European Journal of Personality
Eur. J. Pers. **22**: 37–54 (2008)
Published online 21 August 2007 in Wiley InterScience
(www.interscience.wiley.com) **DOI**: 10.1002/per.662

Single-item Big Five Ratings in a Social Network Design

JAAP J. A. DENISSEN[1]*, RINIE GEENEN[3],
MAARTEN SELFHOUT[2] and MARCEL A. G. VAN AKEN[1]

[1] *Department of Developmental Psychology, Utrecht University, The Netherlands*
[2] *Department of Child and Youth Studies, Utrecht University, The Netherlands*
[3] *Department of Clinical and Health Psychology, Utrecht University, The Netherlands*

Abstract

To develop and validate an ultra-short measure to assess the Big Five in social network designs, the unipolar items of the Ten-Item Personality Inventory were adapted to create a bipolar single-item scale (TIPI-r), including a new Openness item. Reliability was examined in terms of the internal consistency and test–retest stability of self-ratings and peer-rating composites (trait reputations). Validity was examined by means of convergence between TIPI-r and Big Five Inventory (BFI) scores, self-peer agreement and projection (intra-individual correlation between self- and peer-ratings). The psychometric quality of the TIPI-r differed somewhat between scales and the different reliability and validity criteria. The high reliability of the peer-rating composites motivates to use the TIPI-r in future studies employing social network designs. Copyright © 2007 John Wiley & Sons, Ltd.

Key words: social groups; personality scales and inventories; multilevel analysis

INTRODUCTION

A relative consensus has emerged about the usefulness of the Five Factor Model (FFM) to measure personality traits (Costa & McCrae, 1995). However, traditional FFM questionnaires such as the NEO-PI-R (Costa & McCrae, 1992) have a large number of items, which is not practical in situations with high demands on participants' time, motivation and cognitive resources. Shorter versions have been introduced, such as the Big Five Inventory (BFI; John & Srivastava, 1999) with 44 items and the Big Five mini-markers with 40 items (Saucier, 1994). Ultra-short versions with one or two items for each FFM dimension have been developed (Gosling, Rentfrow, & Swann, 2003; Rammstedt & John, 2007; Rammstedt, Koch, Borg, & Reitz, 2004; Woods & Hampson, 2005). These can be economically used to assess personality traits using traditional self-reports but also within social networks employing peer-rating composites (trait reputations).

*Correspondence to: Jaap J. A. Denissen, Department of Developmental Psychology, Utrecht University, The Netherlands. E-mail: j.j.a.denissen@fss.uu.nl

Received 8 March 2007
Revised 2 July 2007
Accepted 10 July 2007

The current paper evaluates an adaptation of the Ten Item Personality Inventory (TIPI; Gosling et al., 2003) in a social network design. Having discussed the nature of ultra-short instruments, and systematically reviewed reliability and validity criteria to assess the psychometric properties of such measures, we will discuss unique opportunities and demands of using ultra-short instruments in a social network design, both in terms of the information processing demands they impose on participants and the kind of statistical analysis needed. We will suggest that bipolar single-item FFM indicators can be employed in a social network design without a deterioration of their psychometric properties when compared to previous ultra-short Big Five measures employed in single-target designs.

To reduce the number of items needed to assess a construct while avoiding acquiescence bias, some researchers have combined adjectives that are semantic opposites into bipolar scales (e.g. Goldberg, 1992). Designers of ultra-short FFM scales have gone even further by combining more than two adjectives into a single item (SI). For example, the Extraversion item of Gosling et al.'s (2003) single-item measure consists of the following eight adjectives: 'extraverted, enthusiastic (i.e. sociable, assertive, talkative, active, NOT reserved or shy)'. SI questionnaires pose a methodological challenge, as the standard practice of traditional personality scales is to use singular adjectives or statements to avoid interpretation ambiguities. The large number of adjectives per item requires that raters are able to mentally construct a valid representation of the underlying personality dimension.

The combination of traditional and unique psychometric standards is needed to examine reliability and validity of ultra-short FFM measures. Although reliability is often treated as a unitary construct, there are three types of reliability stemming from different sources of error (Charter, 2003). First, internal consistency refers to the degree of content overlap between the items of a test and if most often assessed as Cronbach's alpha (though it can also be assessed as split-half reliability). In the case of the 2-item scales of the TIPI, the amount of internal consistency usually fails to meet traditional benchmarks (e.g. ranging between .40 and .73 in the study by Gosling et al., 2003). It is, however, unrealistic to expect high alphas when using short instruments that are designed to measure very broad domains (Gosling et al., 2003).

An average inter-item correlation is useful for comparing across measures of different length, but in the case of SIs, it is impossible to calculate an average correlation. As an alternative, Woods and Hampson (2005) proposed to include SIs in a factor analysis of the items of a longer FFM instrument and use their communalities as a lower-bound estimate of internal consistency. Communalities indicate the share of the variance in an item that is explained by all extracted factors. However, they actually represent *upper-bound* estimates of internal consistency because they are inflated by secondary loadings on factors to which they do not conceptually belong. Instead, we suggest to use the square of the main factor loading of a SI as an estimate of reliability, as this can be constructed (using simple path logic) as the correlation between two parallel item versions. From the formula to calculate Cronbach's alpha,[1] it follows that this estimate equals the proportion of true score variance captured by a SI ('internal consistency').

[1] Equation (1):

$$\alpha = \frac{N\bar{r}}{(1 + (N-1)\bar{r})}$$

where N is the number of items in a scale and $\bar{r}$ is the average inter-item correlation. For a single item scale ($N = 1$), α thus equals $\bar{r}$.

A second class of reliability indices looks at the degree to which a construct is temporally stable and relatively unaffected by random (state) fluctuations around an individual's (trait) mean. The short-term test–retest correlation of a measure is usually used as an index of this kind of reliability. To estimate the test–retest reliability of existing ultra-short FFM measures, we carried out an overview across six different studies (Table 1). An average retest correlation of .73 was found, demonstrating adequate or perhaps even high reliability given the brevity of short FFM measures and that participants in most studies assessed a single target individual. It cannot be excluded that memory effects contribute to an overestimation of retest reliability.

