Criterion correlations between SIMP-G factors and external criteria. *p*_{diff} = significance of pairwise difference between the validity correlation of the SIMP-G and the validity correlation of the multi-item measure with the external criteria. The alpha level of the *p* values was adjusted for multiple testing using Bonferroni's adjustment procedure. Boldfaced values indicate that the direction of the correlation is in line with previous research as indicated by the lowercase letters; italic values indicate that there is no direction of the correlation given in the prior research. In parentheses: Criterion correlations between factors of alternative personality measures (as indicated in parentheses in the first column) and external criteria. PVQ = Portrait Value Questionnaire; EQ-SEA = Emotional Intelligence-Self-Emotion Appraisal; EQ-OEA = Emotional Intelligence-Other-Emotion Appraisal; EQ-UOE = Emotional Intelligence-Use of Emotions; EQ-ROE = Emotional Intelligence-Regulation of Emotion. The direction of the relations was predicted based on ^aDeNeve and Cooper (1998); ^bRoccas, Sagiv, Schwartz, and Knafo (2002); ^cLopes, Salovey, and Straus (2003); ^dNeuman and Kickul (1998); and ^eWoods and Hampson (2005).

p* ≤ .05. *p* ≤ .01. ****p* ≤ .001.

Table 5. Predictive Validity: Regression Analyses.

Criterion	Block 1	Block 2	Criterion	Block 1	Block 2
	R^2_{adj}	ΔR^2_{adj}		R^2_{adj}	ΔR^2_{adj}
Self-esteem, $n = 331$ (TIPI)	.13***	.27***	PVQ–Security, $n = 331$ (TIPI)	.16***	.07***
PVQ–power, $n = 331$ (TIPI)	.23***	.01*	Life Satisfaction, $n = 1,559$ (NEO-FFI)	.10***	.25***
PVQ–Achievement, $n = 331$ (TIPI)	.14***	.02*	EQ–SEA, $n = 1,559$ (NEO-FFI)	.04***	.16***
PVQ–Hedonism, $n = 331$ (TIPI)	.07***	.06***	EQ–OEA, $n = 1,559$ (NEO-FFI)	.05***	.11***
PVQ–Stimulation, $n = 331$ (TIPI)	.14***	.15***	EQ–UOE, $n = 1,559$ (NEO-FFI)	.09***	.27***
PVQ–Self-Direction, $n = 331$ (TIPI)	.20***	.21***	EQ–ROE, $n = 1,559$ (NEO-FFI)	.21***	.11***
PVQ–Universalism, $n = 331$ (TIPI)	.06***	.04**	Positive Affect Trait, $n = 404$ (BFI-10)	.12***	.15***
PVQ–Benevolence, $n = 331$ (TIPI)	.07***	.05***	Negative Affect Trait, $n = 404$ (BFI-10)	.10***	.16***
PVQ–Tradition, $n = 331$ (TIPI)	.09***	.02*	Altruistic Values, $n = 122$ (MMs)	.07*	.00
PVQ–Conformity, $n = 331$ (TIPI)	.14***	.05***			
			Mean	.12	.11

Note. Hierarchical regression: Block 1 includes the five SIMP-G items, Block 2 additionally adds another personality measure as indicated in the brackets behind the criterion.

* $p \leq .05$. ** $p \leq .01$. *** $p \leq .001$.

between 0 and 0.1 or between 0.1 and 0.3 or between 0.3 and 0.5 or above 0.5. In the second and third group, the coefficients of the SIMP-G and the multi-item measure did not fall into the same effect size interval: In the second group, the coefficient for the multi-item measure was larger than for the SIMP-G (i.e., in 40 out of 95 cases). In the third group, the coefficient for the SIMP-G was larger than for the multi-item measure (i.e., in 9 out of 95 cases). Summarizing, in the majority of cases the correlation between the external criterion and the SIMP-G did not significantly differ from the correlation between the same external criterion and the multi-item measure of personality. When comparing the correlations on basis of established effect size intervals most correlations between the external criterion and the SIMP-G were either equal or even larger (i.e., 55 out of 95 cases) than the correlation between the external criterion and the multi-item measure. Thus, the SIMP-G overall showed a pattern of external correlations in line with existing findings and similar to those of the longer measures used in our surveys.

