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)

Peter M. Muck[1], Benedikt Hell[2], and Samuel D. Gosling[3]

[1]University of Bielefeld, Germany, [2]University of Hohenheim, Germany , [3]University of Texas, Austin, TX, USA

Abstract. The five-factor model (FFM) is currently the predominant model in trait psychology. To meet the need for an extremely brief measure of the FFM, Gosling, Rentfrow, and Swann (2003) developed the Ten-Item Personality Inventory (TIPI), which can be administered in about a minute. Here we describe the development and construct validation of a German version of the TIPI (the TIPI-G). Using a multijudge (self and peer), multiinstrument (TIPI-G and the German version of the NEO-PI-R) design, we evaluated the TIPI-G in terms of internal consistency, factor structure, convergent and discriminant validity, and coverage of the NEO-PI-R facets. Together the analyses suggest that the 10 unipolar items of the TIPI-G can provide an efficient approximation for longer measures of the FFM personality constructs. As such, the TIPI-G is recommended for research where time is limited, where the primary theoretical focus is on other constructs, or where it is desirable to reduce the testing burden on participants.

Keywords: five-factor model, Ten-Item Personality Inventory, personality assessment, applied context

The five-factor model (FFM), or Big Five, is currently the predominant model of personality in trait psychology. The gold standard for measuring the dimensions of the FFM (Emotional Stability = ES [vs. Neuroticism = N]; Extraversion = E; Openness to experience = O; Agreeableness = A; Conscientiousness = C) is the 240-item NEO-PI-R (Costa & McCrae, 1992). However, this instrument is too lengthy for many research and applied uses. Therefore, several shorter instruments were developed, including the 60-item NEO-FFI (Costa & McCrae, 1992) and the 44-item Big Five Inventory (BFI; John & Srivastava, 1999). However, studies that require participants to rate themselves and multiple others on several occasions may profit from the use of even shorter scales, as would large-scale surveys, prescreening packets, longitudinal studies, and experience-sampling studies. Such instruments were developed and include the Ten-Item Personality Inventory (TIPI [in Germany, the acronym TIPI is already in use for another instrument: Trier Integrated Personality Inventory; Becker, 2003]; Gosling et al., 2003), Langford's (2003) five-item instruments and the Single-Item Measures of Personality (SIMP; Woods & Hampson, 2005). To extend the benefits of these very brief instruments to German contexts, we describe the translation of one very brief instrument (the TIPI) to German and a construct validation of the translated instrument.

Utility of Short Instruments

When developing a test, a balance must be struck between psychometric and practical considerations. The balance between these two concerns will lie at different points for different research questions. However, the tradeoff is not always as severe as it is sometimes portrayed. Empirical studies that directly compare short and long tests have shown that the psychometric advantages of long tests are not always as great as one would predict purely on theoretical grounds (Burisch, 1984). More important, a loss in reliability does not always lead to a loss in predictive validity (e.g., Gardner, Cummings, Dunham, & Pierce, 1998).

As a consequence of the benefits of brevity, researchers have sought to realize the benefits of short measures in assessing basic personality dimensions. In the context of the Big Five, Saucier (1994) condensed Goldberg's (1992) 100-item Big Five marker set to 40 so-called Mini-Markers. Shafer (1999) identified pairs of adjectives used by at least three previous investigators (e.g., Goldberg, 1992) and reduced them to a set of 30 bipolar rating scales. However, even these shorter scales are too long for some purposes, so several attempts have been made to develop even briefer measures of the FFM.

Langford (2003) developed several abridged scales of Shafer's (1999) 30-item FFM instrument. Langford

European Journal of Psychological Assessment 2007; Vol. 23(3):166–175
DOI 10.1027/1015-5759.23.3.166

showed that these short scales were reliable (i.e., temporally stable), converged with the original scales, and did not lose much of their power to predict dependent variables like job satisfaction, job stress, and a diverse array of leadership constructs. Rammstedt, Koch, Borg, and Reitz (2004) describe a rather similar approach. However, they based their scales on previously published adjectives; a self-report version of their instrument yielded satisfactory patterns of discriminant and convergent correlations with two other FFM instruments, and was as successful as the longer instruments in predicting such variables as job satisfaction, absenteeism, and self-reported job performance. Woods and Hampson (2005) developed a five-item instrument (the SIMP) by providing characterizations of each of the 10 poles of the FFM dimensions; the SIMP measures were reliable (i.e., temporally stable), showed reasonable self/other agreement in four of the five domains, converged with other five-factor measures, and had a similar pattern of external correlations (e.g., with good work habits, perceived stress, intrinsic work motivation, etc.) compared to longer measures.

Gosling et al. (2003) developed two short unipolar instruments, the Five-Item Personality Inventory (FIPI) and the Ten-Item Personality Inventory (TIPI). The FIPI items had a hierarchical structure with two main descriptors and six additional clarifying descriptors for each item. The TIPI consisted of only two descriptors for each item. It had twice as many items as the five-item FIPI, but the TIPI has only two descriptors per item so has half as many descriptors overall and is less cognitively complex than the FIPI. The TIPI uses a 7-point Likert-type scale ranging from 1 (*disagree strongly*) to 7 (*agree strongly*). Gosling et al. conducted a series of studies using self-, peer, and observer reports for the FIPI. However, only self-ratings have been published so far for the TIPT. The convergent validity for the TIPI with the 44-item BFI (John & Srivastava, 1999) was better than that of the FIPI and comparable to the convergent validity of other longer multi-item FFM measures. Discriminant correlations were substantially lower than convergent correlations and the TIPI demonstrated good stability as indexed by test-retest correlations. The patterns of the TIPI's external correlates with variables like self-esteem, physical attractiveness, or political values matched those of the BFI, although the magnitude of these correlations was slightly stronger for the BFI. The analyses suggested that the TIPI is psychometrically superior to the FIPI and was recommended for research where a very brief personality scale is needed. Because of its psychometric quality and current efforts to establish the instrument across cultures we chose to translate the TIPI.