A third way to assess reliability is to look at the degree to which different judges of a target agree with their assessments instead of being guided by idiosyncratic considerations. The composite of their ratings of judges for a particular target individual can be regarded as that individual's trait *reputation* (Hogan, 1996). Every peer who contributes to this composite can be seen as an item in the psychometric sense of the word, contributing to both the true score (i.e. a person's actual reputation) and error variance (i.e. idiosyncratic rating tendencies). These influences are comparable to target and perceiver effects in Kenny's (1994) social relations model framework (SRM), though they are based on different computations.[2]

If the number of peers is equal across social networks, Cronbach's alpha can be computed to compare the reliability of the trait reputation composites. Hierarchical linear modelling (HLM) offers a novel way to deal with an unequal number of raters (as was the case in the current study), because it takes into account the nested structure of the data (i.e. every person is assessed by multiple peers, whose ratings are therefore interdependent). For every target person, an intercept is estimated based on the average peer-ratings of that person's personality.[3] HLM calculates the reliability of this intercept by considering both the number of data points (i.e. raters) on which it is based as well as the variance around it (i.e. deviations from this reputation in the eye of individual peers).[4]

[2] Using only peer ratings, average perceiver effects differ somewhat between target individuals as they are based on a different set of raters (they exclude the target individual). To compensate for this bias, the calculation of target effects in SRM includes the self-rating of the corresponding person (Warner, Kenny, & Stoto, 1979, p. 1747). In our study, these fluctuations in perceiver effects are likely diluted by the large pool of peer raters.

[3] HLM uses the following equations to model specific trait ratings as well as people's average trait reputation intercept:
Equation (2):

$$y_{ij} = \beta_{0j} + r_{ij}$$

Equation (3):

$$\beta_{0j} = \gamma_{00} + u_{0j}$$

where y is a the observed rating of peer i of target individual j, β_{0j} is the average peer rating (i.e. trait reputation) for that target individual, r_{ij} is the deviation of peer i's rating of target j from that average level, γ_{00} is the average rating across all targets and raters, and u_{0j} is the deviation of target individual j's departure from that average.

[4] HLM uses the following equation to calculate the reliability of trait reputations:
Equation (4):

$$\beta_{0j} = \sum_{1}^{j} \frac{\tau_{00}/[\tau_{00} + (\sigma^2/n_j)]}{j}$$

where j is the total number of Level 2 units (i.e. participants), τ_{00} is the variance between these units (i.e. individual differences in reputation), σ^2 is the variance within these units (i.e. deviations from this reputation in the eyes of individual peers) and n_j is the sample size of a particular Level 2 unit (i.e. the number of peer ratings per participant).

DOI: 10.1002/per

Table 1. Overview of psychometric properties of ultra-short FFM Instruments across studies

		TIPI								
	FIPI (Gosling et al., 2003)[†]	(Muck et al., in press)	(Herzberg & Brähler, 2006)	(Gosling et al., 2003)[†]	(Woods & Hampson, 2005)	SIMP (Woods & Hampson, 2005)	(no name) (Rammstedt et al., 2004)	BFI-10 (Rammstedt & John, 2007)[¶]	Average of previous studies	TIPI-r Current study (self-ratings)
Convergent *r*	.66	.62	.40	.77	.66	.64	.66[§]	.68[‖]	.65	.61
Discriminant *r*	.27	.13	.07	.20	.14	.11	.18[§]	.14	.16	.08
Retest interval (weeks)	2	—	16.4	6	—	4	6	6–8	6.9	4
Retest *r*	.68	—	.78	.72	—	.71[‡]	.74	.75	.73	.68
Peer rater	S	R, F, or C	F or R	—	—	F/S	—	F or P		C
Self-peer agreement	.26	.43	.67	—	—	.38/.22	—	.44	.43	.40

Note: R, relative; F, friend; C, colleague/fellow student; S, stranger; P, partner.

[†] Across four samples.

[‡] Values were also reported for 3 months, 9 months and 1 year; results were almost identical.

[§] Average across BFI/NEO-FFI.

[¶] Across five samples.

[‖] Correlations with NEO-PI-R (correlations with the BFI full scale are inflated because of item overlap).

Besides the use of a measure in predictive research to assess its criterion validity, the validity of single-item measures can be assessed in one of three ways. First, as a measure of *convergent validity*, the correlation between SIs and the corresponding scales of longer FFM instruments indicates the ability of a SI to capture the (semantic or psychological) core construct underlying a FFM dimension as operationalized by multi-item scales. Complementary to convergent validity, *discriminant validity* is shown when the correlations between SIs and multi-item scales tapping into other FFM factors (i.e. off-diagonal correlations) are low (Campbell & Fiske, 1959). In eight studies using FFM questionnaires, the SIs had an average convergent validity correlation of .65 and a discriminant validity correlation of .16 (see Table 1), which is quite acceptable given their short length.

The only exception to the generally acceptable level of convergent validity is the Openness factor. In the studies by Gosling *et al.* (2003), Muck, Hell and Gosling (Muck, Hell, & Gosling, in press) and Woods and Hampson (2005), correlations between SIs and multi-item scales ranged between .41 and .64, whereas Herzberg and Brähler (2006) reported a correlation of only .23. In our own pilot work, the Dutch translation of the Openness item also performed poorly compared to the other scales. According to Muck et al. (in press), this lack of convergent validity is due to conceptual ambiguities regarding the content of this trait. The current study examined whether a more valid single-item Openness measure can be construed by focusing on the core of the construct of the multi-item scale identified by careful content analysis of the BFI.

A second way to assess the validity of SIs is by examining the degree of convergence between self and peers. The logic behind this approach is that single-item measures capture something ‘real’ when they converge across different observers. If self-ratings do not converge with peer-ratings, however, this does not necessarily disprove validity, as it may be that peer raters base their judgments on a different source of information than the self-rater (Kenny, 2004, offers an overview of different influences on trait judgment). For example, participants may use perceptions of their own feelings of depression and anxiety to infer their level of neuroticism, but this source of information may not be accessible to peers. Of course, the visibility of a trait also depends on whether raters have observed the targets in contexts that allow the enactment of valid behavioural cues (Funder & Dobroth, 1987; John & Robins, 1993).