Predictive Validity

We followed a two-step regression approach: In the first step, we used the SIMP-G as predictors (i.e., five predictors), in the second step we added the additional Big Five measure to determine the incremental variance explained by the additional instrument (i.e., a total of 10 predictors).

As can be seen in Table 5, the SIMP-G predicted a significant amount of variance of all outcome variables. Based on Cohen (1988) almost half of these proportions (i.e., 8 out of 19) are to be regarded as medium effects (i.e., $.13 \leq R^2 \leq .26$). When examining the additional variance explained by the multi-item personality measures in the second step,

results indicated that in more than half of the cases (i.e., 11 out of 19) the multi-item measure explained less than 13% of variance (which is equal to a small-to-medium effect size) in addition to the variance already explained by the SIMP-G; in 10 out of 19 cases the multi-item measure explained less additional variance than was explained by the SIMP-G in the first step.

Across all criterion and instrument constellations both instruments (i.e., the SIMP-G and a second Big Five measure) together explained on average a total of 23% of the variance of the criterion. Of the total variance explained by the two personality instruments the SIMP-G explained, in the first step, on average 50% of the totally explained variance ($R^2_{\text{adj}} = .12$, [.04; .23]) and the second instrument explained on average additional 50% of the criterion variance ($\Delta R^2_{\text{adj}} = .11$, [.00; .27]). Thus, adding at least twice as many items (the shortest additional Big Five instruments has 10 items) resulted on average in a duplication of the explained variance.

The proportion of variance explained clearly varied as a function of the length of the second Big Five instrument: Instruments with 10 or 11 items explained on average an additional 10% of the total variance explained by personality. Thus, doubling the number of items resulted in less than doubling the variance explained. On the contrary, instruments with 40 or more items explained only an additional 15% of variance to the variance explained by the SIMP-G.

On an aggregate level the SIMP-G predicted 50% of the total explained variance and the additional multi-item instrument predicted another 50%. Thus, the longer instruments we used provided approximately the same incremental variance as our single-item measure was able to provide in the first step. The SIMP-G, therefore, seems to provide a

good reference baseline for a general analysis of the predictive value of the Big Five personality dimensions.

Biological Sex and Age Effects

There were significant differences between female ($n = 1,874$) and male ($n = 2,389$) participants for the factors agreeableness, openness, and emotional stability in line with previous research (for similar results, see Chapman, Duberstein, Sörensen, & Lyness, 2007; Costa, Terracciano, & McCrae, 2001): Female participants reported higher mean values on agreeableness ($M_A = 5.71$, $SD_A = 2.03$) and openness ($M_O = 5.49$, $SD_O = 2.01$) and lower mean values on emotional stability ($M_{ES} = 4.31$, $SD_{ES} = 1.96$) compared with men ($M_A = 5.28$, $SD_A = 2.06$; $M_O = 5.27$, $SD_O = 2.12$; $M_{ES} = 5.17$, $SD_{ES} = 2.03$, $ps \leq .001$; Cohen's d effect size: $d_A = 0.21$, $d_O = 0.11$, $d_{ES} = 0.43$).

For women we detected a positive relation between emotional stability and age, $r = .09$, $p \leq .001$, whereas for men there was a negative association between openness and age, $r = -.09$, $p \leq .001$, and a positive correlation between conscientiousness and age, $r = .07$, $p \leq .001$ (for similar results, see Goldberg, Sweeney, Merenda, & Hughes, 1998; Lucas & Donnellan, 2009). Overall, these findings are totally consistent with existing evidence, thus further corroborating the valid applicability of the SIMP-G.

Discussion

It was our aim to develop and validate the German version of the SIMP by Woods and Hampson (2005) and to provide first guidelines for the interpretation of the various single-item reliability approaches for the field of personality research. Given its brevity the SIMP-G proved to be a relatively reliable and valid measure of personality: The SIMP-G's single-item reliabilities were similar to those of its English original version, and it showed good stability over time. The convergent validity, too, was good for the SIMP-G with medium to strong convergent correlations with a broad range of alternative personality measures and small discriminant correlations.