Translation

The ten items of the English language TIPI were first translated independently into German by the first two authors of the present article. The translators' mother tongue is German and they both have a fluent understanding of the English language. Next the first two authors consensually derived a combined version of the questionnaire, the TIPI-German (TIPI-G). Additionally, the translation tables by Hofstee, Kiers, De Raad, Goldberg, and Ostendorf (1997) and Ostendorf (1990) were consulted. In 12 of 20 cases, the tables offered the same translation. Slightly different translations were found for eight descriptors. Although such tables can provide a useful guide to translations they should be used with great caution because the verbatim translations can result in subtle but consequential discrepancies between the English and German instruments. For example, the translation tables translate *sympathetic* as *einfühlsam*, which has a meaning closer to empathy; we illustrate this point with four examples of cases where we chose not to adopt the translations proposed by the tables.

1. *Complex* had been translated as *komplex* or *kompliziert* in the translation tables. We opted for *vielschichtig* (synonymous to *multilayer*) instead because complex is an indicator of openness. *Komplex* as a verbatim translation is too abstract a term whereas *kompliziert* has a potentially negative meaning like complicated.
2. *Warm* had been translated as *warm*. However, the German word *warm* is rarely used to describe persons (and when it is, it sometimes has a meaning like *gay* in the sense of homosexual). Therefore, we used warmherzig (i.e., *warm-hearted*).
3. *Careless* had been translated as *unsorgfältig* but this use is uncommon in German. It results from a frequently applied negation form (the prefix *un*) because a translation of *careful* – the opposite of careless – is *sorgfältig*. Our translation *achtlos* can be found in several dictionaries as a translation of *careless* and is a more sophisticated expression.
4. *Emotionally stable* had been translated as *gefühlsstabil* but again this use is uncommon in German. We chose *emotional stabil*. This exactly represents the original meaning and the relevant construct. As emotion and feeling (= Gefühl) are used synonymously in German everyday language the connotation is not altered.

The four other terms that did not exactly match the translation tables were enthusiastic, easily upset, sympathetic, and disorganized. We did not use a bilingual sample because one-to-one trait adjective correlations are generally rather low (e.g., John, Goldberg, & Angleitner, 1984). Indeed, recent research has suggested that bilinguals may actually change their personalities slightly as they switch between languages, undermining the use of bilinguals as the gold standard for evaluating instruments (Ramírez-Esparza, Gosling, Benet-Martínez, Potter, & Pennebaker, 2006). Thus, we favor the practice of long-term parallel construct validation procedures, comparing the nomological networks of the two instruments in the two languages. However, as a further test of the translation and to make sure the language was not too technical, a German-English bilingual speaker with no knowledge of the FFM compared

the original and the German translation. No changes had to be made. The resulting questionnaire (see Appendix) was used for the empirical study. Previous validation research on the TIPI had used only self-reports. To improve upon this design, we collected validation data using a self/peer design.

Method

Recruitment and Participants

To help obtain a heterogeneous sample, a snowball recruitment procedure was used. Twenty students of economics collected data from persons they knew as well as from persons these acquaintances knew. Each student tried to find 10 volunteers who would then ask two relatives, friends, or colleagues to complete the peer evaluation form. In total, 180 self- and 359 peer reports were returned representing a rate of return of approximately 90 percent. The students received course credit for collecting the data. No other compensation was given.

Of the participants providing self-reports, 93 (51.7%) were male and 87 (48.3%) were female, 78 (43.3%) were students, 92 (51.1%) were employed, and 10 did not specify their status. Of those who reported age (97.8%), the mean was 31.84 years (SD = 13.56; M = 25; range = 17–75).

Of the peers who indicated their gender, 151 (42.1%) were male and 200 (55.7%) were female. Of those who reported their age (98.1%), the mean age was 32.31 years (SD = 13.67; M = 25; range = 13–74). Of the 347 respondents who indicated how they knew the self-rater, 262 (75.5%) knew the self-raters from private contexts only, 32 (9.2%) knew the self-raters from work contexts only, and 53 (15.3%) knew the self-raters from both contexts. The median length of acquaintance was 6.25 years (range = 0.25–57 years). The average quality of the relationship was deemed to be very good (M = 4.49 on a scale from 1 = *neutral* to 6 = *exceptionally good*; SD = 1.19).

Instruments

Two instruments were used. The first was the German translation of the TIPI (i.e., the TIPI-G), which is the focus of the present study. The second instrument was the German adaptation of the NEO-Personality Inventory (NEO-PI-R; Ostendorf & Angleitner, 2004; original by Costa & McCrae, 1992); this instrument was chosen as the validity criterion because it is the most comprehensive measure of the FFM in the German language. In addition to measuring the five broad dimensions, the NEO-PI-R also assesses the 30 facets (six for each dimension) developed by Costa and McCrae (1992). The inclusion of these facets allowed us to position the instrument in relation to both the five dimensions and the 30 facets.

Some additional instruments were also administered: a job-related inventory for the assessment of the Big Five (AT-B; cf. Höft, 2002), and a new scale format for the measurement of the Big Five (Muck, Hell, & Höft, in press) tested on the NEO-PI-R (NEO BARS) and the AT-B (AT-B BARS). These instruments are of no direct relevance here but are mentioned because different participants took different combinations of the questionnaires. All participants completed the TIPI-G but only half of them completed the NEO-PI-R (and the AT-B BARS; the other half completed the AT-B and the NEO BARS). Every peer rater completed the TIPI-G but only one of the peers completed the NEO-PI-R (and the AT-B BARS; the other peer rater completed the AT-B and the NEO BARS). Thus, an aggregate value of the TIPI-G peer evaluations could be used for most analyses.

Results

The results are ordered in the following way: First, internal TIPI-G analyses are reported. Second, analyses based on both the TIPI-G and the NEO-PI-R are presented. In each case, the self-perspective is analyzed first, the peer perspective next, and then the self/peer comparisons, if available.