Table 1 summarizes self-peer agreement correlations across five different studies. An average agreement correlation of .43 is reached, which is comparable to the average agreement correlation of .37 between self and parental personality ratings with traditional multi-item scales (Funder, Kolar, & Blackman, 1995). It should be noted, however, that previous studies mostly let participants nominate friends or family members to provide peer-ratings. Using raters who are well-acquainted with the target may lead to higher levels of agreement (because people gain access to valid cues to each other’s personality). However, participants may selectively invite others who agree with their self-ratings to provide personality judgments, leading to inflated levels of self-peer agreement over and above the level predicted by shared access to valid behavioural cues by more neutral observers. To avoid this possible source of bias in the present study, we used raters who were randomly assigned to their targets yet knew them well enough to make informed judgments.

When individuals provide both self- and peer-ratings (e.g. in round robin designs, see below), a third way to assess an item’s validity is by calculating the degree of projection (also called assumed similarity by Kenny, 1994). When an item is clearly and

DOI: 10.1002/per

understandably formulated, most participants will use more or less similar available behavioural cues to rate the corresponding trait level of each network partner (Funder, 1995). If the formulation of an item is ambiguous, however, participants are not able to make valid distinctions between peers. One way to fill this informational gap is to *project* their own values onto their peers, leading to a high correlation between participants' self-ratings and the average rating level of their peer-ratings (projection; Kruger & Clement, 1994). A similar high correlation can occur if the rater has difficulty observing valid diagnostic behaviours for the trait in question, either because the trait description taps into covert psychological processes or because the target individual exhibits these behaviours in settings in which the rater is not present.

Opportunities and demands of ultra-short questionnaires in social network designs

Social network designs are frequently employed in the social and behavioural sciences (Wasserman & Faust, 1994). In some cases, such designs involve the generation of a list of all members of their ego-centred social network, such as parents, friends and colleagues (Neyer, 1997). In other cases, the composition of the social network is known in advance and participants complete ratings of every network partner on a on a number of separate dimensions (see Appendix A for an example). Using paper-and-pencil measures, these ratings are typically applied in separate columns. Computer-based methods allows for separate screens for each rating dimension.

The data collected with social network designs can be analysed in a number of exciting ways. Van Duijn, van Busschbach, and Snijders (1999) demonstrated that social network data can be modelled using a multilevel framework. The sample size of this statistically powerful approach (on the so-called Level 1) is $N \times k$, where N is the number of participants, and k is the average number of network partners per participant. When participants rate both themselves and their network partners on the same dimensions, the difference between self- and peer-ratings reflects the degree of dyadic similarity. In a round robin design, every individual within a social network rates every other individual as well as him-/herself (Kenny, 1996). This design allows to disentangle actor effects (individual differences in rating tendencies), partner effects (differences in the way people are seen by others) and relationship effects (influences that cannot be reduced to actor or partner effects because they are dyadic in nature) (Kenny, 1994).

In traditional designs, respondents rate a single target (i.e. themselves or a peer) on a number of characteristics (i.e. items). Usually, the instruction asks participants to compare the target person to a specific reference group, such as peers of the same age and gender. In social network designs, this logic is turned upside down, since respondents rate a single characteristic (i.e. a trait) in numerous targets (i.e. network partners). These other members of the social network form an explicit reference group. If this reference group is large and representative enough, this may lead to more concrete and, thus, accurate ratings than in the traditional design.

The use of single-item ratings in social network designs places a number of demands on participants that are not apparent in more traditional designs. Participants in social network designs have to compare numerous target individuals on a rating dimension, which is more cognitively demanding than rating a single target. Single-item scales that include a large number of adjectives may overburden participants in such a case. For example, Rammstedt et al.'s (2004) bipolar measure consists of a minimum of 10 adjectives per item. As humans

DOI: 10.1002/per

are only capable of holding 7 ± 2 discrete units of information in memory (Miller, 1956), participants are expected to take time to 'chunk' this information into more manageable units. The relatively long time it takes to complete some of the short FFM measures (e.g. 4 minutes in the study by Rammstedt et al., 2004) suggests this is not an easy task. While most participants are apparently able to do this when rating single targets, it is doubtful whether this functions equally well when rating larger groups of individuals. In the current paper, we therefore relied on the TIPI (Gosling et al., 2003), which consists of only two adjectives per FFM factor and takes only 1 minute per target person to complete.

The current study

The current study examines the reliability and validity of the TIPI-r, including a new item to measure Openness. Freshman students used this instrument to rate their own and their peers' Big Five traits during their first year at the university. We will examine the reliability of a single-item FFM measure in terms of the traditional benchmarks of internal consistency and retest stability as well as in terms of the consistency of each participant's trait reputation (i.e. composite of all peer-ratings). To assess the validity of the SI measures, we will examine convergent validity between the single-items and longer scales and between self- and peer-ratings. Beyond these usual strategies, we will calculate the degree of projection (i.e. the degree of similarity between individuals'-ratings of themselves and their peers). A unique feature of the current study is that we used previously unacquainted individuals as raters, thus avoiding a reliance on self-selected samples of peers that may be associated with inflated levels of self-peer agreement (Swann, 1987). This is the first evaluation of the psychometric properties of a single-item FFM questionnaire in Dutch. Findings will indicate to what degree our results can be generalized to results obtained with English and German versions.

METHOD

Sample

Participants were psychology freshmen who started their study in the autumn of 2006. The students had been assigned to introduction groups of around 25 people in order to facilitate social adjustment. These groups work together during the remainder of the year to complete a substantial part of the psychology curriculum. A total of 489 individuals were assigned to one of 20 groups. E-mails, flyers, posters and an announcement during one of the first university lectures generated attention for the current study. Participants received 20€ (around 25$), 2 hours of course credit, and a personality feedback profile at the end of the study. Participants registered for the study on a website. The 10 groups in which more than 80% of the participants registered for the current study were selected for participation. Out of 238 active group members (defined as being recognized by more than 80% of peers), 221 individuals registered for the current study (93% enrolment rate). The mean age of these individuals was 18.9 (SD = 1.6), with 181 (82%) females. The majority of participants (92%) were of Dutch origin. Only five pairs of group members reported that they had known each other before the start of the study.

After 4 months (the time of the current Wave 5 retest, see below), 13 individuals had quit their psychology education. Of the 225 remaining group members, 205 individuals

DOI: 10.1002/per

continued as participants (91% participation rate). Compared to these 205 participants, the remaining 20 non-participating group members were rated by their peers as significantly less neurotic (3.35 vs. 3.66, $F = 5.67$, $p = .02$) and conscientious (3.94 vs. 4.72, $F = 22.74$, $p < .01$). No differences were found for the other Big Five factors.