The external correlations of the SIMP-G with various outcome variables were mainly in line with existing evidence—even though they tended to be smaller in magnitude—thus confirming its validity in terms of external criteria. Future research should extend our preliminary evidence by examining associations between the SIMP-G factors and additional criteria, especially in fields where short instruments are particularly needed. This is the case, for example, in the field of work and occupational or personnel psychology or management for which job performance (Barrick & Mount, 1991), job satisfaction (Judge, Heller, & Mount, 2002), self-efficacy (Strobel, Tumasjan, & Spörrle, 2011), or leadership styles (Judge, Bono, Ilies, & Gerhardt, 2002) might be relevant criteria to examine.

In line with previous findings demonstrating the usefulness of short measures (Möttus, Pullmann, & Allik, 2006; Yarkoni, 2010), our research provides a reliable and valid German version of the SIMP. The SIMP-G provides a good baseline for a general analysis of the predictive value of the Big Five personality dimensions. Given these encouraging findings it seems promising to validate additional versions of the SIMP in other common languages (e.g., Spanish or Chinese) and to transfer this approach of single-item measurement into the domain of facets of personality (see Mullins-Sweatt, Jamerson, Samuel, Olson, & Widiger, 2006, for a first approach on this). Moreover, even though self-reports are the usual basis of personality research other-person evaluations of personality were also found to have predictive value across various fields of psychological research (Bekk & Spörrle, 2010). Thus, future research might examine the psychometric properties of the SIMP-G from an other-person perspective. Irrespective of the perspective, the SIMP-G is not designed to assess a differentiated personality profile or to examine specific personality facets; only multi-item measures are appropriate for these purposes.

The SIMP-G provides researchers from various research areas with an economic first indicator of personality, for example, to statistically control for personality differences. This can be done by either using the SIMP-G items as covariates in multiple regression analyses (Spörrle, Strobel, & Tumasjan, 2010), or by building a residualized index to examine if personality accounts for the effects of interest (Mehl, Vazire, Holleran, & Clark, 2010). Additionally, the SIMP-G items can be used as covariates to control for personality factors in more complex models, like (multiple) mediation models (for using covariates within mediation models, see Preacher & Hayes, 2008). Furthermore personality factors—even though not affecting the dependent variable—might be correlated with the predictor variable, thus acting as a suppressor variable. In such cases, the SIMP-G might be used in semipartial correlations by partialing personality out (Ones, Viswesvaran, & Reiss, 1996).

This research also provides the first meta-analytic guidelines for the interpretation of single-item reliabilities in the context of personality assessment. Such guidelines are important as single items will most likely be less reliable than multiple-item measures, and can, thus, not be compared directly with multiple-item measures with regard to their reliability. Moreover, we found the consistency-based reliability of single items to depend on the reliability of the multiple-item reference measure used to estimate the single-item reliability (i.e., longer reference instruments result in lower reliability estimates). Therefore, meta-analytic guidelines comprising single-item reliability estimates resulting from reference measures of different length provide researchers with a tool to better interpret the single-item reliabilities of their

personality measures. Several conclusions can be drawn from our meta-analytic review: First, it might be possible to use established evaluation rules of retest-reliabilities to evaluate the stability-based single-item reliability, as we could show that these estimates have on average a single-item reliability of .76. Moreover, concerning retest reliabilities, the two different approaches (i.e., retest correlation and estimation models based on linear structural quasi-Markov simplex procedures) seem to result in somewhat different stability estimates: Controlling for the true score variation (which seems reasonable because reliability is an accuracy indicator of the instrument but not of the trait's fluctuations) increased stability estimates.

Second, with an average value of .54 most of the consistency-based single-item reliabilities were below the accepted reliability levels of multiple-item measures (even though, at least for some extraversion single items, surprisingly good reliability estimates higher than .70 were obtained). Thus, in line with previous research on single-item reliabilities of job satisfaction (Wanous et al., 1997), we provide evidence that it might be necessary to evaluate consistency-based single-item reliabilities on the basis of guidelines different from those of multiple-item measures.

