Behav Res (2011) 43:548–567
DOI 10.3758/s13428-011-0066-z

Short assessment of the Big Five: robust across survey methods except telephone interviewing

Frieder R. Lang · Dennis John · Oliver Lüdtke · Jürgen Schupp · Gert G. Wagner

Published online: 18 March 2011
© The Author(s) 2011. This article is published with open access at Springerlink.com

Abstract We examined measurement invariance and age-related robustness of a short 15-item Big Five Inventory (BFI–S) of personality dimensions, which is well suited for applications in large-scale multidisciplinary surveys. The BFI–S was assessed in three different interviewing conditions: computer-assisted or paper-assisted face-to-face interviewing, computer-assisted telephone interviewing, and a self-administered questionnaire. Randomized probability samples from a large-scale German panel survey and a related probability telephone study were used in order to test method effects on self-report measures of personality characteristics across early, middle, and late adulthood. Exploratory structural equation modeling was used in order to test for measurement invariance of the five-factor model of personality trait domains across different assessment methods. For the short inventory, findings suggest strong robustness of self-report measures of personality dimensions among young and middle-aged adults. In old age, telephone interviewing was associated with greater distortions in reliable personality assessment. It is concluded that the greater mental workload of telephone interviewing limits the reliability of self-report personality assessment. Face-to-face surveys and self-administrated questionnaire completion are clearly better suited than phone surveys when personality traits in age-heterogeneous samples are assessed.

Keywords Big Five · Personality · Survey method · Face-to-face · Telephone interview · Self-administrated questionnaire · Exploratory structural equation modeling (ESEM)

F. R. Lang (✉) · D. John
Institute of Psychogerontology,
University of Erlangen-Nuremberg,
Nägelsbachstr. 25,
91052 Erlangen, Germany
e-mail: flang@geronto.uni-erlangen.de
URL: www.geronto.uni-erlangen.de

O. Lüdtke
University of Tübingen,
Tübingen, Germany

J. Schupp · G. G. Wagner
German Institute of Economic Research,
Berlin, Germany

J. Schupp
Free University Berlin,
Berlin, Germany

G. G. Wagner
Berlin University of Technology,
Berlin, Germany

G. G. Wagner
Max-Planck-Institute for Human Development Berlin,
Berlin, Germany

In recent years, a broad range of psychological constructs such as the Big Five personality traits have had strong repercussions in the social and economic sciences, which has led to an increasing use of psychological self-report instruments in large-scale panels such as the German Socio-Economic Panel (SOEP; Siedler, Schupp, & Wagner, 2011; Wagner, Frick, & Schupp, 2007), the British Household Panel Study (BHPS; Taylor, Brice, Buck, & Prentice-Lane, 2009), or the Household, Income, and Labour Dynamics in Australia survey (HILDA; Lucas & Donnellan, 2009). The

wealth of new insights and seminal findings across disciplines that can be gained by including short self-report personality measures in large-scale surveys depends on the extent to which such measures prove robust and reliable within age-heterogeneous samples and across different assessment techniques. Whereas the differences in the direct costs of different survey techniques are obvious, the indirect costs with respect to measurement artifacts are not obvious at all. Face-to-face techniques, such as paper-and-pencil personal interviewing (PAPI), computer-assisted personal interviewing (CAPI), and interviewer-supported self-administration (SELF), are much more expensive than computer-assisted telephone interviewing (CATI). The costs of different interviewing techniques with regard to robustness of self-report personality constructs (i.e., measurement invariance and retest reliability) are not yet well understood. This is of particular relevance when personality traits are measured in representative samples across a broad range of ages encompassing the life span.

For example, telephone interviewing typically requires that respondents comprehend a verbally presented item while reflecting on the correct self-related verbal response on an imaginary rating scale (Bauer, Truxillo, Paronto, Weekley, & Campion, 2004). In this case, the costs of cognitive processing are associated with limitations of working memory. Similarly, in situ personal interviewing may depend on the communication skills of the interviewer. It is an open question whether method effects arise in the assessment of self-reported personality constructs. To the best of our knowledge, we do not know of a study that has explored potential method effects of standard interviewing techniques on assessment of the Big Five personality constructs. The present research investigated the robustness of assessing the Big Five personality constructs with three different interviewing methods (face-to-face interviewing, telephone interviewing, and self-administered questionnaire completion) in two age-heterogeneous, population-representative German samples of adults from early to late adulthood. Specifically, we expected that the measurement demands of the telephone interviewing method would be highest among older adults, leading to less reliable personality assessment.

This implies that the mental workload associated with different assessment procedures may also lead to flaws in studies on aging personality—for example, in research on age differences in personality (Donnellan & Lucas, 2008). In the present research, we investigated the measurement invariance and test–retest reliability of a widespread short 15-item personality inventory (Gerlitz & Schupp, 2005; Lang, 2005) across three assessment conditions: self-administered completion of a questionnaire, computer-assisted personal interviewing, and telephone interviewing. More precisely, we explored the extent to which the challenges of these differential methods would prove *robust* with respect to age-differential comparisons across young, middle, and late adulthood. Here, we refer to *assessment robustness* as an indication of different levels of factorial invariance and absence of method-related differences in test–retest stability. Furthermore, we are fully aware that measurement equivalence of personality assessment across adulthood is a different research issue addressed elsewhere (e.g., Lang, Lüdtke, & Asendorpf, 2001; Marsh, Nagengast, & Morin, 2011). Rather, the present study targets the differential effects of assessment conditions within and across age groups. In detail, we expected that the cognitive demands of personal interviewing situations would involve challenges for the robustness of personality assessment among older adults, as compared with the use of self-administered questionnaires. While the latter represents the standard in most psychological assessment contexts, large-scale surveys with probability samples (Groves et al., 2009), such as national General Social Surveys (e.g., NORC GSS, Davis & Smith, 1992; or the European Social Survey, Stoop, Billiet, Koch, & Fitzgerald, 2010), have typically relied on personal interviewing methods—typically, in face-to-face contexts or via the telephone.

In sum, we addressed two major research questions. First, how robust is the identification of the Big Five structure of personality trait dimensions when assessed via telephone (CATI), self-administrated completion of questionnaires (SELF), and CAPI? Second, are older adults, due to the higher cognitive constraints of telephone interviews, more susceptible to effects of CATI methods, as compared with young and middle-aged adults?

Assessing the Big Five personality dimensions

The Big Five personality constructs represent a powerful frame of reference in psychological reasoning about the structure of interindividual differences in personality dimensions (John & Srivastava, 1999; McCrae & Costa, 1997). A reason for the enormous vigor, spread, and acceptance of the five-factor model in personality research may be related to the notable convergence of research findings that have emerged from theoretical backgrounds in the tradition of eminent scholars such as Hans Eysenck (e.g., 1947) and Ludwig Klages (1926) or from psycholexical theory in the tradition of Gordon Allport (Allport & Odbert, 1936), Raymond B. Cattel (1946), and Lewis R. Goldberg (1990).

Exploratory factor analysis has lent wings to the confluence of these distinct endeavors in identifying the fundamental dimensions of consistent personality differences. In 1985, Costa and McCrae proposed that individual

differences in personality characteristics generally reflect five broad trait dimensions: neuroticism (N), extraversion (E), openness to experience (O), agreeableness (A), and conscientiousness (C). As a consequence, this assumed universal structure of the Big Five domains of individual differences in personality dimensions was reconfirmed in numerous empirical studies from different cultures (for reviews, see John & Srivastava, 1999; McCrae & Costa, 1997). To date, the Big Five personality constructs represent a widely accepted, comprehensive, and ample frame for delineating the structure of core personality traits across adulthood.

As a consequence, there also exist several widely used self-report instruments that allow for reliable and content-valid assessment of the Big Five personality constructs (e.g., Costa & McCrae, 1995; John & Srivastava, 1999). Not surprisingly, the wealth of findings that result from such reliable self- or other-report assessments of the Big Five personality constructs has instigated much research in the neighboring disciplines of psychology—for example, with respect to economics (Borghans, Duckworth, Heckman, & ter Weel, 2008; Heineck & Anger, 2010), education (Marsh, Trautwein, Lüdtke, Köller, & Baumert, 2006; Swanberg & Martinsen, 2010), health (Smith & Williams, 1992), and social resources (Headey, Muffels, & Wagner, 2010). Reliable and robust assessment of the Big Five personality constructs thus promotes strong and insightful cross-disciplinary transfers and exchanges between the behavioral, economic, and social disciplines that also generate and instigate new research questions. For example, one critical issue is whether individual differences in personality domains contribute to an improved understanding of processes that produce or reproduce social inequality in modern societies (e.g., Borghans et al., 2008; Heineck & Anger, 2010; Sutin, Costa, Miech, & Eaton, 2009). Such findings have raised strong demands to include self-report measures of the Big Five personality trait dimension in large-scale national surveys and household panels. As a consequence, there is a growing need for efficient and short measures of the Big Five personality constructs that fit well with the enormous constraints of the survey context (Donnellan, Oswald, Baird, & Lucas, 2006; Gosling, Rentfrow, & Swann, 2003). Such brief measures allow for a rough and valid assessment of the Big Five personality domains. Potential costs of short personality measures are that the breadth and facets of the Big Five constructs may not be fully represented. Moreover, ultra-short measures of the Big Five constructs with two or only one item each (Rammstedt & John, 2007) may not be suitable for latent factor modeling. Typically, three items per construct represent a minimum for identification of the five-factor structure of personality trait domains (Gagné & Hancock, 2006; Marsh, Hau, Balla, & Grayson, 1998).