Instruments

Big Five Inventory

Participants completed the 44-item Dutch translation (Denissen, Geenen, van Aken, Gosling, & Potter, in press) of the BFI (John & Srivastava, 1999). This instrument consists of eight statements for the factors Extraversion (sample item: 'is talkative') and Neuroticism (sample item: 'can be moody'), nine statements for the factors Conscientiousness (sample item: 'does a thorough job') and Agreeableness (sample item: 'is generally trusting') and 10 statements for the factor Openness (sample item: 'values artistic, aesthetic experiences'). Participants indicated their agreement regarding each statement on a 1 ('strongly disagree') to 5 ('strongly agree') Likert scale. Table 2 presents means, standard deviations and internal consistencies (Cronbach's alphas) of the scales.

Translation and adaptation of Ten Item Personality Inventory

To construct SI Big Five indicators that are relatively low in complexity, the first, second and fourth authors translated the 10 items of the TIPI (Gosling et al., 2003), consisting of two adjectives per FFM factor. This translation was back-translated by one English native speaker and one Dutch person living in the United States. Differences were discussed and resolved by consensus by the Dutch authors and English speakers. To further reduce time demands on participants, both items belonging to a FFM domain were combined into a single bipolar rating scale (Extraversion: 'extraverted, enthusiastic' vs. 'reserved, quiet'; Agreeableness: 'critical, quarrelsome' vs. 'sympathetic, warm'; Conscientiousness: 'dependable, self-disciplined' vs. 'disorganized, careless'; Neuroticism: 'anxious, easily upset' vs. 'calm, emotionally stable'; Openness to Experience: 'open to new experiences, complex' vs. 'conventional, uncreative').

New openness item

Because of the low psychometric performance of the original TIPI Openness scale, we created a new Openness item. To maximize convergent validity, a content analysis of the BFI Openness scale was carried out, resulting in four clusters of meaning: three items reflect artistic appeal ('values artistic, aesthetic experiences', 'has few artistic interests', and 'is sophisticated in art, music, or literature'), two items tap into the propensity to engage in cognitive activity ('is ingenious, a deep thinker', 'likes to reflect, play with ideas'), two items target inventiveness and creativity ('is original, comes up with new ideas', 'is inventive') and two items cover imagination and curiosity ('has an active imagination', 'is curious about many different things'). The item 'prefers work that is routine' was not further analysed because of its poor psychometric properties (Denissen et al., in press).

For every cluster, an adjective was sought that covered the corresponding domain as closely as possible while at the same time being coherent with the other three adjectives. In order to fit the bipolar format of the current rating instrument, two of these adjectives needed to be keyed towards low openness. An additional criterion was that ratings were to

Table 2. Means (M), standard deviations (SD), retest reliabilities (r(w4, w5)), internal consistencies and convergent and divergent validities of BFI scales, single-item self-ratings and single-item peer-rating composites

				BFI					TIPI-r self-ratings						TIPI-r peer composites					
	M	SD	*r*(w4, w5)	E	A	C	N	O	E	A	C	N	O′	O	E	A	C	N	O′	O
BFI																				
-E	3.45	.57	—	.83	.13	.06	−.26**	.11	**.68****	.11	.14*	−.21**	−.07	.31**	**.63****	.03	−.05	.11	.04	.41**
-A	3.58	.50	—		.76	.02	−.16*	.08	.12	**.59****	.19**	.00	.14	.13	.08	**.44****	.15*	.02	.16*	.12
-C	3.22	.64	—			.86	.08	−.02	.07	.11	**.66****	.09	.02	−.01	−.11	.11	**.50****	.19**	.22**	−.11
-N	3.02	.70	—				.88	−.20**	−.15*	−.13	.10	**.70****	−.01	−.16*	−.06	−.06	.11	**.37****	.00	−.14
-O	3.53	.57	—					.82	.11	.04	−.02	−.11	**.68****	**.28****	.07	−.04	.05	−.16*	**.38****	**.16***
TIPI-r self-ratings																				
-E	4.80	1.29	.75						.56	.20**	.26**	−.16*	.03	.35**	**.56****	.09	.02	.10	.09	.41**
-A	5.13	1.02	.58							.50	.30**	−.09	.11	.31**	.01	**.27****	.13	−.02	.14*	.07
-C	4.74	1.36	.71								.52	.06	.07	.16*	.02	.21**	**.54****	.15*	.29**	.04
-N	3.64	1.43	.73									.61	.09	−.22**	−.05	.03	.13	**.40****	.03	−.10
-O’	4.62	1.31	.70										.59	.25**	−.01	.05	.12	−.04	**.39****	.10
-O	5.10	1.19	.58											.10	.21**	.03	.00	−.14*	.18**	**.21****
TIPI-r peer composites																				
-E	4.37	.82	.96												.78	.08	−.04	.12	.16*	.72**
-A	4.70	.45	.85													.91	.33**	−.14*	.31**	.27**
-C	4.65	.74	.91														.86	.21**	.49**	.05
-N	3.63	.56	.89															.88	.01	−.14*
-O′	4.09	.52	.87																.82	.45**
-O	4.59	.51	.83																	.91

Note: E, Extraversion; A, Agreeableness; C, Conscientiousness; N, Neuroticism; O′, new Openness; O, old Openness; BFI, Big Five Inventory; reliability estimates (BFI: Cronbach’s alpha, TIPI-r self-ratings: squared factor loadings, TIPI-r peer composites: HLM intercept reliabilities) are displayed on the diagonals; convergent validities are displayed in bold; except for the retest stability, all estimates are based on data from Wave 5.

$^{*}p < .05$; $^{**}p < .01$.

be based on behaviours that are both observable and relatively consistent across situations (e.g. ruling out descriptors of 'curiosity', which is too dependent on the level of interest in the topics that are discussed during study groups; Renninger, Ewen, & Lasher, 2002). With these criteria, 'interested in art' was chosen as a marker of the artistic openness facet, reflecting an interest in art and culture and 'deep' as a marker for the propensity to engage in deep thinking. On the low openness side, 'conventional' (antonyms: 'new', 'original', 'Roget's New Millennium Thesaurus', 2006) was selected as a marker for low creativity, and 'pragmatic' ('matter-of-fact', 'sober', 'Roget's New Millennium Thesaurus', 2006) as a marker for low imaginativeness ('pragmatic' was also selected as a low-openness marker by Woods & Hampson, 2005).