Third, our analysis is the first to include all three indicators of consistency-based single-item reliability suggested to date, that is, the correction for attenuation formula (r_{SI}), the communalities (h^2), and the squared main factor loadings (a^2). Our findings suggest that on aggregated levels these indicators resulted in similar reliability estimates with the communality (h^2) tending to be an upper bound of single-item reliability estimates. Nevertheless, their sometimes pronounced differences (e.g., for extraversion) point to the necessity of reporting all of them when addressing single-item reliability.

Fourth, across different single-item measures of personality, consistency-based single-item reliabilities were systematically lower than test-retest reliability coefficients, thus indicating that representatives of both reliability concepts should be considered when evaluating an instrument. Finally, the meta-analytic findings indicate that some personality dimensions (e.g., extraversion, stability-based $r_{Mean} = .81$, consistency-based $r_{Mean} = .70$) are consistently assessed more reliably by means of single-item instruments than others (e.g., agreeableness, stability-based $r_{Mean} = .71$, consistency-based $r_{Mean} = .45$). This result points to the possibility that the latter dimensions might either be perceived more heterogeneously or that these dimensions are more normatively evaluative, that is, people stick to social norms and do not dare answer to be disagreeable, which would lead to less reliable answers than for less evaluative dimensions. Thus, it seems to be useful to not only look at overall guidelines but also at guidelines for each personality dimension separately. Future research should examine whether these findings can be consistently

replicated and whether they are based on differences in the item characteristics or in the respondents' ability to evaluate the respective personality trait.

Concluding, we can state that the SIMP-G is a relatively reliable and valid measure of personality. The SIMP-G provides researchers from various research areas with an efficient first indicator of personality. Translating and validating the original English measure into German is a first step to provide international researchers all over the globe with a psychometrically sound and short personality instrument. Using this instrument will allow future research to compare the results gained from it more easily even across cultures. Furthermore this research provides some first meta-analytic guidelines on single-item reliability for personality measures. This provides researchers with the means to evaluate the strength of their instrument's single-item reliability, thus helping them to conduct research with very short scales whenever this is needed.

Appendix

German Adaptation of the Single-Item Measures of Personality (SIMP-G)

Auf dieser Seite finden Sie fünf Paare von Aussagen, die allgemein benutzt werden, um verschiedene Menschen zu beschreiben.

Bitte umkreisen Sie jeweils eine Markierungslinie auf der Skala, je nachdem wie sehr jede der beiden Beschreibungen eines Paares Ihrer Meinung nach auf Sie generell zutrifft.

Dies kann z. B. folgendermaßen aussehen:

- Beschreibt ein Paar Sie gleichermaßen gut, dann markieren Sie bitte den Punkt in der Mitte der Skala.

Beschreibung 1 |---|⊕---| Beschreibung 2

- Sollte Beschreibung 1 etwas besser auf Sie persönlich zutreffen, als Beschreibung 2, dann markieren Sie bitte einen Punkt, der etwas näher an Beschreibung 1 liegt.

Beschreibung 1 |---|⊙---| Beschreibung 2

- Sollte Beschreibung 2 genau auf Sie zutreffen und Beschreibung 1 überhaupt nicht, dann markieren Sie bitte den Punkt direkt neben Beschreibung 2.

Beschreibung 1 |---|---|---|⊕ Beschreibung 2

Wie sehr treffen die folgenden Aussagen auf Sie zu? Bitte denken Sie dabei nicht an spezifische Situationen, sondern ganz allgemein, wie sehr die Aussagen Sie selbst in den meisten Bereichen und Situationen in Ihrem Leben beschreiben.

Allgemein wirke ich tendenziell eher wie eine Person, die . . .