Methods for assessing self-report measures in survey research: face-to-face, telephone, and self-administration

Most surveys rely on personal-interviewing methods such as paper- or computer-assisted face-to-face interviews (FACE), telephone interviewing (CATI), or interviewer-assisted self-administrated completion of questionnaires (SELF). In face-to-face or telephone assessments, interviewers typically read questions and then record the respondents' answers on paper or on a computer. As compared with self-administrated questionnaires, such procedures require more time, thus reducing the number of items that can be included.

Face-to-face personal interviewing

In face-to-face procedures (FACE), interviewers typically read questions and answer formats aloud to respondents and then note or mark responses either on a paper questionnaire (paper-assisted personal interviewing; PAPI) or with a special CAPI software program on a portable computer. Differences between PAPI and CAPI procedures are marginal with respect to self-report personality measures. CAPI software procedures typically entail plausibility checks and automatic response-dependent selection of questions in the course of the interview. One critical common feature of face-to-face personal interviewing methods is that respondents report their self-descriptive ratings aloud to the interviewer, which may instigate reactivity or social desirability (McHorney, Kosinski, & Ware, 1994; Richman, Kiesler, Weisband, & Drasgow, 1999).

Self-administered paper-and-pencil mail-in questionnaire

In the field of personality research, self-administrated completion of self-report questionnaires is a popular and widely used assessment method for psychological assessment (Holden & Troister, 2009). Different self-administration procedures in personality assessment based on either paper-and-pencil or computerized self-report questionnaires have been shown to generate equivalent means and covariances (Rammstedt, Holzinger, & Rammsayer, 2004). Thus, self-administration procedures represent a valuation standard of personality assessment methods. This is not to say that self-administrated questionnaire completion does not entail risks of response biases. For example, respondents may overlook items or may not read statements carefully enough. The advantage of self-administration is that the respondents do not have to disclose themselves to an interviewer and are

able to rate self-descriptive statements in a discrete context. Self-completion methods may also be used for group testing or for mail-in studies.

Computer-assisted telephone interviewing

CATI is a common and widely used method in the social and economic sciences (Groves & Mathiowetz, 1984). While there have been some explorations of possible method effects of telephone interviewing in the 1970s and 1980s (Hansell, Sparacino, Ronchi, & Strodtbeck, 1984; Herman, 1977), technology-assisted application of telephone interviewing methods have recently found more use in psychological research—for example, in studies with large national samples (e.g., Tun & Lachman, 2008) or in research with high-risk populations (Cercone, Danielson, Ruggiero, & Kilpatrick, 2009). One advantage of CATI procedures is related to time and cost efficiency. Typically, CATI involves only a short initial contact and no transportation; however, it allows for only a few questions (Bauer, et al., 2004). In general, interviewers read each single item aloud on the phone. together with response options. Participants then rate their agreement with each self-report statement. Interviewers document the responses with special CATI software on a computer. The cognitive task of such procedures involves, for example, that participants simultaneously represent the units of a rating scale (e.g., 7-point) while cognitively processing the degree of agreement with the respective statement. Naturally, this involves constraints of auditory information processing, which may be a particular challenge when reflecting responses that involve complex descriptions of social behaviors and personality characteristics.

Is there measurement invariance across assessment methods?

It is an open question whether self-administered questionnaire assessment, telephone interviewing, or face-to-face interviewing generate mode biases with respect to the consistency and reliability of personality constructs. In a meta-analysis comparing studies that used computerized and paper-and-pencil questionnaires, as well as face-to-face interviews, Richman et al., (1999; see also Bauer et al., 2004; Holbrook, Green, & Krosnick, 2003; Marshall, De Fruyt, Rolland, & Bagby, 2005) reported mean-level differences in personality measures related to social desirability. As was expected, self-administered questionnaire procedures were associated with generally less positive self-reports of personality characteristics, as compared with face-to-face interviewing procedures. Similarly, McHorney et al. (1994) conducted a study comparing the effects of a mail-in assessment versus telephone interviewing. In this study, telephone interviewing resulted in more favorable self-descriptions of health, as compared with the self-administered mailing procedure. Whereas an interviewing bias of mean levels of constructs may be corrected, the study did not demonstrate whether there were equivalent variances. More important, no study has compared the factorial structure of the involved personality measures across assessment methods. To the best of our knowledge, we do not know of a study that has demonstrated measurement equivalence of the Big Five personality dimensions for CAPI, telephone interviewing, and conventional self-administered paper-and-pencil assessment contexts. All these survey methods are widely applied in major large-scale surveys.

Equivalence of the factorial structure is particularly relevant whenever individual differences of personality trait domains are associated with other variables such as gender, education, or chronological age. For example, several studies have reported contradictory findings on age-related differences in the Big Five personality constructs in age-heterogeneous samples of young, middle-aged, and older adults (Allemand, Zimprich, & Hendriks, 2008; Allemand, Zimprich, & Hertzog, 2007; Roberts, Walton, & Viechtbauer, 2006). Generally, findings on age differences in personality domains may be interpreted only with caution as long as it has not been shown that the underlying factorial structure is equivalent across age groups. One reason may be that the cognitive processing of items may be limited in older age, particularly in personal- or telephone-interviewing contexts. Another reason is that the meaning of some statements may have shifted across cohorts, causing different patterns of factor loadings between birth cohort groups. More generally, effects of age, gender, education, or experimental conditions will be valid only as long as it is shown that measurement equivalence holds across all comparison units.

In this research, we applied exploratory structural equation modeling (ESEM) for analyzing the Big Five structure of personality domains. This new approach extends and advances traditional approaches that have identified the Big Five personality structure with exploratory factor analyses (EFA; McCrae & Costa, 1997). In this vein, there has been a long debate as to whether confirmatory factor analyses (CFA) or structural equation modeling (SEM) may also be applied for identifying the five-factor model of personality domains at all (e.g., Aluja, García, García, & Seisdedos, 2005; Borkenau & Ostendorf, 1990; Church & Burke, 1994; McCrae, Zonderman, Costa, Bond, & Paunonen, 1996; Vassend & Skrondal, 1997). SEM and CFA typically require specifying loadings for each indicator on its congruent factor, while noncongruent factor loadings are constrained to zero. As a consequence,

CFA may not obtain an adequate fit of the Big Five model to the data, while, at the same time, artificially inflating correlations of the five factors (cf. Hopwood & Donnellan, 2010). It was therefore concluded that CFA and SEM are not suitable for identifying the Big Five personality constructs (McCrae et al., 1996). By contrast, traditional EFA has limitations with respect to comparing equivalence of factor structures and, thus, proving measurement invariance across multiple groups. In this vein, Marsh and his colleagues (Marsh et al., 2010; Marsh et al., 2009) previously outlined a new approach that integrates EFA and CFA in the framework of SEM—that is, ESEM, recently implemented in the Mplus statistical software package (Mplus 5.2; Muthén & Muthén, 2008; see also Asparouhov & Muthén, 2009). The general idea of ESEM implies that factor loadings of items are specified as in a traditional EFA, while at the same time giving access to parameter estimates, standard errors, and goodness-of-fit statistics that are typically associated with CFA (Asparouhov & Muthén, 2009; Marsh et al., 2010). Most important, ESEM provides comfortable and efficient tools for multigroup comparisons of mean levels and tests of measurement equivalence of factor structures.

The present research

In the German SOEP (Wagner et al., 2007) and a related telephone-interviewing study, a short 15-item version of the Big Five Inventory (BFI–S) was assessed with different survey methods such as CAPI, PAPI, and CATI, as well as traditional self-administered paper-and-pencil questionnaires (SELF) with or without an interviewer present. Schräpler, Schupp, and Wagner (2010) did not find any mode effects in different outcome variables of the study due to the mixed-method approach in the survey.