Participants rated both themselves (TIPI-r self-ratings) and their peers (TIPI-r peer composites) on each of these bipolar items, using a 1 (extremely like the left adjective pair) to 7 (extremely like the right adjective pair) scale. Following Woods and Hampson (2005), we varied the location of the socially desirable pole, with Extraversion, Openness and Conscientiousness items having the desirable pole on the left side, and Neuroticism and Agreeableness having this pole on the right side.

Procedure

For four consecutive months starting the second week of their university freshman year, participants filled out five waves of online questionnaires by accessing the website of the study using a personal password. In Wave 1, participants completed the BFI as well as the single-item ratings. In monthly intervals, they completed Waves 2, 3 and 4, which consisted only of the single-item ratings. Again after 1 month, in Wave 5, 4 months after the beginning of the study, participants completed the BFI in addition to the single-item ratings. For two reasons, the current analyses are based on Waves 4–5. First, as the participants were previously unacquainted, they likely had only a limited empirical basis to provide accurate peer-ratings of the FFM factors during the early waves. Second, the modified Openness item was only added in Wave 4.

In Wave 4, participants filled out a short questionnaire that took a median of 19 minutes to complete. The TIPI-r was part of a larger battery. The ratings were presented in randomized order to avoid response sets. For every Big Five dimension, a list of all group members (including the participants themselves) was presented on a separate page, alongside the rating scales (see Appendix A). In Wave 5, 1 month later, participants first completed self-ratings using the BFI, followed by self- and peer-ratings using the 5 SI ratings, which were presented in randomized order. Both the BFI and the TIPI-r ratings were part of a larger battery that took a median of 40 minutes to complete.

RESULTS

Descriptive statistics

Table 3 shows the distribution characteristics of the five TIPI-r items used to rate the personalities of all members of participants' social networks in Wave 5. For comparative purposed, we also report results for the old Openness item. All items were slightly tilted towards the socially desirable side of the scale (the theoretical mean being 4), though this

Table 3. Single-item statistics of TIPI-r scales assessing social network members (Wave 5)

	E	A	C	N	O′	O
Mean	4.39	4.71	4.66	3.63	4.11	4.61
SD	1.37	1.18	1.33	1.26	1.18	1.27
Skewness	−.27	−.19	−.30	.06	−.05	−.30
Kurtosis	−.39	−.19	−.29	−.20	−.11	−.10

Note: E, Extraversion; A, Agreeableness; C, Conscientiousness; N, Neuroticism; O′, new Openness; O, old Openness.

did not result in worrisome skewness or kurtosis values. Standard deviations ranged between 1.18 for Agreeableness/new Openness and 1.37 for Extraversion.

Reliability

Proportion of true score variance ('internal consistency')

Following the procedure recommended by Woods and Hampson (2005), the five SIs were inserted together with the 44 items of the BFI in a joint principal component analysis with a forced five-factor solution, explaining 50% of the variance. Factor loadings of the SIs were .75, .78, .77, .72 and .71, and .32 for Extraversion, Neuroticism, new Openness, Conscientiousness, Agreeableness and old Openness, respectively (for purposes of comparison: Denissen et al., in press, reported an average loading of .62 for the items of the Dutch BFI). By squaring these loadings, reliability estimates were calculated. As can be seen in the corresponding diagonal in Table 2 (block TIPI-r self-ratings), they range between .10 for old Openness and .61 for Neuroticism ($M = .48$). These estimates are similar to the communalities of these items ($M = .54$), except for old Openness, for which an inflated communality (.27) was found due to this item's substantial secondary loadings. In total, the reliability estimates of the single-item TIPI-r self-ratings are relatively low, especially when compared to the internal consistency (Cronbach's alpha) of the multi-item BFI, which ranged from .76 to .88.

One-month test–retest stability

The fourth column of Table 2 shows the 1-month test–etest stabilities. Retest correlations were moderate to high for the TIPI-r self-ratings, with correlations ranging between .58 for Agreeableness/old Openness and .75 for Extraversion ($\bar{r} = .68$ after Fisher r-to-Z transformation and back-transformation, as in all further instances). Very high retest correlations were obtained for the TIPI-r peer composites, with values ranging between .83 for old Openness and .96 for Extraversion ($\bar{r} = .90$), demonstrating excellent test–retest stability.

Reliability of peer-rating composite

To examine the reliability of the peer-ratings, a two-level multilevel model was specified, with peer-ratings (Level 1) being nested within participants (Level 2). Three-level models with a higher-order group level led to identical results. Intercept-only models were estimated for all Big Five factors. As two participants failed to complete the peer-ratings, and 1 participant failed to produce any variation in her ratings, there were 203 raters offering 4341 peer-ratings in Wave 5 (results for Wave 4 were virtually identical and will not be reported here). The corresponding diagonal of the TIPI-r peer composite ratings in

DOI: 10.1002/per

Table 2 shows the corresponding reliability estimates. The composite peer-ratings were highly reliable, with an average value of .87 (range .78–.91).

Given that the average number of peer raters for each group member was 21.4, the Spearman–Brown formula[5] estimated (on the basis of the HLM reliability coefficients) that the average agreement between pairs of raters would be .19, with values of .14, .32, .22, .26, .32 and .18, for Extraversion, Agreeableness, Conscientiousness, Neuroticism, old Openness and new Openness, respectively. To corroborate these estimates, the data matrix was restructured so that every line represented a participant and the columns represented the corresponding peer-ratings for that target individual. The average agreement level between all possible rater pairs was almost identical to the estimates obtained with multilevel analysis (.16, .33, .24, .26, .35 and .18, respectively).