Gesprächig und aufgeschlossen ist, und sich in Gruppen wohlfühlt, die aber auch laut sein kann und Aufmerksamkeit bekommen möchte.		Reserviert und zurückhaltend ist, die sich eher im privaten Kreis wohlfühlt, die nicht gerne Aufmerksamkeit auf sich lenkt und sich gegenüber Fremden manchmal schüchtern verhält.
Anderen geradeheraus sagt, was sie denkt, dazu neigt, andere zu kritisieren, häufig Fehler bei anderen findet und nicht sehr tolerant gegenüber Menschen ist, die sich dumm verhalten.		Generell anderen Menschen vertraut, ihnen Fehler nachsieht und an ihnen interessiert ist, auf die man sich verlassen kann und der es schwer fällt, nein zu sagen.
Empfindsam und leicht aufzuregen ist, aber auch angespannt sein kann.		Entspannt ist und wenig emotional, die selten irritiert, durcheinander oder traurig ist.
Gerne Dinge plant, Sachen ordentlich hält und auf Details achtet, die aber auch unnachgiebig und unflexibel sein kann.		Nicht unbedingt nach Plan arbeitet, zwar flexibel ist, aber manchmal auch unorganisiert und häufig vergisst, Sachen an ihren ursprünglichen Platz zurückzustellen.
Praxisorientiert ist, kein Interesse an abstrakten Ideen hat, bekannte und vertraute Arbeiten bevorzugt und wenig künstlerische Interessen besitzt.		Gerne Zeit damit verbringt, über Dinge nachzudenken, eine ausgeprägte Phantasie und Vorstellungskraft hat, der es Spaß macht über neue Wege wie man Dinge tun könnte nachzudenken, aber der ein gewisser Pragmatismus fehlen kann.

Acknowledgments

The authors would like to thank Aaron L. Pincus and three anonymous reviewers as well as Prisca Brosi, Markus Bühner, Georg Felser, Rudolf Kerschreiter, Joachim Kruse, Jon K. Maner, Jason D. Shaw, Maria Strobel, Stephen A. Woods, and Matthias Ziegler for their insightful comments on earlier versions of this article.

Declaration of Conflicting Interests

The authors declared no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.

Funding

The authors received no financial support for the research and/or authorship. The fee for the open access publication of this article was paid by the authors' institutions.

Notes

1. An interesting alternative to this facet-aggregating approach might be the search for a specific facet so strongly correlated with the overall dimension that it would be justified to use this single facet as an indicator of the overall dimension. To our knowledge, such facets have not been proposed so far.
2. Throughout the article squared brackets contain the range of the respective statistics (e.g., Cronbach's alpha or correlation coefficient) with the first value representing the smallest and the second value representing the largest value obtained.
3. Throughout this article all mean correlations are based on Fisher's r to z transformations.
4. The number is not a multiple of five (for the five dimensions) as Denissen et al. (2008) changed, during their scale development process, the single-item for the dimension openness, thus providing us with values for two different single openness items. Thus we have three additional openness values for the stability-based and two additional values for the consistency-based (correction for attenuation approach and factor loadings) methods.

References

- Alwin, D. F. (2007). *Margins of error: A study of reliability in survey measurement*. Hoboken, NJ: Wiley.
- Bäckman, C., & Carlstedt, B. (2010). A construct validation of a profession-focused personality questionnaire (PQ) versus the FFPI and the SIMP. *European Journal of Psychological Assessment*, 26, 136-142.
- Barrick, M. R., & Mount, M. K. (1991). The Big Five personality dimensions and job performance. A meta-analysis. *Personnel Psychology*, 44, 1-26.
- Bekk, M., & Spörrle, M. (2010). The influence of perceived personality characteristics on positive attitude towards and suitability of a celebrity as a marketing campaign endorser. *Open Psychology Journal*, 3, 54-66.
- Bergkvist, L., & Rossiter, J. R. (2007). The predictive validity of multiple-item versus single-item measures of the same constructs. *Journal of Marketing Research*, 44, 175-184.
- Bernard, L. C., Walsh, R. P., & Mills, M. (2005). Ask once, may tell: Comparative validity of single and multiple item measurement of the Big-Five personality factors. *Counseling and Clinical Psychology Journal*, 2, 40-57.
- Bonett, D. G. (2010). Varying coefficient meta-analytic methods for alpha reliability. *Psychological Methods*, 15, 368-385.
- Borkenau, P., & Ostendorf, F. (2008). *NEO-FFI: NEO-Fünf-Faktoren-Inventar nach Costa und McCrae (2., neu normierte und vollständig überarbeitete Auflage)* [NEO-FFI: NEO-five-factor-inventory by Costa and McCrae; 2nd ed. including new norms]. Göttingen, Germany: Hogrefe Verlag.
- Bruni, L., & Stanca, L. (2006). Income aspiration, television and happiness: Evidence from the World Values Survey. *Kyklos*, 59, 209-225.
- Burisch, M. (1984a). Approaches to personality inventory construction. A comparison of merits. *American Psychologist*, 39, 214-227.
- Burisch, M. (1984b). You don't always get what you pay for: Measuring depression with short and simple versus long and