In the present research, we examined the extent to which use of different assessment methods would have an effect on the reliability, and equivalence of the Big Five factor structure of personality trait domains across early, middle, and late adulthood. In particular, we expected that because telephone interviewing entails greater cognitive demands in later adulthood, this might challenge the five-factorial structure of the five personality constructs. While face-to-face interviewing has been found to have small effects on the reliability of Big Five assessment across adulthood, the telephone interviewing method was expected to be associated with reduced factorial invariance in later adulthood. In the SOEP study, we were also able to compare the test–retest stability of the Big Five personality assessment with two method conditions (FACE, SELF) across a 5-year interval for three birth cohort groups.

Method

Our research was based on two studies with representative German samples that included different types of interviewing methods. First, the German SOEP (Wagner et al., 2007) is a representative household panel study with probability sampling that has been conducted annually since 1984. Big Five constructs were included in the SOEP waves of 2005 and 2009. In 2005, 19,351 participants completed the BFI questionnaire across conditions. Of these, a total of 13,459 participants (69.6%) had completed the BFI questionnaire in the same or different assessment conditions, allowing for tests of the rank order stability coefficients of Big Five constructs. Second, as a study related to the SOEP, we conducted a cross-sectional, representative multipurpose telephone survey including the Big Five–S inventory, using a CATI method (for a description, see Lang, Baltes, & Wagner, 2007). The CATI study was conducted in 2005 and included 1,200 participants, who responded to a telephone survey. In Germany, more than 97% of households have a telephone. The probability sample represents the heterogeneity of the German population 20–80 years of age with a telephone in the household with respect to central demographic variables.

Description of study participants In the face-to-face interviewing conditions (i.e., PAPI, CAPI) of the SOEP 2005, complete data for the Big Five SOEP inventory and for the method group indicator variable were available for 11,266 participants, and complete data were available for 8,085 participants in the SELF condition. In the CATI study, 1,178 respondents with complete data participated. We excluded participants who did not have complete data for the BFI–S. Selectivity analyses revealed minor differences for the excluded participants with respect to sociodemographic variables. Excluded participants were older (M = 56.92, SE = 0.69) than participants with complete data (M = 48.38, SE = 0.12), $t(21179) = 12.95$, $p < .001$. Additionally, excluded participants were less educated (i.e., fewer educational years; M = 9.65, SE = 0.03) than participants with full data (M = 9.92, SE = 0.01), $t(19253) = 7.37$, $p < .001$. There was no group difference with respect to gender composition (chi-square/df = 3/1, n.s.). Possible resulting confounds of missingness with assessment methods are addressed in the presentation of the results.

With respect to chronological age, the sample was representative for the age heterogeneity in the population. Categorization of a continuous age variable in three, ten or more groups is arbitrary (e.g., Baltes, Reese, & Nesselroade, 1977). In order to maximize statistical power of within age-group comparisons, we defined three age brackets consisting of young adulthood from 20 to 39 years (M = 30.24, SD =

5.96), middle adulthood from 40 to 59 years (M = 48.66, SD = 5.65), and old adulthood at 60 years and older (M = 69.6, SD = 7.15). Table 1 gives an overview of sample characteristics of gender, age, and education across the assessment conditions. There was no randomized assignment of participants to the different assessment procedures. Within the SOEP, assignment to either self-administered questionnaire completion or face-to-face personal-interviewing procedures (FACE: PAPI or CAPI) was dependent on the availability of respondents at the time the interviewer visited the household. Participants in the FACE group were less educated (chi-square/df = 1,148/6, p < .001) and older (chi-square/df = 734/4, p < .001), as compared with the two other conditions.

Interviewing and assessment procedures The SOEP study has used different methods of assessment, such as CAPI, PAPI, and self-administered completion of questionnaires with or without an interviewer present (SELF–alone, SELF–not-alone). We decided to concatenate the two face-to-face personal interviewing conditions, as well as the self-administration conditions with and without the interviewer present in the following analyses. We decided to pool these related method conditions for two reasons: First, procedures in the two face-to-face conditions (i.e., CAPI, PAPI), and, respectively, in the two SELF conditions (i.e., SELF–alone, SELF–not-alone) were nearly identical. For example, PAPI means that the interviewer marks the participant's responses on paper (i.e., PAPI without self-completion) or, alternatively, assists the respondent with completing the questionnaire (i.e., PAPI assisted). In all, the interviewer's direct assistance is embedded in an identical face-to-face interviewing process. In sum, 6,532 participants took part in the paper-assisted conditions in 2005 (PAPI without self-completion: N = 5,137; PAPI assisted: N = 1,395). In the SELF–not-alone condition, participants completed questionnaires while the interviewer was present but did not assist (N = 5,434). In the SELF–alone condition, 2,651 participants completed the questionnaire without the assistance of an interviewer. Completion of self-administered paper-and-pencil questionnaires (SELF) was implemented in the SOEP study as an alternative assessment method when face-to-face interviewing was not feasible (e.g., participants were unavailable in the household). In these cases, respondents were also able to fill out a paper questionnaire alone—that is, without the assistance of an interviewer—and send it back via regular mail in a stamped envelope. The selection of this procedure depended on arbitrary reasons, because the interviewers visited the households without advance notice. Selectivity analyses have not shown differences between samples of respondents who completed the questionnaire with and without the interviewer. We consider these two procedures to be very similar. Thus, we did not have reason to expect method differences on self-report assessments. Consequently, we assumed a strong concordance between the two FACE conditions, on the one hand, and the two SELF conditions, on the other hand. Since our aim was to challenge the robustness of personality assessment with different interviewing methods, we attempted to maximize the potential of identifying effects of diverse procedures. Therefore, we combined methods from identical branches in order to identify clear-cut groups of interviewing methods—that is, face-to-face interviewing (FACE), telephone interviewing (CATI), and self-administered questionnaires (SELF). With this distinction, we wanted to reduce the risk of overlooking or not detecting a truly existing method effect between the procedures. Second, in order to make sure that none of the information was lost, we conducted separate statistical analyses testing for the measurement invariance between CAPI and PAPI conditions, as well as between the SELF–alone and the SELF–not-alone conditions. These tests did not reveal any substantive differences between the respective procedures. (Supplementary information material describing these results—here, referring to Tables SM4–SM6—may be downloaded at http://www.geronto.uni-erlangen.de/pdfs/Langetal_BigFive_SIM.pdf).

Table 1 Overview of percentages of sample characteristics in the FACE, CATI, and SELF conditions

	FACE	SELF	CATI	χ^2/df
Gender				
Male	47.2	48.8	47.5	
Female	52.8	51.2	52.5	5/2
Age				
20 – 39 years	28.1	38.7	44.7	
40 – 59 years	36.7	42.4	36.2	
≥ 60 years	35.2	18.9	19.1	734/4***
Education in years				
Other[a]	10.7	7.4	0.3	
8 – 9 years	42.7	25.8	19.8	
10 years	25.9	33.6	35.0	
≥ 11 years	20.7	33.2	44.9	1148/6***

FACE, N = 11,266; CATI, N = 1,178; SELF, N = 8,085. *** p < .001,
[a] includes participants with continued education or not identified educational degree

The short Big Five SOEP Inventory The Big Five personality constructs were assessed at two measurement occasions: in 2005 and in 2009. The Big Five personality structure was assessed with a 15-item German adoption of the Big Five Inventory Version (BFI–S; Gerlitz & Schupp 2005; John & Srivastava, 1999). With regard to reliability and validity, the BFI–S is a reasonable, short instrument

designed to measure the Big Five personality factors in large surveys (Lang, 2005; Lang, Lüdtke, & Asendorpf, 2001, Rammstedt, Goldberg, & Borg, 2010). The validity and reliability of a paper-and-pencil version of the BFI–S proved acceptable in comparison with the German version of the NEO-FFI and other external criteria of validity (Gerlitz & Schupp, 2005; Lang, 2005). Within the BFI–S, each of the five personality factors is measured with 3 items on a 7-point Likert scale from 1 (*strongly disagree*) to 7 (*strongly agree*). Cronbach's alpha values for the short 3-item subscales of the Big Five constructs were low, reflecting the width of these broad constructs (i.e., neuroticism, α = .60; extraversion, α = .66; openness, α = .63; agreeableness, α = .50; conscientiousness, α = .60). Table 2 gives an overview of the 15 items that assess each of the five personality constructs.

Effects of changes of interviewing procedures after 5 years in the SOEP study Another test of reliability relates to the 5-year retest stability of the short Big Five SOEP inventory across interviewing procedures. Across the two waves (2005 and 2009), 1,823 participants completed the questionnaire with different procedures (i.e., change from FACE to SELF or vice versa). In the panel data, 6,768 participants completed the interview with the same FACE procedure at both waves, and 4,868 participants completed the questionnaire in the SELF condition at both waves. In this context, we explored the stability coefficients for those participants that were included in either the CAPI or the SELF conditions of the 2005 waves of the SOEP and responded to the same or a different interviewing (CHANGE group) procedure in 2009 (see the supplementary materials, Table SM9).