Validity

Discriminant scale intercorrelations

The multitrait-multimethod matrix in Table 2 shows that the correlations between the different BFI scales were generally small, with levels ranging between $-.26$ and .13 ($\bar{r}$ after reverse-coding Neuroticism $=.09$). For the TIPI-r self-ratings, correlations $>.30$ were found between old Openness and both Extraversion ($r=.35$) and Agreeableness ($r=.31$), with an average value (after reverse-coding Neuroticism and excluding the correlation between old and new Openness) of .15. The average discriminant validity of the TIPI-r peer composites (after reverse-coding Neuroticism and excluding the correlation between old and new Openness) was .18. We observed substantial correlations of .72 between Extraversion and old Openness, .49 between new Openness and Conscientiousness, .31 between new Openness and Agreeableness and .33 between Agreeableness and Conscientiousness. Corrected for unreliability (the square root of the product of the reliabilities of the corresponding scales), the average discriminant validity of the TIPI-r self-ratings is .31 (.15/.48) against a value of .21 (.18/.87) for the TIPI-r peer composites composite. For the BFI scales, a corresponding value of .11 (.09/.83) was found. Accordingly, halo bias seems a somewhat greater problem for SIs, especially when used for self-ratings.

Convergent and discriminant correlations between single items and BFI scales

Table 2 presents correlations between the TIPI-r self-ratings, TIPI-r peer composites and BFI scales. Convergent correlations between the TIPI-r self-ratings and the BFI scales were all significant and moderate to high in size ($\bar{r}=.61$), except in the case of old Openness,

[5] The Spearman–Brown formula estimates the reliability of a new scale (r_{kk}) that is created by multiplying the known reliability (r_{11}) by a factor k, which represents that the test is made k times as long:
Equation (5):

$$r_{kk} = \frac{kr_{11}}{1+(k-1)r_{11}}$$

because indices of internal consistency can be regarded as the correlation of a measure with itself, the agreement between two peers can be regarded as the internal consistency of a single peer rater. Following this logic, the average peer–peer agreement (r_{11}) can be calculated with the following formula:
Equation (6):

$$r_{11} = \frac{r_{kk}}{k-kr_{kk}+r_{kk}}$$

where n is the average number of peer raters, and r_{kk} is the reliability of the peer composite.

which correlated only .28 with the corresponding BFI scale. Corrected for unreliability, convergent validity coefficients approached unity in all cases (range .96–1.00). Discriminant validities ranged between −.21 and .31 and averaged (after reverse-coding Neuroticism) .08.

Turning to the convergence between BFI scales and the TIPI-r peer aggregates, significant correlations were found for all FFM factors, with values ranging between .16 for old Openness and .63 for Extraversion ($\bar{r} = .43$). Correlations corrected for unreliability ranged between .19 for old Openness and .78 for Extraversion ($\bar{r} = .52$). Discriminant validities ranged between −.16 and .41 ($\bar{r} = .07$). The off-diagonal correlation of .41 between old Openness and BFI Extraversion exceeded the correlation between the old Openness item and the BFI Openness.

Self-peer agreement

The agreement between self-ratings and aggregated peer-ratings was significant, with correlations ranging between .21 for old Openness and .56 for Extraversion ($\bar{r} = .40$). The off-diagonal correlation between self-rated Extraversion and peer-rated old Openness ($r = .41$) was almost double the value of the convergent correlation for the old Openness item ($r = .21$).

Degree of projection

Finally, we calculated the correlation between participants' single-item ratings of themselves and their peers. If this correlation is high, participants' peer-ratings on average do not deviate much from their self-ratings, which can be treated as an index of projection given the fact that there was random assignment to groups and thus no objective personality similarity. The degree of projection was .30, .33, .40, .43, .37 and .52 for Extraversion, Agreeableness, Conscientiousness, Neuroticism, new Openness and old Openness, respectively.

Summary

Table 4 summarizes the different reliability and validity results, which differed somewhat between our scales. Only Neuroticism received at least acceptable psychometric support across all criteria. Neuroticism had minimally acceptable communalities and projection

Table 4. Summary of reliability and validity results for single-item Big Five Indicators

	Reliability				Validity		
	True score variance/s	Retest/s	Retest/p	Consistency/p	Convergence BFI-/s	Agreement /p-/s	Low projection
Extraversion	−	+	+ +	+	+/−	+ +	+
Agreeableness	−	−	+ +	+ +	−	+/−	+/−
Conscientiousness	−	+	+ +	+ +	+/−	+ +	+/−
Neuroticism	+/−	+	+ +	+ +	+	+	+/−
New Openness	−	+	+ +	+ +	+/−	+	+/−
Old Openness	− −	−	+ +	+ +	− −	+/−	−

Note: /p, peer composite, /s, self, SI, single item, BFI, Big Five Inventory, + +, very good, +, respectable, +/−, minimally acceptable, −, unacceptable, − −, very poor; for the evaluation of the reliability and convergent validity estimates, the following criteria by Nunnally and Bernstein (1994) were used: + + $\alpha \geq .80$, + $.70 \leq \alpha < .80$, (+) $.65 \leq \alpha < .70$, +/− $.60 \leq \alpha < .65$, −$.40 \leq \alpha < .60$, − − $\alpha < .40$. For the evaluation of the agreement

DOI: 10.1002/per

coefficients, respectable convergent and agreement validity and test–retest reliability of single-item self-ratings and very good internal consistencies and test–retest reliability of aggregate peer composites. Extraversion and Conscientiousness, and new Openness had unacceptable communalities, but otherwise their psychometric properties were acceptable and even surpassed those of Neuroticism in terms of self-peer agreement. Finally, Agreeableness had unacceptable values in terms of the communality, test–retest reliability and convergent validity of the single-item self-ratings, though it fared quite well in terms of the psychometric properties of the peer-rating aggregates. The old Openness item had very poor psychometric properties in terms of the proportion of true score variance (‘internal consistency’) and convergent validity of the single-item self-rating and a high level of projection, which was only compensated by the high internal consistency and test–retest reliability of the aggregate peer-ratings.

DISCUSSION

The current study evaluated the use of the TIPI-r in a social network design. The basic psychometric properties of the TIPI were not distorted by the translation into Dutch, the construction of bipolar single-item format and the use within a social network design. At ‘traditional’ psychometric criteria such as proportion of true score variance (‘internal consistency’), retest reliability and convergent and discriminant validity, the TIPI-r displayed a similar level of performance as other ultra-short Big Five measures (see Table 1). In the social network design, the TIPI-r fared relatively well with respect to consistency and stability of people’s trait reputations and degree of projection that raters used to infer their peers’ trait levels.