Internal Analyses of the German TIPI

The descriptive statistics of the self-ratings are presented in Table 1. We report the TIPI-G self-ratings alongside the self-ratings from the original English TIPI; the German sample was overwhelmingly white making the white subgroup in the original TIPI normative data the most appropriate comparison group. Except for Agreeableness, the TIPI-G self-ratings are slightly higher than in the original TIPI and have slightly lower standard deviations. However, the rank order of means is the same across languages and the grouping of the standard deviations is similar (ES and E > O, A, and C). The internal consistency coefficients (Cronbach's αs) are not very different from the original version of the TIPI. It should be borne in mind that the original TIPI was designed to optimize content validity; with only two items to tap each broad dimension, the goal of maximizing content validity inevitably comes at the expense of internal consistency (John & Benet-Martínez, 2000).

The means of the peer reports are very similar to the self-reports. The only significant difference between self- and peer reports is for Conscientiousness (t_{paired} = 2.08; p < .05; cf. Table 1) but even this difference is small (d = .23). The averaged internal consistency coefficients of the two peer raters are slightly higher for three of the scales (ES, E, C) compared to the self-reports. The αs of the aggregat-

Table 1. Descriptive statistics and Cronbach's αs of the TIPI-G scales

	Self-reports for the TIPI-G and TIPI						Peer reports for the TIPI-G			TIPI-G self-peer differences		
Big 5	Mean		*SD*		α		Mean	*SD*	α	t_p	*p*	*d*
dimen.	TIPI-G	TIPI	TIPI-G	TIPI	TIPI-G	TIPI						
ES	5.10	4.85	1.20	1.45	.67	.73	5.10	1.05	.73/.80	.09	.93	.01
E	4.87	4.56	1.21	1.48	.57	.68	4.83	.99	.60/.65	.43	.67	.04
O	5.49	5.43	.97	1.06	.54	.45	5.43	.83	.53/.55	1.01	.32	.12
A	5.20	5.26	.95	1.12	.42	.40	5.15	.86	.42/.50	.43	.67	.05
C	5.85	5.47	.93	1.13	.66	.50	5.69	1.03	.76/.81	2.08	.04	.23

Notes. ES = Emotional Stability; E = Extraversion; O = Openness; A = Agreeableness; C = Conscientiousness. *N* = 175 for TIPI-G self, *N* = 181 for TIPI-G peer. Values for the TIPI represent the normative data for the white ethnicity in the Mean and SD columns. Two values for the Cronbach's αs for the peer raters are reported: The first is calculated as the average of the two peer raters, the second is the α of the aggregated values. t_p = *t* value for paired samples; *p* = significance level of t_p; *d* = effect size. *df* for *t*-test = 174.

Table 2. Correlations of the TIPI-G scales

		Self-ratings					Peer ratings				
		ES	E	O	A	C	ES	E	O	A	C
Self											
	ES	1.00					.38***	.15*	.02	.00	.19*
	E	.33*** (.23)	1.00				.19*	.56***	.32***	.01	–.02
	O	.20** (.21)	.42*** (.36)	1.00			.02	.21**	.40***	.06	–.05
	A	.10 (.31)	–.03 (.08)	.16* (.19)	1.00		–.12	–.02	.07	.32***	–.04
	C	.39*** (.21)	.09 (.10)	.21** (.12)	.25** (.17)	1.00	.12	–.07	.02	.12	.48***
Peer											
	ES						1.00				
	E						.29***	1.00			
	O						.25**	.44***	1.00		
	A						.15*	–.07	.28***	1.00	
	C						.34***	.01	.16*	.24**	1.00

Notes. ES = Emotional Stability; E = Extraversion; O = Openness; A = Agreeableness; C = Conscientiousness. *N* (self-self) = 175; *N* (self-peer) = 175; *N* (peer-peer) = 181. * = $p < .05$; ** = $p < .01$; *** = $p < .001$. Numbers in brackets denote correlations of the original TIPI based on *N* = 1817 self-reports.

ed values are higher than the averaged αs of the single peer raters, reflecting the gain in reliability attributable to aggregating the scores of two peers.

Table 2 reports the intercorrelations among the TIPI-G scales. The pattern of correlations is comparable to the original TIPI. The mean discriminant correlation is .22 for the TIPI-G and .20 for the original TIPI. Especially high correlations can be found between E and O as well as between ES and C. This is true for both self- and peer reports. Reassuringly, the convergent correlations were consistently and substantially stronger than their discriminant counterparts. The average self/peer convergent validity correlation was .43, substantially higher than the average absolute discriminant validity correlation of .09.[1] Similarly, the peer-peer convergent correlations averaged .36 whereas the discriminant correlations averaged .11.[2]

To test whether each scale accounted for the majority of variance in the appropriate validity criterion scale we performed 10 hierarchical regression analyses (one for each of the five scales for the self- and peer reports). Conceptually, we wanted to examine whether one perspective on a trait (e.g., TIPI-G *self*-ratings of A) accounted for the majority of variance in the other perspective on that same trait (e.g., TIPI-G *peer* ratings of A). Independent variables were the five

[1] In all cases, average correlations were computed using Fisher's r-to-z transformation.

[2] Because the peers were randomly assigned (i.e., each of the two peers was arbitrarily designated as either Peer 1 or Peer 2) we also calculated intraclass correlations (ICCs). The mean ICC (1,1) for the five convergent correlations was .35, and the mean ICC (1,1) for the 20 discriminant correlations was .10, supporting the findings of the correlational analyses.