Statistical analysis and procedure for testing measurement equivalence We tested the measurement invariance of the short SOEP Big Five Inventory (BFI–S) in three different assessment conditions across three age groups—that is, FACE, CATI, and SELF. We conducted ESEM analyses with Mplus (Version 5.2; Muthén & Muthén, 2008). We used the maximum likelihood robust (MLR) estimator, which is robust in relation to nonnormality (Muthén & Muthén, 2008). Within the ESEM framework, all factor loadings for an a priori postulated number of factors and the item intercepts, as well as the item uniquenesses are estimated. ESEM provides the possibility of testing the invariance of these parameters across multiple groups (for further details of the ESEM approach, see Asparouhov & Muthén, 2009). Oblique quartimin rotation was conducted, since this rotation technique was more appropriate for our simple structure data than was Geomin rotation. For example, quartimin is the more effective rotation for identifying simple loading structures (see Asparouhov & Muthén, 2009). The loading pattern for the five-factor model in our data was in accordance with this notion. Also, quartimin does not depend on the value for ε (a small constant). Therefore, quartimin results in just one rotation solution, instead of multiple solutions. Regarding models with several factors, multiple solutions are more likely in Geomin (Asparouhov & Muthén, 2009). We argue that the solution for the well-established five-factor model is straightforward in the context of this study.

Evaluation of equivalent measurement models relied on the Comparative Fit Index (CFI), the Tucker–Lewis Index (TLI), and the root mean square error of approximation (RMSEA). Common guidelines for evaluating fit indices are CFI and TLI values greater than .90 and .95 for acceptable and excellent data fits and RMSEA values less than .05 and .08 for close and reasonable fits to the data (Marsh, Hau, & Wen, 2004). SEM tests the assumption that an a priori specified model fits the data. Model fit is classically indicated by a nonsignificant chi-square test. However, chi-square testing is sample size dependent (Tucker & Lewis, 1973). We chose fit indices that are relatively robust to sample size differences (e.g., CFI, TLI, RMSEA; see Marsh et al., 1998). Chen (2007) suggested that a more parsimonious model is supported if the change in the CFI is less than .01 or if the change in the RMSEA is less than .015. A conservative criterion for the more parsimonious model is that the values of the TLI and RMSEA are equal to or even better than the values for the respective less restrictive model (Marsh et al., 2009). Bentler (1990) suggested testing of nested models in order to evaluate the more parsimonious model. However, Brannick (1995) pointed out that evaluation of nested models with chi-square difference testing remains dependent on sample size. We are aware that this issue is an area of on-going debate. We submit that a conservative evaluation of changing fit indexes is an appropriate way for evaluating nested models in studies with large sample sizes. Our evaluation criteria for nested models was based on guidelines derived from methodological research (e.g., Chen, 2007; Cheung & Rensvold, 2001)

Models of measurement equivalence were tested and compared using the following five steps. (1) *Configural invariance* is defined as the least demanding model that imposes no invariance constraints. The configural invariance model was tested to establish the baseline assumption that the five-factor structure was prevalent in all assessment conditions. (2) The w*eak measurement invariance model* constrained factor loadings to be invariant across method groups. (3) In the *strong measurement invariance model*, all item intercepts were additionally constrained to be invariant across the three methods. The rejection of this model implies differential item functioning (i.e., differences in mean levels of items between the three methods cannot be

Table 2 Descriptive statistics for FACE–CATI–SELF for young, middle-aged, and older adults

I See Myself as Someone Who …	Young Adults (N = 6,819)						Middle-Aged Adults (N = 7,982)						Older Adults (N = 5,728)					
	FACE		CATI		SELF		FACE		CATI		SELF		FACE		CATI		SELF	
	M	*SD*	*M*	*SD*	*M*	*SD*	*M*	*SD*	*M*	*SD*	*M*	*SD*	*M*	*SD*	*M*	*SD*	*M*	*SD*
Worries a lot (N)	4.52	1.72	4.74	1.84	4.64	1.61	4.75	1.66	4.98	1.77	4.83	1.58	4.86	1.70	4.98	1.81	4.89	1.57
Gets nervous easily (N)	3.52	1.67	3.58	1.78	3.73	1.65	3.59	1.77	3.74	1.78	3.79	1.67	3.78	1.84	3.90	1.94	4.02	1.66
Remains calm in tense situations (N, recoded)	3.36	1.50	3.29	1.48	3.58	1.47	3.34	1.52	3.16	1.47	3.58	1.48	3.40	1.54	3.15	1.45	3.53	1.47
Is talkative (E)	5.64	1.30	5.58	1.28	5.45	1.29	5.60	1.32	5.66	1.26	5.42	1.28	5.49	1.42	5.86	1.19	5.26	1.37
Is outgoing, sociable (E)	5.26	1.37	5.26	1.43	5.14	1.40	5.15	1.44	5.25	1.46	4.92	1.43	5.02	1.51	5.53	1.40	4.92	1.48
Is reserved (E, recoded)	4.02	1.62	4.29	1.83	4.17	1.67	3.86	1.63	3.83	1.73	3.99	1.59	3.57	1.59	3.39	1.67	3.66	1.58
Is original, comes up with new ideas (O)	4.78	1.39	5.02	1.21	4.73	1.26	4.65	1.43	4.96	1.31	4.68	1.31	4.18	1.60	4.96	1.50	4.35	1.47
Values artistic, aesthetic experiences (O)	4.07	1.78	4.42	1.78	3.92	1.80	4.21	1.83	4.77	1.71	4.07	1.76	4.24	1.86	5.16	1.59	4.21	1.75
Has an active imagination (O)	5.05	1.44	5.25	1.46	4.98	1.48	4.79	1.50	5.24	1.44	4.73	1.49	4.54	1.63	5.23	1.48	4.63	1.53
Is sometimes rude to others (A, recoded)	5.09	1.67	4.50	1.76	4.79	1.61	5.13	1.66	4.65	1.87	4.93	1.63	5.35	1.67	4.45	2.08	5.07	1.68
Has a forgiving nature (A)	5.50	1.31	5.61	1.35	5.40	1.26	5.56	1.31	5.74	1.27	5.45	1.32	5.61	1.34	5.96	1.21	5.36	1.41
Is considerate and kind to almost everyone (A)	5.81	1.09	5.83	1.07	5.72	1.07	5.81	1.09	5.81	1.10	5.73	1.10	5.91	1.09	6.00	1.10	5.64	1.24
Does a thorough job (C)	6.20	0.99	6.17	0.91	6.07	0.96	6.35	0.92	6.26	0.90	6.26	0.91	6.20	1.10	6.24	0.95	6.09	1.15
Tends to be lazy (C, recoded)	5.65	1.54	5.17	1.76	5.27	1.60	6.05	1.35	5.47	1.76	5.83	1.45	6.09	1.38	5.36	1.87	5.75	1.53
Does things efficiently (C)	5.86	1.07	5.80	1.00	5.76	0.97	5.93	1.08	5.82	1.05	5.80	1.04	5.72	1.26	5.96	1.03	5.66	1.25

Young adults, N (FACE) = 3,164; N (CATI) = 527; N (SELF) = 3,128. Middle-aged adults, N (FACE) = 4,131; N (CATI) = 426; N (SELF) = 3,425. Older adults, N (FACE) = 3,971; N (CATI) = 225; N (SELF) = 1,532. N Neuroticism, E Extraversion, O Openness to experience, A Agreeableness, C Conscientiousness

explained merely by differences at the factorial mean level). Finally, (4) *Strict measurement invariance* was tested, requiring invariance of factor loadings, item intercepts, and item uniquenesses across FACE, CATI, and SELF. The rejection of this model implies that there are measurement error differences between the three methods. Changes in *df* = 30 would result from equality constraints on uniquenesses of items. (5) In a last step, we then added a test of the invariance of factor means across the three methods.

Results

Table 2 gives an overview of the means and standard deviations for the 15 items of the BFI–SOEP in three assessment conditions for three age cohort groups. We conducted ESEM in the same set of four measurement models (configural, weak, strong, strict) in four data-analytical steps. First, we tested measurement invariance across all samples in order to establish the general robustness of the assessment methods across all age groups. In a next step, we subsequently tested the measurement invariance of the assessment methods within each age group of young adults, middle-aged adults, and older adults. In this research, we examined measurement invariance and reliability of the Big Five structure of personality domains across assessment context for different age groups from early to late adulthood.