In agreement with another study (Muck et al., in press), our study showed that the original TIPI Openness item performs worse than the other TIPI scales. Three modifications to the original TIPI can explain the relatively poor psychometric performance of the old Openness item. First, inadequate translation of the English item into Dutch is an unlikely explanation, because our openness translation was virtually identical to the German translation (Muck et al., in press). Second, it is also not likely that the use of a social network design negatively affected the old Openness item, because the other items did relatively well. Third, and most likely, the poor performance of the old Openness item stems from the construction of a single bipolar scale from two opposite items. This requires that both poles are semantic and psychological opposites. The two Openness items of the TIPI may have been too heterogeneous. For example, individuals can be conventional in politics but creative in arts. This is consistent with Gosling et al.’s (2003) aim while developing the TIPI to maximize content validity by covering different facets of the FFM factors instead of maximizing internal consistency. Moreover, participants seem to have reacted primarily to the ‘open to experience’ part of the old Openness item, which is reminiscent of the sensation-seeking facet of Extraversion (e.g. ‘excitement-seeking’ is a facet of the NEO-PI-R Extraversion scale; Costa & McCrae, 1992) and would explain the relatively high correlation (especially in terms of trait reputations) between the old Openness and Extraversion SI ratings in our study. In terms of psychometric properties, the new Openness performed better than the old Openness item and is recommend if researchers want to assess this factor with a bipolar single-item self-rating scale.

DOI: 10.1002/per

In terms of reliability, differences were found between single-items' ability to tap individual differences in self-rated personality and their ability to capture people's trait reputations. For example, internal consistency and temporal stability reliabilities of the old Openness and Agreeableness items suggest that these items could better be used to form peer composites than for self-ratings. When the adjectives of a scale, like the Old openness scale, are not psychological opposites, the scale will be most applicable to individuals who have high values on one pole and low values on the opposite one, while individuals who have high or low values on *both* poles will likely obtain intermediate values from raters on the scale. If raters are consistent in this tendency, poor psychometric properties on the level of individual self-ratings can give rise to spuriously superior properties on the level of peer-rating composites.

In the current study, the composite of all peer-ratings reliably assessed *trait reputations*, which are crucial to understanding the very nature of personality assessment according to some authors (e.g. Hogan, 1996). Indeed, according to Hofstee (1994, p. 149), 'the averaged judgment of knowledgeable others provides the best available point of reference both for the definition of personality structure in general and for assessing someone's personality in particular'. As peer-ratings are less subjected to self-serving response biases, they may form an attractive alternative to self-rating instruments, especially if they are assessed with ultra-short measures such as the TIPI or TIPI-r. This does not have to entail a very large rater pool as the one used in the current study: Applying the Spearman–Brown formula predicts that reliable (alpha = .70) peer-rating composites can be achieved with between 5 and 14 raters for Neuroticism, Openness, Conscientiousness and Agreeableness. One could use surplus rater capacity in a planned missing data design in which groups of raters assess different items (e.g. if eight raters assess the assertiveness facet of extraversion; the next eight raters assess the gregariousness facet; etc.). The composite of these peer-ratings would then provide a more multifaceted informant report of personality.[6]

In terms of the validity of the single-items, results diverged across criteria. For example, agreement between self and peers was at least minimally acceptable for all scales, but convergence between single-item and BFI self-ratings was unacceptable for Agreeableness[7], and very poor for old Openness. We tested the items' susceptibility for projection as an additional aspect of content validity that can only be assessed within a social network design. The results for this validity test were consistent with the agreement between self and peer-ratings as an estimate of convergent validity, which are deemed so crucial in assessing the validity of SIs. In both cases, validity estimates were satisfactory for Extraversion, and minimally acceptable for Neuroticism, Openness, Conscientiousness and Agreeableness and unacceptable for old Openness. This highly similar pattern bolsters our faith in the usefulness of projection as an index of validity. However, there may be circumstances in which applying the convergent validity and projection criteria leads to different results. For example, SIs may tap into well-observable, highly specific traits (e.g. smiling a lot) that do not converge with the FFM factor they are supposed to assess (e.g. agreeableness).

Besides allowing the calculation of new reliability (consistency of peer composite) and validity (projection) indicators, social network designs bring about several new advantages and limitations to the assessment of personality. Time is the first obvious limitation, as it

[6] We thank an anonymous reviewer for this suggestion.

[7] The relatively low reliability of the BFI Agreeableness scale (see Table 2) likely attenuated the convergent correlation with the single-item Agreeableness scale.

DOI: 10.1002/per

would simply be unfeasible to use traditional multi-item scales to assess personality traits in social network designs. Simple bipolar indicators limit the time needed for participants to rate every network partner. We did not track response times on an item-to-item basis, but observed that in Wave 4 participants needed 1.42 minutes on average for every rating dimension (50% of dimensions being FFM indicators), whereas in Wave 5 this was 1.53 minutes (32% of dimensions being FFM indicators). Accordingly, 7–8 minutes are needed when using five SIs to assess the FFM factors of all members of a social network with between 23 and 25 members. Given that the number of peer raters needed to form reliable peer-rating composites using ultra-short FFM measures is relatively modest, it is feasible to use single-items in round-robin designs to assess individual differences in peer reputations.

Social network designs also place limitations on the complexity of items. SI scales cannot replace existing bipolar FFM measures that consist of eight or more adjectives per factor, but that are difficult to implement in designs with multiple ratings. More complex single-item measures, such as the one by Rammstedt et al. (2004) taking 4 minutes to complete would have been not efficient in our design. We used the items of the simple and well-validated unipolar measure, the TIPI (Gosling et al., 2003) for the construction of bipolar single-item indicators of the Big Five. Given that the average time to complete the entire rating procedure was around 21 seconds per target (7.4 minutes/21.4 network partners), we remained well within the complexity limits imposed by the social network design.

Some limitations of our study are worth mentioning. One obvious limitation is the reliance on psychology undergraduates. However, we succeeded in voluntarily enlisting almost all members of the introductory groups for our study, so that we were able to draw a representative sample from this subpopulation. Research with different types of social networks will assure the generalizability of the psychometric properties of the SIs. Second, we included the BFI as a validation measure of the FFM, neglecting alternative measures that differ in content and scope. Research with other instruments is needed to expand the evidence regarding the convergent validity of the current single-item FFM measure. Third, although the current paper is the first to develop and validate a single-item instrument in a language other than English and German, more research is needed in non-Germanic, non-Western languages. For such an enterprise, it may be helpful to base future SIs on the kind of universal processes that form the conceptual core behind each of the FFM factors (Denissen and Penke, in preparation) instead of on linguistic markers that were derived from Western lexicons.