Table 3. Confirmatory factor analyses of the TIPI-G scales

	Factors (Self)					Factors (Peer)				
	1	2	3	4	5	1	2	3	4	5
ES1: ängstlich, leicht aus der Fassung zu bringen*	.67					.81				
ES2: gelassen, emotional stabil	.71					.81				
E1: extravertiert, begeistert		.72					.75			
E2: zurückhaltend, still*		.56					.60			
O1: offen für neue Erfahrungen, vielschichtig			.69					.74		
O2: konventionell, unkreativ*			.55					.56		
A1: kritisch, streitsüchtig*				.42					.50	
A2: verständnisvoll, warmherzig				.63					.74	
C1: zuverlässig, selbstdiszipliniert					.72					.85
C2: unorganisiert, achtlos*					.67					.81

	Self-ratings					Peer ratings				
	ES	E	O	A	C	ES	E	O	A	C
ES	1.00	.28*	.48***	.21	.66***	1.00	.36***	.42***	.22*	.44***
E		1.00	.87***	.25	.11		1.00	.79***	.22	.01
O			1.00	.40**	.32**			1.00	.55***	.29**
A				1.00	.43**				1.00	.42***
C					1.00					1.00

Notes. ES = Emotional Stability; E = Extraversion; O = Openness; A = Agreeableness; C = Conscientiousness. Numbers denote whether item is the first or the second item of the corresponding scale. *N* (self) = 166; *N* (peer) = 180. Overall goodness of fit indices for the self-reports: χ^2 = 43.53 (df = 28; $p < .05$); χ^2/df = 1.56; goodness of fit index (GFI) = .95; Tucker-Lewis fit index (TLI) = .92; comparative fit index (CFI) = .95; root mean square error of approximation (RMSEA) = .06. Overall goodness of fit indices for the peer reports: χ^2 = 55.20 (df = 28; $p < .01$); χ^2/df = 1.97; goodness of fit index (GFI) = .94; Tucker-Lewis fit index (TLI) = .91; comparative fit index (CFI) = .95; root mean square error of approximation (RMSEA) = .07. Both models are specified with intercorrelated factors, no secondary loadings, and two correlated residuals. The correlated residuals for the self-reports were those of ES1/E2 (.42) and ES2/O2 (–.28). The correlated residuals for the peer reports were those of ES1/E2 (.33) and E2/A1 (–.29). The values for the peer sample are based on aggregated values.
* The item is recoded to conform to the scale name.

scales of the other perspective. First, the dependent variable was regressed upon the corresponding scale of the different perspective. In the next step, the four remaining scales were added to the regression equation. In 9 out of 10 cases the change in F was not significant, indicating that the predicted scales accounted for the majority of variance in the validity criterion scales. The one exception was peer-assessed TIPI-G C ($\Delta F = 2.66$; $p < .05$), which was predicted by both self-assessed TIPI-G C ($\beta = .53$) and (negatively) by self-assessed TIPI-G A ($\beta = -.16$), though it should be noted the impact of A was minimal with a corrected R^2 increase of only .03, from .22 to .25. For purposes of comparison the same procedure was carried out for the NEO-PI-R using the corresponding NEO scales of the other perspective. Here, 3 of the 10 comparisons reveal a second predictor.[3]

A confirmatory factor analysis (CFA) was performed on the TIPI-G self-reports. In accordance with other research concerning five-factor inventories (e.g., Church & Burke, 1994; Vassend & Skrondal, 1995) a model with correlated factors instead of a simple structure model was specified (cf. Table 3). Only two residual covariances had to be added. Kline (2005, p. 172) has pointed out that "models with factors that have only two indicators are more prone to estimation problems, especially when the sample size is small." Therefore, he recommends at least three indicators per factor. Although we are aware of this problem, we provide the CFA as an additional piece of evidence for the construct validity of the TIPI-G. Even if a model with five factors and two items each is theoretically overidentified such models are still susceptible to empirical underidentification. Such underidentification can occur if the correlation between some of the factors is close to zero. As predicted by these arguments, when we tried fitting a model we produced a Heywood case (negative variance); we, therefore, constrained the values of each pair of indicators to the same value. The goodness of fit indices for this model are good compared to other CFAs for five-factor inventories (cf. Table 3). Similar results can be shown for the aggregated peer reports: A model with correlated factors (no secondary loadings) and two residual covariances again

[3] Only one peer rater was available for the NEO-PI-R so we repeated the regression analyses for the TIPI-G using only the peer perspective of the one rater who had also completed the NEO-PI-R. The results did not change. Again, only peer-assessed TIPI-G C was predicted not only by the corresponding self-assessed TIPI-G C scale ($\beta = .47$) but also (negatively) by self-assessed TIPI-G A ($\beta = -.16$).

Table 4. Convergent and discriminant correlations between TIPI-G and NEO-PI-R

		TIPI self-ratings					TIPI peer ratings				
		ES_T	E_T	O_T	A_T	C_T	ES_T	E_T	O_T	A_T	C_T
NEO self	N_N	–.76***	–.12	–.08	–.05	–.28**	–.38***	.05	.29**	.09	–.07
	E_N	.17	.69***	.52***	.07	.03	.04	.38***	.21*	.16	–.16
	O_N	.02	.39***	.41***	–.07	–.04	–.12	.30**	.37***	–.05	–.19
	A_N	–.05	.00	.11	.51***	.13	–.14	–.05	.15	.40***	–.02
	C_N	.28**	–.07	–.03	.03	.68***	.03	–.14	–.21*	–.19	.45***
NEO peer	N_N	–.33***	–.09	–.03	.02	–.22**	–.77***	–.14	–.19*	–.23**	–.41***
	E_N	.10	.40***	.22**	.03	–.04	.35***	.69***	.52***	.20*	.06
	O_N	.04	.25**	.34***	.06	–.10	.09	.43***	.62***	.15	.02
	A_N	–.03	–.03	.01	.26**	–.05	.17*	–.11	.10	.70***	.12
	C_N	.21**	.02	–.01	–.05	.40***	.38***	.01	.19*	.19*	.76***

Notes. N = Neuroticism; ES = Emotional Stability; E = Extraversion; O = Openness; A = Agreeableness; C = Conscientiousness. *N* (TIPI self – NEO self) = 88; *N* (TIPI self – NEO peer) = 171; *N* (TIPI peer – NEO self) = 90; *N* (TIPI peer – NEO peer) = 172. For the peer-peer comparison only the TIPI-G values of the peer who also completed the NEO-PI-R were considered. * = $p < .05$; ** = $p < .01$; *** = $p < .001$. The subscripts T and N stand for the TIPI-G (T) and the NEO-PI-R (N).

produced good fit indices (Table 3). Both RMSEA values did not differ significantly from .05 (which can be considered a close fit; Browne & Cudeck, 1992) indicated by PCLOSE values of .32 for the self-reports and .09 for the peer reports.