- sophisticated scales. *Journal of Research in Personality*, 18, 81-98.
- Burisch, M. (1997). Test length and validity revisited. *European Journal of Personality*, 11, 303-315.
- Chapman, B. P., Duberstein, P. R., Sörensen, S., & Lyness, J. M. (2007). Gender differences in Five Factor Model personality traits in an elderly cohort. *Personality and Individual Differences*, 43, 1594-1603.
- Cohen, J. (1988). *Statistical power analysis for the behavioral sciences* (2nd ed.). Hillsdale, NJ: Lawrence Erlbaum.
- Costa, P. T., Terracciano, A., & McCrae, R. R. (2001). Gender differences in personality traits across cultures: Robust and surprising findings. *Journal of Personality and Social Psychology*, 81, 322-331.
- DeNeve, K. M., & Cooper, H. (1998). The happy personality: A meta-analysis of 137 personality traits and subjective well-being. *Psychological Bulletin*, 124, 197-229.
- Denissen, J. J. A., Geenen, R., Selfhout, M., & Aken, M. A. G. (2008). Single-item Big Five ratings in a social network design. *European Journal of Personality*, 22, 37-54.
- Diener, E., Emmons, R. A., Larsen, R. J., & Griffin, S. (1985). The satisfaction with life scale. *Journal of Personality Assessment*, 49, 71-75.
- Drolet, A. L., & Morrison, D. G. (2001). Do we really need multiple-item measures in service research? *Journal of Service Research*, 3, 196-204.
- European Commission. (2010). *Euromosaik-Studie: Die europäischen Sprachen* [Euromosaik-Study: The European languages]. Retrieved from http://ec.europa.eu/education/languages/euromosaik/doc4658_de.htm#3
- Finkel, E. J., Burnette, J. L., & Scissors, L. E. (2007). Vengefully ever after: Destiny beliefs, state attachment anxiety, and forgiveness. *Journal of Personality and Social Psychology*, 92, 871-886.
- Georgellis, Y., Tsitsianis, N., & Yin, Y. P. (2009). Personal values as mitigating factors in the link between income and life satisfaction: Evidence from the European Social Survey. *Social Indicators Research*, 91, 329-344.
- Gerstorf, D., Ram, N., Estabrook, R., Schupp, J., Wagner, G. G., & Lindenberger, U. (2008). Life satisfaction shows terminal decline in old age: Longitudinal evidence from the German Socio-Economic Panel Study (SOEP). *Developmental Psychology*, 44, 1148-1159.
- Goldberg, L. R., Sweeney, D., Merenda, P. F., & Hughes, J. E. (1998). Demographic variables and personality: The effects of gender, age, education, and ethnic/racial status on self-descriptions of personality attributes. *Personality and Individual Differences*, 24, 393-403.
- Gosling, S. D., Rentfrow, P. J., & Swann, W. B. (2003). A very brief measure of the Big-Five personality domains. *Journal of Research in Personality*, 37, 504-528.
- Heise, D. R. (1969). Separating reliability and stability in test-retest correlation. *American Sociological Review*, 34, 93-101.
- Jordan, J. S., & Turner, B. A. (2008). The feasibility of single-item measures for organizational justice. *Measurement in Physical Education and Exercise Science*, 12, 237-257.
- Judge, T. A., Bono, J. E., Ilies, R., & Gerhardt, M. W. (2002). Personality and leadership: A qualitative and quantitative review. *Journal of Applied Psychology*, 87, 765-780.