Measurement invariance comparing FACE, CATI, and SELF contexts Table 3 summarizes the goodness-of-fit statistics for the four invariance models across all samples of the SOEP and the related CATI study.[1] Fit indices showed close fit of the configural model (CFI = .969, TLI = .918, RMSEA = .044; see Table 3) and for the *weak measurement invariance model* of invariant factor loadings (chi-square/*df* = 1,843/220, CFI/TLI = .967/.953, RMSEA = .033). Fit indices of the *weak measurement invariance model* compared with the *configural invariance model* indicated improvement (TLI, .953 vs. .918; RMSEA, .033 vs. .044), while the difference in the CFI (.967 vs. .969) was less than .01. Fit indices for the *strong measurement invariance model* again revealed good model fit (CFI = .959, TLI = .946, RMSEA = .035). The difference between the CFI values of the strong measurement and the weak measurement invariance models (.959 vs. .967) was less than .01, and the TLI (.946 vs. .953) and RMSEA (.035 vs. .033) remained essentially stable. The fit of the *strict measurement invariance model* (CFI = .952, TLI = .943, RMSEA = .036) proved satisfactory as well. A comparison of fit indices of the strict and the strong measurement invariance models revealed no CFI change greater than .01 (.952 vs. .959). Additionally, the TLI (.943 vs. .946) and RMSEA (.036 vs. .035) indicated improved model parsimony.

As was stated above, we excluded participants who did not have complete data for the Big Five inventory. We considered this the more conservative test of assessment robustness. In order to avoid a possible confounding of listwise deletion with effects of assessment methods, we reanalyzed the data using full information maximum likelihood (FIML): Again, we observed close model fits for the configural, weak, strong, and strict invariance models. The fit indices correspond closely to model results obtained for listwise deletion. In conclusion, results do not differ between listwise deletion and FIML.

In conclusion, the juxtaposition of the four models of measurement invariance supports the assumption of equal factor loadings, equal item intercepts, and equal item uniquensses (i.e., strict measurement invariance) of method conditions. Finally, we tested the *invariance of factor means* across methods. The strict measurement invariance models constrain mean differences between the three methods to be invariant. Therefore, we compared the strict measurement invariance model without (model a, as above) and with (model b) constrained factor means. The fit indices showed a close fit of model b (chi-square/*df* = 3,220/280, CFI = .941, TLI = .934, RMSEA = .039). Comparison of strict invariance model a with model b revealed a CFI change of .011 (.941 vs. .952) and a change in the TLI (.934 vs. .943) and RMSEA (.039 vs. .036). Table 4 displays the effect sizes for the factor mean differences between FACE, CATI, and SELF. In Table 4, factor means of FACE are the reference means, while the factor means of CATI and SELF are the comparison means. For standardizing the factor mean differences, we used the pooled standard deviation of the reference group (FACE)

[1] Factor variance–covariance invariance is not in the center of measurement invariance studies, because they typically focus on unidimensional models. However, this matters with respect to convergent and discriminant validity. Therefore, we tested extended weak measurement invariance models with invariant factor loadings, factor variances, and factor covariances. These extended weak measurement invariance models across method groups (FACE–CATI–SELF) revealed close model fit for young adults (chi-square/*df* 891/250, CFI/TLI = .962/.952, RMSEA = .034), for middle-aged adults (chi-square/*df* = 907/250, CFI/TLI = .966/.957, RMSEA = .031), and for old adults (chi-square/*df* = 750/250, CFI/TLI = .965/.956, RMSEA = .033). Comparisons of these extended models with weak measurement invariance models support the notion that factor variances and factor covariances are invariant across method groups in young, middle-aged, and old adults. Supplementary information material (SIM) is made available online for download at the following URL: http://www.geronto.uni-erlangen.de/pdfs/Langetal_BigFive_SIM.pdf. The supplement (Tables SM2 and SM3) reports standardized factor loadings across age groups and factor correlations across age groups. The supplement (Tables SM15 and SM 16) also provides detailed information on measurement invariance of age by method group comparisons.

Table 3 Summary of goodness-of-fit statistics for measurement invariance across the total sample in young, middle-aged, and older adults (FACE–CATI–SELF)[a]

Model	Total Sample					Young Adults					Middle-Aged Adults					Older Adults				
	MLR/df	Nfp	CFI	TLI	RMSEA	MLR/df	Nfp	CFI	TLI	RMSEA	MLR/df	Nfp	CFI	TLI	RMSEA	MLR/df	Nfp	CFI	TLI	RMSEA
Configural invariance	1676/120	285	.969	.918	.044	602/120	285	.971	.925	.042	722/120	285	.969	.918	.043	587/120	285	.968	.915	.045
Weak invariance	1843/220	185	.967	.953	.033	785/220	185	.966	.952	.034	847/220	185	.967	.953	.033	695/220	185	.967	.953	.034
Strong invariance	2291/240	165	.959	.946	.035	959/240	165	.957	.944	.036	1021/240	165	.959	.947	.035	777/240	165	.963	.951	.034
Strict invariance	2687/270	135	.952	.943	.036	1102/270	135	.950	.942	.037	1158/270	135	.954	.946	.035	916/270	135	.955	.948	.036
Factor mean invariance	3220/280	125	.941	.934	.039	1234/280	125	.943	.936	.039	1326/280	125	.946	.939	.037	1099/280	125	.943	.936	.039

Total sample: for FACE, *N*=11,266; for CATI, *N* = 1,178; for SELF, *N* = 8,085. Young adults: for FACE, *N* = 3,164; for CATI, *N* = 527; for SELF, *N* = 3,128. Middle-aged adults: for FACE, *N* = 4,131; for CATI, *N* = 426; for SELF, *N* = 3,425. Old adults: for FACE, *N* = 3,971; for CATI, *N* = 150; for SELF, *N* = 1,532. MLR/*df* = maximum likelihood robust chi-square/degrees of freedom ratio; *Nfp* = number of free parameters; CFI = comparative fit index; TLI = Tucker–Lewis-Index; RMSEA = root mean square error of approximation

[a] For older adults: only highly educated participants of the CATI study were included

and the comparison group (CATI or SELF). Findings point to invariance of factor means across methods, with the exception of the openness-to-experience construct. In conclusion, the means for openness to experience were higher in the CATI condition, as compared with the other two method conditions.

In sum, factor means in the telephone interviewing (CATI) study appear to generate less stable factor means with regard to the openness-to-experience personality domain. In a next step, we tested for measurement invariance separately within each age cohort group, with particular attention to the openness construct and potential differences related to CATI assessment contexts.

Measurement invariance in young adulthood Table 3 summarizes the goodness-of-fit statistics for the four invariance models in young adulthood. Goodness-of-fit criteria suggested close fit for the *configural invariance model* (CFI = .971, TLI = .925, RMSEA = .042), as well as for the *weak measurement invariance model* (chi-square/*df* = 785/220, CFI/TLI = .966/.952, RMSEA = .034). Comparing the fit of the *weak invariance model* with that of the *configural invariance model* indicated improved model parsimony (i.e., TLI, .952 vs. .925; RMSEA, .034 vs. 042), while the CFI difference was less than .01 (.966 vs. .971). Fit indices for *strong measurement invariance* were satisfactory (CFI = .957, TLI = .944, RMSEA = .036). The difference in the CFI between the strong and the weak measurement invariance models was less than .01 (.957 vs. .966), while the TLI (.944 vs. .952) and RMSEA (.036 vs. .034) proved stable. Fit indices for the *strict measurement invariance model* (CFI = .950, TLI = .942, RMSEA = .037) were again satisfactory. Fit indices between the strict and the strong measurement invariance models revealed a change in the CFI of less than .01 (.950 vs. .957), with the TLI (.942 vs. .944) and RMSEA (.037 vs. .036) fairly unchanged. In conclusion, the four measurement invariance models support the assumption of equal factor loadings, equal item intercepts, and equal item uniquenesses (i.e., strict measurement invariance) across method conditions in young adulthood. In addition, we also tested *invariance of factor means* across methods. In this model, factor means were constrained to be invariant across methods (strict invariance model b). Fit indices showed close fit of model b (chi-square/*df* = 1,234/280, CFI = .943, TLI =. 936, RMSEA = .039). Differences in fit indices compared with the strict invariance model revealed a CFI change (.943 vs. .950) of less than .01. Finally, the TLI (.936 vs. .942) and RMSEA (.039 vs. .037) values remained essentially the same. In conclusion, as is shown in Table 4, the data support the assumption that factor means are generally invariant and fairly robust across methods in young adulthood. The standardized factor loadings were in accordance with the

Table 4 Standardized differences of factor means between FACE–CATI–SELF for the strict measurement invariance model[a]

	Total Sample			Young Adults			Middle-Aged Adults			Older Adults		
	FACE	CATI	SELF	FACE	CATI	SELF	FACE	CATI	SELF	FACE	CATI	SELF
N	.000	.074	.127	.000	.087	.170	.000	.117	.151	.000	.051	.151
E	.000	.097	-.069	.000	.014	-.070	.000	.010	-.130	.000	.194	-.123
O	.000	.563	.051	.000	.331	-.050	.000	.536	-.034	.000	.757	.116
A	.000	-.251	-.281	.000	-.243	-.208	.000	-.134	-.162	.000	-.299	-.364
C	.000	-.126	-.169	.000	-.123	-.198	.000	-.195	-.161	.000	.015	-.146

N Neuroticism, *E* Extraversion, *O* Openness, *A* Agreeableness, *C* Conscientiousness

[a] For older adults: only highly educated participants of the CATI study were included

expected simple structure model for each age group (see Table 5). Only 3 out of 45 convergent loadings were lower than .43. All diverging loadings were lower than .30 (with only 8 of 180 loadings higher than .20). In addition, factor correlations across the three method conditions among the young, middle-aged, and older adults (see Table 6) were in accordance with findings using longer Big Five measures (Benet-Martinez & John, 1998; Lang et al., 2001). Moreover, factor correlations (Table 6) do not point to a method-related pattern of differentiation between the three age groups, after excluding low-educated older adults in the CATI condition.