Convergence Between the TIPI-G and the NEO-PI-R

Given our multijudge (self and peers), multiinstrument (TIPI-G and NEO-PI-R) design, four different convergent analyses are possible: Two analyses compare the instruments within a judge type (i.e., self-self; peer-peer) and two analyses compare the instruments across judge types (i.e., self-peer; peer-self). Overall, the convergent correlations are consistently and substantially stronger than the discriminant correlations (Table 4). As should be expected, the convergences computed within a judge type are stronger than the convergences computed across judge types. In only one case was a discriminant correlation stronger than an associated convergent correlation: the TIPI-G self O correlated .52 with the NEO self E but only .41 with the NEO self O. However, all other 79 discriminant correlations were lower than their related convergent counterparts. One interesting finding concerns the dimensions of Openness and Agreeableness: Whereas the convergent correlations of the peer ratings are rather high (.62 and .70, respectively) the convergent correlations of the self-ratings are lower, also in comparison with the other three dimensions (.41 and .51, respectively). All in all, the averaged convergent and discriminant correlations were: .62 vs. .13 (TIPI self/NEO self), .71 vs. .21 (TIPI peer/NEO peer), .35 vs. .08 (TIPI self/NEO peer), and .39 vs. .13 (TIPI peer/NEO self).

These results are supported by a series of 20 hierarchical regression analyses (five for each combination of judge type for the TIPI-G and the NEO-PI-R: self-self, peer-peer [where only TIPI-G values of the peer who completed the NEO-PI-R were included], self-peer, peer-self). In the first step, the TIPI-G scale was regressed upon the corresponding NEO-PI-R scale. In the second step, the remaining four NEO scales were added to the regression equation. The F change was significant for the second step in only four out of the 20 cases, and in each of these cases only one new predictor was significant. Specifically, in predicting TIPI-G self O, NEO self E (in addition to NEO self O) entered the regression equation ($\Delta F = 5.15$; $p < .01$); in predicting TIPI-G peer O, NEO peer E (in addition to NEO peer O) entered the equation ($\Delta F = 3.73$; $p < .01$), in another perspective combination in predicting TIPI-G peer O, NEO self N (in addition to NEO self O) entered the equation ($\Delta F = 2.93$; $p < .05$), and in predicting TIPI-G peer E, NEO peer A (in addition to NEO peer E) entered the equation ($\Delta F = 5.44$; $p < .01$).

To investigate the specific facets of the FFM tapped by the TIPI-G, we correlated the TIPI-G scales with the NEO-PI-R facet scores. The combination of six facets, five TIPI-G scales, and four judge perspectives results in 120 convergent correlations and 480 discriminant correlations, vastly increasing the possibility of Type I errors. Therefore, only correlations exceeding a significance level of .001 were considered; correlations meeting this standard are reported in Table 5.[4]

We first examined the convergent correlations. As shown in the first column of Table 6, there are four possible judge-instrument combinations. Here we highlight only those TIPI-G correlations with NEO-PI-R facets that reached the .001 significance threshold in more than one of the judge-instrument combinations. To emphasize the most replicable correlations, we have also italicized those correlations that occurred in more than two of the judge-instrument combinations:

[4] The full correlation table is available from the first author.

Table 5. Represented facets of the NEO-PI-R in the TIPI-G

	Combination	ES_T	E_T	O_T	A_T	C_T
Convergent	TIPI self/NEO self	N1 (–.70), N2 (–.61), N3 (–.68), N4 (–.52), N6 (–.73)	E1 (.40), E2 (.52), E3 (.48), E4 (.44), E6 (.57)	O4 (.37)	A3 (.51), A4 (.50)	C1 (.45), C2 (.62), C3 (.60), C4 (.44), C5 (.51), C6 (.47)
	TIPI peer/NEO peer	N1 (–.73), N2 (–.58), N3 (–.63), N4 (–.55), N6 (–.73)	E1 (.44), E2 (.50), E3 (.51), E4 (.47), E5 (.34), E6 (.54)	O1 (.31), O2 (.48), O3 (.44), O4 (.52), O5 (.41), O6 (.29)	A1 (.54), A2 (.42), A3 (.60), A4 (.53), A5 (.52), A6 (.43)	C1 (.58), C2 (.63), C3 (.74), C4 (.59), C5 (.71), C6 (.51)
	TIPI self/NEO peer	N1 (–.37), N3 (–.37), N6 (–.38)	E1 (.31), E3 (.27), E4 (.29), E6 (.32)	O2 (.28)	–	C1 (.31), C2 (.40), C3 (.34), C5 (.40), C6 (.28)
	TIPI peer/NEO self	N1 (–.43), N6 (–.45)	–	O2 (.39)	A5 (.39)	C2 (.44), C3 (.40), C5 (.39)
Discriminant	TIPI self/NEO self	C1 (.45)	–	E1 (.40), E2 (.42), E6 (.42)	E1 (.42), N2 (–.39)	N2 (–.41)
	TIPI peer/NEO peer	C1 (.47), C3 (.32), C4 (.29), C5 (.42), E1 (.32), E3 (.52), O5 (.27), O6 (.30)	A2 (–.28), A5 (–.31), C6 (–.37), N3 (–.27), N4 (–.39), O1 (.27), O2 (.29), O3 (.34), O4 (.34)	C1 (.29), C4 (.31), E1 (.41), E2 (.27), E3 (.38), E4 (.47), E6 (.40), N4 (–.27)	C3 (.30), E1 (.52), E6 (.33), N2 (–.53)	E3 (.27), E4 (.29), N2 (–.32), N3 (–.28), N5 (–.42), N6 (–.47), O1 (–.36)
	TIPI self/NEO peer	C1 (.31)	N5 (.27)	–	–	N5 (–.28), O1 (–.29)
	TIPI peer/NEO self	–	C6 (–.38), N5 (.38)	–	–	E5 (–.37)