- Judge, T. A., Heller, D., & Mount, M. K. (2002). Five-Factor model of personality and job satisfaction: A meta-analysis. *Journal of Applied Psychology*, 87, 530-541.
- Langford, P. H. (2003). A one-minute measure of the Big Five? Evaluating and abridging Shafer's (1999a) Big Five markers. *Personality and Individual Differences*, 35, 1127-1140.
- Larsen, J. T., Norris, C. J., McGraw, A. P., Hawkey, L. C., & Cacioppo, J. T. (2009). The evaluative space grid: A single-item measure of positivity and negativity. *Cognition and Emotion*, 23, 453-480.
- Lopes, P. N., Salovey, P., & Straus, R. (2003). Emotional intelligence, personality, and the perceived quality of social relationships. *Personality and Individual Differences*, 35, 641-658.
- Lucas, R. E., & Donnellan, M. B. (2009). Age differences in personality: Evidence from a nationally representative Australian sample. *Developmental Psychology*, 45, 1353-1363.
- McCrae, R. R., Kurtz, J. E., Yamagata, S., & Terracciano, A. (2011). Internal consistency, retest reliability, and their implications for personality scale validity. *Personality and Social Psychology Review*, 15, 28-50.
- Mehl, M. R., Vazire, S., Holleran, S. E., & Clark, C. S. (2010). Eavesdropping on happiness: Well-being is related to having less small talk and more substantive conversations. *Psychological Science*, 21, 539-541.
- Möttus, R., Pullmann, H., & Allik, J. (2006). Toward more readable Big Five personality inventories. *European Journal of Psychological Assessment*, 22, 149-157.
- Muck, P. M., Hell, B., & Gosling, S. D. (2007). Construct validation of a short Five-Factor model instrument. A self-peer study on the German adaptation of the Ten-Item Personality Inventory (TIPI-G). *European Journal of Psychological Assessment*, 23, 166-175.
- Mullins-Sweatt, S. N., Jamerson, J. E., Samuel, D. B., Olson, D. R., & Widiger, T. A. (2006). Psychometric properties of an abbreviated instrument of the five-factor model. *Assessment*, 13, 119-137.
- Muthén, B., & Kaplan, D. (1985). A comparison of some methodologies for the factor analysis of non-normal Likert variables. *British Journal of Mathematical and Statistical Psychology*, 38, 171-189.
- Nagy, M. S. (2002). Using a single-item approach to measure facet job satisfaction. *Journal of Occupational and Organizational Psychology*, 75, 77-86.
- Neuman, G. A., & Kickul, J. R. (1998). Organizational citizenship behaviors: Achievement orientation and personality. *Journal of Business and Psychology*, 13, 263-279.
- Ones, D. S., Viswesvaran, C., & Reiss, A. D. (1996). Role of social desirability in personality testing for personnel selection: The red herring. *Journal of Applied Psychology*, 81, 660-679.
- Osburn, H. G. (2000). Coefficient alpha and related internal consistency reliability coefficients. *Psychological Methods*, 5, 343-335.
- Ostendorf, F., & Angleitner, A. (2004). *NEO-PI-R: NEO-Persönlichkeitsinventar nach Costa und McCrae* [NEO-PI-R: NEO-Personality Inventory by Costa and McCrae]. Göttingen, Germany: Hogrefe.
- Peterson, R. A. (1994). A meta-analysis of Cronbach's coefficient alpha. *Journal of Consumer Research*, 21, 381-391.