Measurement invariance in middle adulthood Table 3 summarizes the goodness-of-fit statistics for the four invariance models in middle adulthood. The fit indices (CFI = .969, TLI = .918, RMSEA =. 043; see Table 3) showed good model fit of the configural model. The *weak measurement invariance model* (i.e., invariant factor loadings across the three groups) obtained satisfactory model fit (CFI = .967, TLI = .953, RMSEA = .033). Comparison of fit indices of the weak invariance model with the configural model indicated improved model parsimony (TLI, .953 vs. .918; RMSEA, .033 vs. .043). Also, the difference in CFI (.967 vs. .969) was less than .01. The *strong measurement invariance model* revealed close model fit (CFI = .959, TLI = .947, RMSEA = .035; see Table 3). The difference in CFI values for the strong measurement invariance model, as compared with the weak measurement invariance model, was negligible at less than .01 (.959 vs. .967). The TLI (.947 vs. .953) and RMSEA (.035 vs. .033) values remained essentially the same. This supports the assumption of strong measurement invariance across methods in middle adulthood. The *strict measurement invariance model* (i.e., invariant factor loadings, item intercepts, and item uniqueness) showed satisfactory model fit (CFI = .954, TLI = .946, RMSEA = .035; see Table 3). Differences in fit indices between the strict invariance model and the strong invariance model revealed no substantive CFI change (.954 vs. .959; ≤ .01). The TLI (.946 vs. .947) and RMSEA (.035 vs. .035) did not differ substantively. Results confirm the assumption of equal factor loadings, equal item intercepts, and equal item uniqueness (i.e., strict measurement invariance) across methods in middle adulthood. In a last step, we again tested for invariance of factor means. We compared the strict measurement invariance without (model a; see above) and with (model b) constrained factor means, which indicated a close fit of model b (chi-square/*df* = 1,326/280, CFI = .946, TLI =. 939, RMSEA = .037). Comparison of fit indices between strict invariance models a and b revealed a change in the CFI of less than .01 (.946 vs. .954). Also, the TLI (.939 vs. .946) and RMSEA (.037 vs. .035) pointed to invariance of factor means. In conclusion, as is shown in Table 4, the data support the assumption that factor means are generally invariant and fairly robust across methods in middle adulthood.

Measurement invariance in late adulthood: the challenges of telephone interviewing In a last step, we tested measurement invariance of the five personality factors across FACE, CATI, and SELF conditions in later adulthood only (i.e., 60 years and older). Fit indices pointed to close fit of the baseline *configural measurement invariance model* (chi-square/*df* = 562/120, CFI/TLI = .969/.920, RMSEA = .044). However, in the CATI study, two items had negative residual variances, leading to a nonpositive definite covariance matrix.

We were able to rule out the possibility of linear dependency and multicollinearity after checking item covariances and correlations. Due to skewed items, we chose MLR estimators to counteract possible distortions. Consequently, the nonpositive definite covariance matrix may also point to a model misspecification in the CATI assessment with older adults, as we had expected. Therefore, we tested for measurement invariance of FACE and SELF conditions only, while excluding the CATI condition. Also, fit of the model (chi-square/*df* = 83/40, CFI/TLI =

.909/.761, RMSEA = .069) for older participants in the telephone interview was not acceptable. Consequently, further measurement invariance testing was skipped (Meredith, 1993).[2]

However, test of measurement invariance among older adults for FACE versus SELF conditions reached close fit for the configural invariance model (chi-sqare/*df* = 462/80, CFI = .972, TLI =.927, RMSEA = .042) and for the weak invariance model (chi-sqare/*df* = 545/130, CFI = .970, TLI = .951, RMSEA = .034). Comparison of the weak and the configural invariance models indicated improvement (TLI, .951 vs. .927; RMSEA, .034 vs. .042), while the CFI difference was less than .01 (.970 vs. .972). Fit indices for the strong measurement invariance model were satisfactory (chi-sqare/*df* = 577/140, CFI = .968, TLI = .952, RMSEA = .034). The CFI difference between the strong and the weak measurement invariance models was less than .01 (.968 vs. .970), and the TLI (.952 vs. .951) and RMSEA (.034 vs. .034) did not differ. Finally, the test of the strict measurement invariance model also showed satisfactory model fit (chi-sqare/*df* = 653/155, CFI = .640, TLI = .951, RMSEA = .034). The CFI change, as compared with the strong invariance model (.964 vs. .968), was less than .01, while the TLI (.51 vs. .952) and RMSEA (.034 vs. .034) remained unchanged. Indices for the strict measurement invariance model with additionally constrained factor means (strict invariance model b) revealed a close fit (chi-square/*df* = 780/160, CFI = .955, TLI =. 940, RMSEA = .038) and no substantive difference from the strict invariance model (CFI, .955 vs. .964; < .01; TLI, .951 vs. .90; RMSEA: .034 vs. .038). Thus, results indicate invariant factor means, factor loadings, item intercepts, and item uniqueness (i.e., strict measurement invariance) across FACE and SELF conditions in older adulthood.

In the following, we tested two possible explanations for the observed distortions in the CATI assessment of the Big Five personality dimensions among older adults. First, we expected that mental workload of the telephone interviewing context would preclude valid self-report responses, since it requires listening to interviewers while reflecting responses on a 7-point rating scale. An implication is that more highly educated older adults—for example, due to better training with such contexts—fare better with mastering this task. A second assumption was that the costs of mental workload would show up as greater variability in item responses for older adults within each of the five constructs, which might result in reduced likelihood of identifying the expected five factors in a very brief Big Five inventory.

Exclusion of older respondents with less education from the CATI study Testing the assumption that the higher mental workload of the CATI procedure would distort self-report responses to the BFI–S items among less well-educated older participants, we included only highly educated older participants from the CATI study (i.e., excluding 74 older adults with only 8 or 9 years of education and one person with a missing value for education; thus, we used $N = 150$ with 10 or more years of schooling) in a test of measurement invariance across methods in this age group. In the FACE and SELF groups, all educational levels were included. As is shown in Table 3, fit indices revealed close fit for the configural invariance model (CFI = .968, TLI = .915, RMSEA =.045) and for the weak measurement invariance model (CFI = .967, TLI = .953, RMSEA = .034). Differences between the weak and configural invariance models indicated improved model fit (TLI, .953 vs. .915; RMSEA, .034 vs. 045), with a CFI difference of less than .01 (.967 vs. .968). The strong measurement invariance model showed close fit (CFI = .963, TLI = .951, RMSEA = .034; see Table 3) and remained essentially stable, as compared with the weak invariance model (CFI difference, .963 vs. .967; = .01; TLI: .951 vs. .953; RMSEA, .034 vs. .034). Fit indices for the strict measurement invariance model were satisfactory as well (CFI = .955, TLI = .948, RMSEA = .036), and proved robust in comparison with the strong invariance model (CFI change, .955 vs. .963; TLI, .948 vs. .951; RMSEA, .036 vs. .034). In a final step, constraining the factor means (model b) revealed a close fit for the strict invariance model (chi-square/*df* = 1,099/280, CFI = .943, TLI =. 936, RMSEA = .039).