Notes. *N* (self-self) = 88; *N* (peer-peer) = 172. *N* (self-peer) = 171; *N* (peer-self) = 90. For the peer-peer comparison only the TIPI values of the peer were considered who also completed the NEO-PI-R. ES = Emotional Stability; E = Extraversion; O = Openness; A = Agreeableness; C = Conscientiousness. NEO-PI-R facets: N1: Anxiety, N2: Angry Hostility, N3: Depression, N4: self-Consciousness, N5: Impulsiveness, N6: Vulnerability, E1: Warmth, E2: Gregariousness, E3: Assertiveness, E4: Activity, E5: Excitement-Seeking, E6: Positive Emotions, O1: Openness to Fantasy, O2: Openness to Aesthetics, O3: Openness to Feelings, O4: Openness to Actions, O5: Openness to Ideas, O6: Openness to Values, A1: Trust, A2: Straightforwardness, A3: Altruism, A4: Compliance, A5: Modesty, A6: Tender-Mindedness, C1: Competence, C2: Order, C3: Dutifulness, C4: Achievement Striving, C5: self-Discipline, C6: Deliberation. $p < .001$ for all correlations. The subscript T stands for the TIPI-G.

– for TIPI-G ES: *N1: Anxiety*, N2: Angry Hostility, *N3: Depression*, N4: Self-Consciousness, *N6: Vulnerability*;
– for TIPI-G E: *E1: Warmth*, E2: Gregariousness, *E3: Assertiveness*, *E4: Activity*, *E6: Positive Emotions;*
– for TIPI-G O: *O2: Openness to Aesthetics*, O4: Openness to Actions;
– for TIPI-G A: A3: Altruism, A4: Compliance, A5: Modesty;
– for TIPI-G C: *C1: Competence*, *C2: Order*, *C3: Dutifulness*, C4: Achievement Striving, *C5: Self-Discipline*, *C6: Deliberation*.

In sum, there are strong and consistent convergent relationships between the TIPI-G scales and certain NEO facets. However, as is to be expected for very brief scales, the full range of content captured by the 30 NEO facets cannot be represented by the 5 two-item scales. The TIPI-G scales seem to capture the meanings of ES, E, and C as they are defined by the NEO-PI-R better than they capture the NEO-PI-R definitions of O and A.

We next examined the discriminant correlations. Again, we highlight here only those correlations that reached our .001 significance threshold more than once in the four possible judge-instrument combinations, italicizing those that appeared more than twice. In addition, we denote negative correlations with an "R":

– for TIPI-G ES: *C1: Competence*;
– for TIPI-G E: C6: Deliberation (R), N5: Impulsiveness;
– for TIPI-G O: E1: Warmth, E2: Gregariousness, E6: Positive Emotions;
– for TIPI-G A: E1: Warmth, N2: Angry Hostility (R);
– for TIPI-G C: N2: Angry Hostility (R), N5: Impulsiveness (R), O1: Openness to Fantasy (R).

First, as expected, there are many fewer discriminant correlations between the TIPI-G scales and theoretically unrelated facets. The paucity of significant discriminant correlations is especially striking given that there are four times more discriminant correlations than convergent correlations – whereas 70% (21/30) of the convergent facets demonstrated consistently significant relationships the same is true for only 9% (11/120) of the discriminant facets.[5]

To determine whether the TIPI-G and the NEO-PI-R weighed the facets similarly, we analyzed the pattern of correlations between the NEO-dimensions and the NEO-

[5] Our procedure is rather conservative because we did not correct the *p* level for these analyses to account for the fact that there are four times as many possible discriminant correlations as convergent correlations.

facets (Ostendorf & Angleitner, 2004, p. 109 and p. 134) on the one hand and the correlations between the TIPI-G scores and the NEO-facets on the other. The median of the rank correlation (Spearman's ρ) between the columns of these two matrices was .92 for self-report-data (between .74 for O and .95 for N) and .89 for the peer-data (between .72 for O and .95 for E). Admittedly, a rank correlation does not take into account the absolute strength of the correlations. The alternative use of structural equation modeling techniques either led to calculation errors or showed significant differences between a saturated model and a model where the covariances between the TIPI-G scale and the NEO-PI-R facets were fixed to the estimated covariances between the corresponding NEO-PI-R scale and its facets. Finally, we compared the pattern of correlations between the correlations between the original TIPI self-report scores and the NEO-facets (here administered 6 weeks later) on the one hand and the correlations between the TIPI-G self-report scores and the NEO-facets on the other. The median of the rank correlation between the columns of these two matrices is .81 for self-report-data (between .69 for A and .82 for C). These ρ values are a bit lower, but still quite substantial. It should be noted that this comparison included two instruments in two languages as well as a 6-week interval between the collection of the original TIPI and NEO data.

Discussion

The present study documented the successful transfer of a very brief measure of the FFM personality domains (Gosling et al., 2003) from the United States to Germany. A self/peer study yielded convergent correlations that substantially exceeded the discriminant correlations. Moreover, cross-validation analyses using the NEO-PI-R as the standard corroborated previous findings based on the original TIPI. Together the analyses suggest that the ten unipolar items of the TIPI-G can provide an efficient approximation for longer measures of the FFM personality constructs.