CONCLUSION

The psychometric properties of the TIPI-r are generally comparable to results found by previous investigations of ultra-short Big Five measures. The results differ somewhat between the specific psychometric criteria that are applied (see Table 4). Accordingly, whether the use of single-items is advisable will vary according to the nature of the research question at hand. SIs cannot provide the faceted picture of an individual person, for which the longer FFM questionnaires are needed. The relatively low proportion of true score variance ('internal consistency') of the self-ratings gives rise to serious caution, but the high reliability of the peer-rating composites suggests usefulness in future studies. A main asset of the current paper is that it demonstrates that employing bipolar single-item FFM

DOI: 10.1002/per

indicators in a social network design does not result in a deterioration of their psychometric properties. Such a design allows novel ways to calculate reliability and validity, and can be used to assess individual differences in trait reputations, which have been described as crucial in understanding and conceptualizing the FFM of personality.

REFERENCES

Campbell, D., & Fiske, D. (1959). Convergent and discriminant validation by the multitrait-multimethod matrix. *Psychological Bulletin*, *54*, 81–105.

Charter, R. A. (2003). A breakdown of reliability coefficients by test type and reliability method, and the clinical implications of low reliability. *The Journal of General Psychology*, *130*, 290–304.

Cohen, J. (1988). *Statistical power analysis for the behavioral sciences*. Hillsdale: Erlbaum.

Costa, P. T., & McCrae, R. R. (1992). *Revised NEO Personality Inventory (NEO PI-R) and NEO Five-Factor Inventory (NEO-FFI): Professional manual*. Odessa: Psychological Assessment Resources.

Costa, P. T., & McCrae, R. R. (1995). Solid ground in the wetlands of personality: A reply to Block. *Psychological Bulletin*, *117*, 216–220.

Denissen, J. J. A., Geenen, R., van Aken, M. A. G., Gosling, S. D., & Potter, J. (in press). Development and validation of a Dutch translation of the Big Five Inventory (BFI).

Denissen, J. J. A., & Penke, L. (under revision). Individual reaction norms underlying the Five Factor Model of Personality: First steps towards a theory-based conceptual framework.

Funder, D. C. (1995). On the accuracy of personality judgment: A realistic approach. *Psychological Review*, *102*, 652–670.

Funder, D. C., & Dobroth, K. M. (1987). Differences between traits: Properties associated with interjudge agreement. *Journal of Personality and Social Psychology*, *52*, 409–418.

Funder, D. C., Kolar, D. C., & Blackman, M. C. (1995). Agreement among judges of personality: Interpersonal relations, similarity, and acquaintanceship. *Journal of Personality and Social Psychology*, *69*, 656–672.

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. (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 Personality Assessment*, *22*, 139–148.

Hofstee, W. K. B. (1994). Who should own the definition of personality? *European Journal of Personality*, *8*, 149–162.

Hogan, R. (1996). A socioanalytic perspective on the five-factor model. In J. S. Wiggins (Ed.), *The five factor model of personality: Theoretical perspectives* (pp. 163–179). New York: Guilford Press.

John, O. P., & Robins, R. W. (1993). Determinants of interjudge agreement on personality traits: The big-five domains, observability, evaluativeness, and the unique perspective on the self. *Journal of Personality*, *61*, 521–551.

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: Theory and Research* (pp. 102–138). New York: Guilford Press.

Kenny, D. A. (1994). *Interpersonal perception: A social relations analysis*. New York: Guilford.

Kenny, D. A. (1996). The design and analysis of social-interaction research. *Annual Review of Psychology*, *47*, 59–86.

Kenny, D. A. (2004). PERSON: A general model of interpersonal perception. *Personality and Social Psychology Review*, *8*, 265–280.

Kruger, J., & Clement, R. W. (1994). The truly false consensus effect: An ineradicable and egocentric bias in social perception. *Journal of Personality and Social Psychology*, *67*, 596–610.

Miller, G. A. (1956). The magical number seven, plus or minus two: Some limits on our capacity for processing information. *Psychological Review*, *63*, 81–97.

DOI: 10.1002/per

Muck, P. M., Hell, B., & Gosling, S. D. (in press). 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 Personality Assessment.*

Neyer, F. J. (1997). Free recall or recognition in collecting egocentered networks: The role of survey techniques. *Journal of Social and Personal Relationships*, *14*, 305–316.

Nunnally, J. C., & Bernstein, I. H. (1994). *Psychometric theory.* New York: McGraw-Hill.

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. *ZUMA Nachrichten*, *55*, 5–28.

Renninger, K. A., Ewen, L., & Lasher, A. K. (2002). Individual interest as context in expository text and mathematical word problems. *Learning & Instruction*, *12*, 467–491.

Roget's New Millennium Thesaurus. (2006). Retrieved December 06, 2006, from http://thesaurus.reference.com/browse/conventional

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

Swann, W. B. (1987). Identity negotiation: Where two roads meet. *Journal of Personality and Social Psychology*, *53*, 1038–1051.

van Duijn, M. A. J., van Busschbach, J. T., & Snijders, T. A. B. (1999). Multilevel analysis of personal networks as dependent variables. *Social Networks*, *21*, 187–209.

Warner, R. M., Kenny, D. A., & Stoto, M. (1979). A new round robin analysis of variance for social interaction data. *Journal of Personality and Social Psychology*, *37*, 1742–1757.

Wasserman, S., & Faust, K. (1994). *Social network analysis*. Cambridge: Cambridge University Press.

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.

APPENDIX A

Example of ratings in a social network design

Instruction: Some people are extraverted and enthusiastic, whereas other people are reserved and quiet. In the following, please rate yourself and your group members with regard to this dimension.

		Extraverted and enthusiastic						Reserved and quiet
Nr.	Name	1	2	3	4	5	6	7
1	John		X					
2	Anne			X				
3	Peter			X				
4	Mary		X					
5	Jack				X			
6	Rose					X		
7	Bill						X	
8	Cindy		X					
9	Alex			X				
...								
25	Joan				X			

DOI: 10.1002/per