However, differences between models did not indicate improved parsimony for the strict measurement invariance model with and without constrained means (CFI change > .01, .943 vs. .955; TLI, .936 vs. .948; RMSEA, .039 vs. .036). In conclusion, factor means may not be invariant across assessment conditions among older adults. When

[2] Despite the recommendation of Meredith (1993), we tested models of weak, strong, and strict measurement invariance across assessment methods in late adulthood. We fixed the error variances for the two unidentified items at 0 for the configural invariance model (i.e., *configural invariance model 2*). For this configural invariance model 2, model fit was improved (chi-square/*df* = 552/122, CFI/TLI = .970/.923, RMSEA = .043). However, the weak measurement invariance model did not converge due to serious iteration problems. The model still did not converge after enhancing the number of iterations and defining start values. A check of the standardized factor loadings in the CATI analysis (see the supplementary material, Table SM7) confirmed that the five factors were not prevalent in the old age group; that is, conscientiousness, extraversion, and agreeableness were not identified well. After lowering the convergence criterion from .00005 (default) to .01, the weak measurement invariance model showed acceptable fit (CFI = .955, TLI = .937, RMSEA = .039). However, a comparison of fit indices between the weak and configural invariance model 2 (CFI: .955 vs. .970) did not suggest that there were invariant factor loadings across methods in old age.

Table 5 Standardized factor loadings and unstandardized item intercepts (iic) for FACE–CATI–SELF[a]

I see Myself as Someone Who …	Young Adults						Middle-Aged Adults						Older Adults					
	N	E	O	A	C	iic	N	E	O	A	C	iic	N	E	O	A	C	Iic
Worries a lot	.534	-.023	.074	.068	.118	4.59	.486	-.030	.058	.078	.128	4.80	.464	-.015	-.006	.154	.049	4.86
Gets nervous easily	.737	-.017	.029	-.005	-.036	3.62	.800	.008	.017	.007	-.022	3.69	.804	.011	.036	-.015	-.002	3.84
Remains calm in tense situations[b]	.541	.027	-.230	-.109	-.055	3.46	.522	-.008	-.020	-.142	-.043	3.43	.478	-.019	-.293	-.117	-.074	3.43
Is talkative	.048	.717	.037	.072	.112	5.55	.023	.663	.049	.065	.120	5.52	.031	.684	.021	.076	.122	5.43
Is outgoing, sociable	-.005	.681	.123	.110	-.043	5.21	.028	.665	.119	.123	-.016	5.06	.007	.664	.140	.108	-.014	5.00
Is reserved[b]	-.101	.632	-.109	-.201	-.073	4.11	-.090	.587	-.112	-.211	-.088	3.92	-.091	.529	-.096	-.291	-.094	3.59
Is original, comes up with new ideas	-.080	.104	.616	-.133	.144	4.78	-.092	.096	.604	-.138	.146	4.68	-.043	.115	.602	-.147	.134	4.24
Values artistic, aesthetic experiences	.036	.013	.459	.114	-.051	4.03	.014	.034	.463	.105	-.049	4.18	.033	.026	.495	.105	.003	4.26
Has an active imagination	.035	.068	.580	.114	-.095	5.03	.029	.062	.632	.073	-.078	4.79	.001	.067	.662	.070	-.035	4.58
Is sometimes rude to others[b]	-.156	-.048	-.177	.543	-.025	4.91	-.150	-.030	-.212	.521	-.036	5.02	-.149	-.006	-.221	.498	-.065	5.25
Has a forgiving nature	-.011	.085	.045	.362	.041	5.46	-.005	.082	.085	.363	.055	5.52	.008	.152	.024	.434	-.011	5.55
Is considerate and kind to almost everyone	.038	.057	.064	.704	.073	5.77	.023	.049	.045	.747	.064	5.78	.010	.052	.058	.714	.086	5.84
Does a thorough job	.019	-.013	-.017	-.014	.816	6.14	.018	.003	-.026	-.019	.776	6.30	-.013	.018	-.048	-.039	.779	6.17
Tends to be lazy[b]	-.005	.074	-.196	.119	.458	5.44	-.059	.085	-.173	.094	.446	5.92	-.066	.062	-.255	.176	.326	5.97
Does things efficiently	-.052	.028	.115	.053	.607	5.81	-.044	-.001	.131	.084	.558	5.87	.017	.010	.149	.095	.600	5.71

N Neuroticism, *E* Extraversion, *O* Openness, *A* Agreeableness, *C* Conscientiousness

[a] For older adults, only highly educated participants from the CATI study were included. [b] Item was recoded (inversed)

Table 6 Standardized factor correlations for FACE–CATI–SELF[a]

	Young Adults					Middle-Aged Adults					Older Adults				
	N	E	O	A	C	N	E	O	A	C	N	E	O	A	C
N	1.00					1.00					1.00				
E	-.200	1.00				-.146	1.00				-.129	1.00			
O	-.015	.442	1.00			-.031	.511	1.00			-.019	.519	1.00		
A	-.036	.115	.213	1.00		-.067	.166	.195	1.00		-.082	.158	.165	1.00	
C	-.116	.230	.234	.358	1.00	-.123	.275	.272	.382	1.00	-.071	.373	.352	.389	1.00

N Neuroticism, *E* Extraversion, *O* Openness to Experience, *A* Agreeableness, *C* Conscientiousness

[a] For older adults, only highly educated participants from the CATI study were included

factor means in FACE and SELF were fixed to be equal, model fit improved for the strict invariance model (chi-square/*df* = 1,045/275, CFI = .947, TLI = .939, RMSEA = .039). As is shown in Table 4, factor means for the openness construct were higher among highly educated older participants who participated in the CATI study. This finding is less surprising considering that openness to experience is generally expected to be higher among highly educated adults.

Convergent construct variability of the Big Five factor structure across adulthood Another possible challenge of CATI assessment procedures may be related to the issue of generating consistent item responses in a telephone situation that involves greater attentional and cognitive resources. To evaluate possible age-associated costs in the CATI self-report assessment of the Big Five personality constructs, we investigated the intraindividual variability (inconsistency) of manifest responses within each of the convergent items as they deviate from the respective scale composite—that is, the variability within converging items of the five constructs (i.e., convergent construct variability [CCV]). We defined the CCV score as the sum of squared differences between each converging item score and the mean composite of the respective personality construct. More specifically, the formula $\sum_{j}^{5}\sum_{i}^{3}\left(x_{ij}-\bar{x}_{j}\right)^{2}$ was used for each person, where x_{ij} is a person's score on item i (i = 1, 2, 3) for construct j (j = 1, . . . 5), and $\bar{x}_{j}$ is the person's average across the three items for each construct j.

CCV indicates how well respondents' ratings of agreement with each item deviate from the expected convergence of the personality scale. CCV thus reflects the mean cross-item consistency of convergent items across all five personality constructs. Following the assumptions of classical test theory, high construct variability scores indicate higher individual measurement errors, whereas low construct variability scores indicate precise true score measures. One possible reason for such intraindividual deviations from the expected construct true score consists of distortions related to mental workload. Therefore, we expected negative age trends with regard to construct variability in CATI settings, indicating that within this survey method, older participants can be differentiated less with respect to the five personality constructs.

We conducted moderated hierarchical regression analyses of the CCV score on chronological age and assessment method (dummy 1, CATI vs. SELF; dummy 2, CATI vs. FACE). Chronological age was mean centered to prevent confounding of main and interaction effects. A first model confirmed significant main effects of age (β = .06, p < .001) and method contrasts (CATI–Self, β = .18, p < .001; CATI–FACE, β = .14, p < .001; SELF–FACE β = .04, p < .001). Thus, respondents differentiate more between items in CATI assessment procedures (i.e., have higher CCV) than in FACE or SELF. Also, older adults differentiate more (i.e., higher CCV) than young adults. In model 2, an interaction term of age × method group was included in the equation, revealing a significant R^2 change of .001 (p < .001). Significant interaction effects were observed for age with method groups, both for the CATI versus SELF dummy (β = .08, p < .001) and for the CATI versus FACE dummy (β = .11, p < .001) on construct variability. Figure 1 illustrates this finding with a fitted regression line to plot age × method conditions on CCV. While the construct differentiation was only slightly associated with age in the FACE condition (β = .06, p < .001) and in the SELF condition (β = .06, p < .001), the largest association between age and CCV was found in the CATI condition (β = .20, p < .001). Figure 1 suggests that older adults deviate from the expected convergence of the personality scale in their responses when participating in a telephone-interviewing procedure. All effects remained stable after controlling for the five personality manifest scale values.