The clear overlap between E and O (which can also be found in the original TIPI) is not uncommon. Actually, the observed correlations of E and O in the German NEO manual are of comparable size: .40 for self-reports (as in the original NEO-PI-R) and .41 for peer reports (Ostendorf & Angleitner, 2004). These correlations are consistent with Digman's (1997) analyses suggesting the existence of two higher order factors of the Big Five: α, a socialization factor, which consists of Conscientiousness, Emotional Stability, and Agreeableness, and Beta, a self-actualizing factor, which consists of Extraversion and Openness.

However, as could be predicted on psychometric grounds, a two-item instrument like the TIPI-G cannot attain the levels of accuracy achieved by longer instruments such as the 48-item scales of the NEO-PI-R (Costa & McCrae, 1992). Unsurprisingly, the 10-item TIPI-G did not capture all of the facets assessed by the 240-item NEO, although the TIPI-G did successfully capture the cores of the broader dimensions. Moreover, the small number of significant discriminant correlations suggests that the TIPI-G is assessing constructs very similar to those assessed by the NEO-PI-R. The finding that the convergent correlations for O and A between the TIPI-G and the NEO-PI-R are higher for the peer ratings than for the self-ratings may suggest that the TIPI-G captures information that is visible from an external point of view and attenuated in the self-perspective. A reason for this could have been that the mean values of the TIPI-G self-reports for both dimensions would be higher and the standard deviations would be lower than the corresponding values for the other three dimensions. However, the mean value for C is even higher and its standard deviation even a bit lower compared to the corresponding values for O and A. Further studies should examine whether these results can be generalized to other samples. Because of the specific sample based on a snowball recruitment procedure by 20 students of economics the generalizability of the results of this study might be limited.

The results of the current research seem to contradict results recently published in this journal by Herzberg and Brähler (2006). They concluded that they "... could not recommend the use of the German TIPI as a proxy for longer Big-Five measures, because of low reliabilities of the scales. Furthermore, based on an independent sample the low convergence with the well-validated Big-Five measure NEO-FFI indicates a lack of validity" (p. 144). Specifically, Herzberg and Brähler attribute these and other observed insufficient psychometric properties to "... the fact that two adjectives have to be evaluated simultaneously" (p. 147). However, the psychometric problems associated with the Herzberg and Brähler instrument might be attributable to their specific translation rather than to the fact that two adjectives were included in each item. There are certainly differences between the Herzberg and Brähler translation and ours. For example, Herzberg and Brähler translated *quiet* as *ruhig* and *sympathetic* as *einfühlsam*; we translated *quiet* as *still* and *sympathetic* as *verständnisvoll*. The internal consistencies are higher for our translation than for Herzberg and Brähler's translation and they are comparable to the original English language TIPI. Furthermore, convergence of our TIPI-G with a NEO Inventory is higher than for Herzberg and Brähler's translation of the TIPI. Additionally, the reasoning that it is the simultaneous evaluation of two adjectives that accounts for the insufficient psychometric properties disregards the successful development of other short Big Five instruments where even more than two descriptors per item are used (e.g., Woods & Hampson, 2005). This reasoning also disregards the validation evidence from the original English language TIPI, which used two descriptors per item. In short, when combined with the results of the current study, the available evidence suggests that the deficiencies reported by Herzberg and Brähler may be the result of their particular translation of the TIPI rather than of the TIPI itself.

Obviously, when comprehensive assessments of personality are required (e.g., for counseling or in-depth diagnostic purposes), the TIPI-G is not an appropriate instrument. Furthermore, the TIPI-G would not be appropriate in situations where facets afford better predictions than do dimensions (e.g, Paunonen, Haddock, Forsterling, & Keinonen, 2003); in such situations, where there are theoretical grounds for predicting the superiority of specific facets, narrower assessments should be made. However, facet-level predictions often cannot be made. In such circumstances, an instrument with a broad coverage could be more appropriate. If so – and if the available time for assessment is short – the TIPI-G is one sensible option.

What are the benefits of the TIPI-G? Given the instrument has only 10 items, it incurs very little fatigue or other costs on participants. These advantages will be particularly salient in research where time is limited or where personality could influence the findings but is not the central focus. More generally, the TIPI-G permits the measurement of the FFM in circumstances where they could not formerly have been measured. The TIPI-G could even be applied in telephone opinion surveys (e.g., for market research), following them up with longer instruments if necessary. However, until the TIPI has been tested in the relevant target populations, these benefits are speculative. The adjectives might be difficult to understand for subjects with low verbal ability and may differ in their meaning across cultural subgroups.

An advantage of the TIPI is its widespread distribution; several translations of the TIPI exist although most of these have yet to be validated (http://homepage.psy.utexas.edu/homepage/faculty/Gosling/scales_we.htm). Nonetheless, the existence of a brief instrument translated into many languages holds great promise for integrating research findings across many domains and cultures.

References

Becker, P. (2003). *TIPI – Trierer Integriertes Persönlichkeitsinventar* [Trier Integrated Personality Inventory]. Göttingen: Hogrefe.

Browne, M.W., & Cudeck, R. (1992). Alternative ways of assessing model fit. *Sociological Methods and Research, 21*, 230–258.

Burisch, M. (1984). 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.

Church, A.T., & Burke, P.J. (1994). Exploratory and confirmatory facts of the Big Five and Tellegen's three- and four-dimensional models. *Journal of Personality and Social Psychology, 66,* 93–114.

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.

Digman, J.M. (1997). Higher-order factors of the Big Five. *Journal of Personality and Social Psychology, 73,* 1246–1256.

Gardner, D.G., Cummings, L.L., Dunham, R.B., & Pierce, J.L. (1998). Single-item versus multiple-item measurement scales: An empirical comparison. *Educational and Psychological Measurement, 58*, 898–915.

Goldberg, L.R. (1992). The development of markers for the Big-Five factor structure. *Psychological Assessment, 4*, 26–42.

Gosling, S.D., Rentfrow, P.J., & Swann, W.B. Jr. (2003). A very brief measure of the Big-Five personality domains. *Journal of Research in Personality, 37*, 504–528.