- Preacher, K. J., & Hayes, A. F. (2008). Asymptotic and resampling strategies for assessing and comparing indirect effects in multiple mediator models. *Behavior Research Methods*, 40, 879-891.
- 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.
- Rammstedt, B., Koch, K., Borg, I., & Reitz, T. (2004). Entwicklung und Validierung einer Kurzskala für die Messung der Big-Five-Persönlichkeitsdimensionen in Umfragen [Development and validation of a short measure of the Big Five personality dimensions in surveys]. *ZUMA Nachrichten, Gesis*, 28, 5-28.
- Robins, R. W., Hendin, H. M., & Trzesniewski, K. H. (2001). Measuring global self-esteem: Construct validation of a single-item measure and the Rosenberg self-esteem scale. *Personality and Social Psychology Bulletin*, 27, 151-161.
- Roccas, S., Sagiv, L., Schwartz, S. H., & Knafo, A. (2002). The Big Five personality factors and personal values. *Personality and Social Psychology Bulletin*, 28, 789-801.
- Schmidt, P., Bamberg, S., Davidov, E., Herrmann, J., & Schwartz, S. H. (2007). Die Messung von Werten mit dem "Portraits Value Questionnaire" [Measuring values using the "Portraits Value Questionnaire"]. *Zeitschrift für Sozialpsychologie*, 38, 261-275.
- Schumacher, J., Klaiberg, A., & Brähler, E. (Eds.). (2003). *Diagnostische Verfahren zu Lebensqualität und Wohlbefinden* [Diagnostic measures for life quality and well-being]. Göttingen, Germany: Hogrefe.
- Spörrle, M., Strobel, M., & Tumasjan, A. (2010). On the incremental validity of irrational beliefs in predicting subjective well-being while controlling for personality factors. *Psicothema*, 22, 543-548.
- Spörrle, M., Welpe, I. M., Ringenberg, I., & Försterling, F. (2008). *Irrationale Kognitionen als Korrelate emotionaler Kompetenzen aus dem Kontext emotionaler Intelligenz und individueller Zufriedenheit am Arbeitsplatz* [Irrational cognitions as correlates of emotional competences in the context of emotional intelligence and individual satisfaction at work]. *Zeitschrift für Personalpsychologie*, 7, 113-128.
- Stern, P. C., Dietz, T., Abel, T., Guagnano, G. A., & Kalof, L. (1999). A value-belief-norm theory of support for social movements: The case of environmentalism. *Human Ecology Review*, 6, 81-97.
- Strobel, M., Tumasjan, A., & Spörrle, M. (2011). Be yourself, believe in yourself, and be happy: Self-efficacy as a mediator between personality factors and subjective well-being. *Scandinavian Journal of Psychology*, 52, 43-48.
- Thompson, E. R. (2007). Development and validation of an internationally reliable short-form of the Positive and Negative Affect Schedule (PANAS). *Journal of Cross-Cultural Psychology*, 38, 227-242.
- TNS Opinion & Social. (2005, Spring). *Eurobarometer 63.4: Europeans and Languages* (National report, executive summary). Brussels, Belgium: European Commission.
- Wanous, J. P., & Hudy, M. J. (2001). Single-item reliability: A replication and extension. *Organizational Research Methods*, 4, 361-375.
- Wanous, J. P., Reichers, A. E., & Hudy, M. J. (1997). Overall job satisfaction: How good are single-item measures? *Journal of Applied Psychology*, 82, 247-252.
- Weller, I., & Matiaske, W. (2008). *Gütekriterien einer deutschsprachigen Version der Mini-Markers zur Erfassung der "Big Five"* [Quality indicators of a German version of "Big Five" Mini-Markers]. Berlin, Germany: Berichte der Werkstatt für Organisations- und Personalforschung e.V. [Reports of the workshop of organizational and personnel research e.V.]
- Wiley, D. E., & Wiley, J. A. (1970). The estimation of measurement error in panel data. *American Sociological Review*, 35, 112-117.
- Wong, C.-S., & Law, K. S. (2002). The effects of leader and follower emotional intelligence on performance and attitude: An exploratory study. *Leadership Quarterly*, 13, 243-274.
- Woods, S. A., & Hampson, S. E. (2005). Measuring the Big Five with single items using a bipolar response scale. *European Journal of Personality*, 19, 373-390.
- Yarkoni, T. (2010). The abbreviation of personality, or how to measure 200 personality scales with 200 items. *Journal of Research in Personality*, 44, 180-198.