Herzberg, P.Y., & Brähler, E. (2006). Assessing the Big-Five personality domains via short forms. A cautionary note and a proposal. *European Journal of Psychological Assessment, 22*, 139–148.

Höft, S. (2002). *Grundlagen einer persönlichkeitsorientierten Berufseignungsdiagnostik. Verhaltens- und berufsbezogene Aspekte des Fünf-Faktoren-Modells der Persönlichkeit* [Fundamentals of a personality-oriented assessment of vocational skills: Behavioral and job-related aspects of the five-factor model of personality]. Berlin: dissertation.de.

Hofstee, W.K.B., Kiers, H.A.L., De Raad, B., Goldberg, L.R., & Ostendorf, F. (1997). A comparison of Big Five structures of personality traits in Dutch, English, and German. *European Journal of Personality, 11*, 15–31.

John, O.P., & Benet-Martinez, V. (2000). Measurement, scale construction, and reliability. In H.T. Reis & C.M. Judd (Eds.), *Handbook of research methods in social psychology* (pp. 339–369). New York: Cambridge University Press.

John, O.P., Goldberg, L.R., & Angleitner, A. (1984). Better than the alphabet: Taxonomies of personality-descriptive terms in English, Dutch, and German. In H. Bonarius, G. Van Heck, & N. Smid (Eds.), *Personality psychology in Europe: Theoretical and empirical developments* (Vol. 1, pp. 83–100). Lisse: Swets and Zeitlinger.

John, O.P., & Srivastava, S. (1999). The Big Five trait taxonomy: History, measurement, and theoretical perspectives. In L.A. Pervin & O.P. John (Eds.), *Handbook of personality* (2nd ed., pp. 102–138). New York: Guilford.

Kline, R.B. (2005). *Principles and practice of structural equation modeling* (2nd ed.). New York: Guilford.

Langford, P.H. (2003). A one-minute measure of the Big Five? Evaluating and abridging Shafer's (1999) Big Five markers. *Personality and Individual Differences, 35*, 1127–1140.

Muck, P.M., Hell, B., & Höft, S. (in press). Application of the principles of Behaviorally Anchored Rating Scales to assess the Big Five personality constructs at work. In J. Deller & D.S. Ones (Eds.), *Personality@work.* Mering: Hampp.

Ostendorf, F. (1990). *Sprache und Persönlichkeitsstruktur: Zur Validität des Fünf-Faktoren-Modells der Persönlichkeit* [Language and personality structure: On the validity of the five-factor model of personality]. Regensburg: Roderer.

Ostendorf, F., & Angleitner, A. (2004). *NEO-Persönlichkeitsinventar nach Costa und McCrae, Revidierte Fassung (NEO-PI-R)* [NEO Personality Inventory by Costa and McCrae, revised version (NEO-PI-R)]. Göttingen: Hogrefe.

Paunonen, S.V., Haddock, G., Forsterling, F., & Keinonen, M. (2003). Broad versus narrow personality measures and the prediction of behavior across cultures. *European Journal of Personality, 17*, 413–433.

Ramírez-Esparza, N., Gosling, S.D., Benet-Martínez, V., Potter,

J.P., & Pennebaker, J.W. (2006). Do bilinguals have two personalities? A special case of cultural frame-switching. *Journal of Research in Personality, 40,* 99–120.

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 scale for the measurement of the Big Five personality dimensions in surveys]. *ZUMA-Nachrichten, 55*, 5–28.

Saucier, G. (1994). Mini-markers: A brief version of Goldberg's unipolar Big Five markers. *Journal of Personality Assessment, 63*, 506–516.

Shafer, A.B. (1999). Brief bipolar markers for the five-factor model of personality. *Psychological Reports, 84*, 1173–1179.

Vassend, O., & Skrondal, A. (1995). Factor analytic studies of the NEO Personality Inventory and the five-factor model: The problem of high structural complexity and conceptual indeterminacy. *Personality and Individual Differences, 19*, 135–147.

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.

Peter M. Muck

University of Bielefeld
Faculty of Psychology and Sports Science
Work and Organizational Psychology
D-33501 Bielefeld
Germany
E-mail peter.muck@uni-bielefeld.de

Benedikt Hell

University of Hohenheim
Department of Psychology (540F)
D-70593 Stuttgart
Germany
E-mail hell@uni-hohenheim.de

Samuel D. Gosling

The University of Texas at Austin
Department of Psychology
1 University Station A8000
Austin, TX 78712-0187
USA
E-mail samg@mail.utexas.edu

Appendix

The Original TIPI items and Their German Translation

Item	Original item	German item
1	extraverted, enthusiastic	extravertiert, begeistert
2	critical, quarrelsome	kritisch, streitsüchtig
3	dependable, self-disciplined	zuverlässig, selbstdiszipliniert
4	anxious, easily upset	ängstlich, leicht aus der Fassung zu bringen
5	open to new experiences, complex	offen für neue Erfahrungen, vielschichtig
6	reserved, quiet	zurückhaltend, still
7	sympathetic, warm	verständnisvoll, warmherzig
8	disorganized, careless	unorganisiert, achtlos
9	calm, emotionally stable	gelassen, emotional stabil
10	conventional, uncreative	konventionell, unkreativ

Note. Scoring ("R" denotes reverse-scored items): Extraversion: 1, 6R; Agreeableness: 2R, 7; Conscientiousness: 3, 8R; Emotional Stability: 4R, 9; Openness to Experience: 5, 10R. The 7-point scale has been translated as follows: 1 = trifft überhaupt nicht zu; 2 = trifft größtenteils nicht zu; 3 = trifft eher nicht zu; 4 = weder zutreffend noch unzutreffend; 5 = trifft eher zu; 6 = trifft größtenteils zu; 7 = trifft voll und ganz zu. The introductory statement is: "Ich sehe mich selbst als:".