Five-year rank order consistency of the BFI–S In 2009, a total of 13,459 participants completed the SOEP Big Five

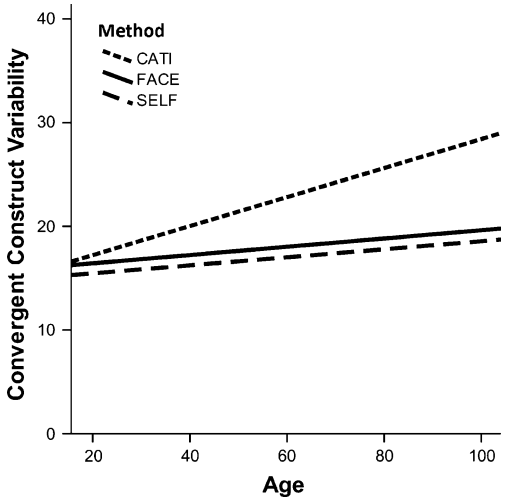

Fig. 1 Construct variability increases with age in computer-assisted telephone interviewing. CATI, computer-assisted telephone interviewing; FACE, face-to-face interviewing; SELF, self-administered questionnaire completion

Inventory (BFI–S) after 5 years either in the same or in a different assessment condition as in 2005. In order to test for possible method effects on the 5-year rank order consistency, we compared retest stability coefficients across age groups and the following four groups of assessment conditions: (1) FACE assessment in 2005 and 2009, (2) SELF assessment in 2005 and 2009, and (3) different assessment conditions in 2005 and 2009 (CHANGE group).

A comparison of test–retest stability coefficients necessarily requires proof of invariant factor loadings (i.e., weak measurement invariance) across the two measurement occasions. We included correlated uniquenesses for the same item at measurement points one and two in our models to avoid inflated test–retest stability coefficients. (for a detailed description of correlated uniquenesses across assessment conditions by age groups, see the supplementary information material, Table SM14). Comparable to the cross-sectional analyses above, we observed close fit for the configural invariance models between T1 and T2 in all three age groups (young adults, chi-square/*df* = 1,849/795, CFI = .967, TLI = .945, RMSEA = .031; middle-aged adults, chi-square/*df* = 1,919/795, CFI = .972, TLI = .954, RMSEA = .028; old adults, chi-square/*df* = 1,704/795, CFI = .965, TLI = .942, RMSEA = .030), as well as for the weak measurement invariance model (young, chi-square/*df* = 2,140/1045, CFI = .965, TLI= .957, RMSEA = .027; middle, chi-square/*df* = 2,285/1045, CFI = .969, TLI = .961, RMSEA = .025; old, chi-square/*df* = 1,996/1045, CFI = .963, TLI = .954, RMSEA = .027). The weak invariance model showed improved parameters, as compared with the configural model (i.e., CFI change < .01; young, .967 vs. .965; middle, .972 vs. 969; old, .965 vs. .963; TLI, young, .945 vs. .957; middle, .954 vs. .961; old, .942 vs. .954; RMSEA, young, .031 vs. .027; middle, .028 vs. .025; old, .030 vs. .027). In conclusion, factor loadings proved invariant within methods across the 5-year interval. This finding justified a comparison of latent test–retest stability coefficients (i.e., bivariate correlations of latent factor scores at T1 and T2), as is shown in Table 7.

In order to clarify the invariance of test–retest stability coefficients across method conditions, we specified the following regression model. All convergent regression paths from factor scores at T1 to factor scores at T2 (e.g., *N* at T2 on *N* at T1) were fixed across method groups. Divergent regression paths from factor scores at T1 to factor scores at T2 (e.g., *N* at T2 on E at T1) were freed up. In comparison with a fully saturated model, the model results revealed no acceptable model fit for young adults (chi-square/*df* = 286/10, CFI = .987, TLI = .862, RMSEA = .140), middle-aged adults (chi-square/*df* = 326/10, CFI = .991, TLI = .902, RMSEA = .131), or older adults (chi-square/*df* = 95/10, CFI = .996, TLI = .959, RMSEA = .083). Therefore, in all age groups, methods differed with regard to test–retest stability. With a few exceptions (e.g., late adulthood, conscientiousness), self-administered personality assessment is associated with a stronger test–retest stability, as compared with personality interviewing methods.

Table 7 Five-year latent factor score test–retest stability coefficients

	Young Adults (*N* = 4,232)					Middle-Aged Adults (*N* = 5,503)					Older Adults (*N* = 3,724)				
	N	E	O	A	C	N	E	O	A	C	N	E	O	A	C
All conditions	.808	.807	.715	.751	.699	.839	.872	.750	.851	.703	.795	.790	.734	.740	.659
Method change	.723	.758	.689	.745	.645	.782	.844	.712	.778	.598	.800	.701	.687	.702	.562
FACE	.810	.714	.676	.720	.662	.787	.850	.703	.835	.702	.769	.766	.736	.711	.699
SELF	.838	.881	.780	.789	.763	.919	.907	.837	.897	.744	.880	.905	.749	.850	.591

Young adults: Change, *N* = 709; FACE, *N* = 1,720; SELF, *N* = 1,803. Middle-aged adults: Change, *N* = 763; FACE, *N* = 2,589; SELF, *N* = 2,151. Older adults: Change, *N* = 351; FACE, *N* = 2,459; SELF, *N* = 914

N Neuroticism, *E* Extraversion, *O* Openness, *A* Agreeableness, *C* Conscientiousness

Discussion

In our research, we generated new evidence that short self-report scales of the Big Five prove fairly robust across adulthood from 18 to 90 years and across different contexts of assessment procedures, such as self-administration or face-to-face interviewing. In all these contexts, the Big Five factorial structure and the 5-year rank-order consistency of the scales proved robust. Even in the context of telephone-interviewing studies, we observed fairly reliable factor structures, with the exception of telephone interviews with older adults. In the following, we discuss in detail the insights that can be gained from this kind of research.

First, we were able to replicate a solid and robust five-factor structure based on an internationally used short Big Five Inventory (BFI–S) in two large probability samples for two methods and in one sample across a 5-year time interval. We relied on a 15-item Big Five inventory that was developed for the large-scale German SOEP study (Gerlitz & Schupp, 2005) and that has generated much validating evidence in several research publications across disciplines (Donnellan & Lucas, 2008; Headey et al., 2010; Winkelmann & Winkelmann, 2008). An equivalent Big Five personality inventory was implemented in the BHPS (Taylor et al., 2009), and a similar personality inventory was included in the HILDA study (Lucas & Donnellan, 2009). Obviously, implementation of such short personality measures of the Big Five personality trait dimensions promises new insights about psychological processes related to personality development across disciplines all over the world (Donnellan et al., 2006; Donnellan & Lucas, 2008; Gosling et al., 2003; Lucas & Donnellan, 2009; Marsh et al., 2006; Rentfrow, 2010). It thus seems of critical relevance to understand whether such universal implementation of a standard psychological measure generates robust and reliable findings that are comparable across broad age ranges, and even when assessed with different assessment procedures and in diverse contexts. Our findings serve to illustrate that there is good reason to assume such robustness when assessing short self-report measures of personality trait characteristics. On the basis of our findings, there is reason to assume that results of cross-cultural comparisons between national survey such as the HILDA study (Australia), the BHPS, and the German SOEP may not be distorted due to use of differential assessment procedures within and across these national studies.

Second, we present the first evidence that the short self-report personality constructs prove robust across a broad range of different assessment methods (computer assisted, paper and pencil, telephone, self-administration). That is, irrespective of the context of interviewing, we observed a widely invariant factor structure of the Big Five model, with equivalent factor loadings and correlations, as well as invariant 5-year stability coefficients. We confirmed such an indication of robust measurement invariance with respect to factor means (with one exception), factor loadings, item intercepts, and item uniqueness. According to our findings, the robustness of a widely applied short Big Five Inventory (BFS–S) generalizes across face-to-face interviewing and self-administrative procedures but does not generalize to phone surveys. There is caution warranted, however, when comparing mean differences across different assessment contexts.

For example, we observed method effects on mean levels for the factor of openness to experience. In the telephone interview, respondents described themselves as being more open than in personal-interviewing or self-administered questionnaire contexts. This finding points to the possibility that assessment contexts (e.g., speaking to an unknown person on the phone) may well generate different self-descriptions that are context specific. We do not have a good explanation for this finding. It may be that there is a potential bias related to social desirability for phone survey studies similar to results found in personal-interviewing studies (Marshall et al., 2005). For example, McHorney and colleagues (1994) compared strategies of correcting for bias due to effects of desirability or reactivity. A critical question is whether a general recommendation should be made to always include explicit and detailed assessments of contextual information (e.g., presence of others in the same room, distracting factors) when conducting telephone interviews. This appears particularly relevant because the use of telephone interviewing has increased in psychological research in recent years (Bauer et al., 2004; Blickle, Kramer, & Mierke, 2010; Cercone et al., 2009; Tun & Lachman, 2008).

A related finding pertains to the generally robust cross-method stability of the Big Five personality trait dimensions across a 5-year time interval. In a longitudinal comparison, we demonstrated the measurement invariance of the Big Five personality factor structure across two measurement occasions for the face-to-face interviewing and for the self-administrative conditions of assessment. Robustness persisted with a few exceptions: Retest stability coefficients were lowest when respondents participated in different method conditions at time 1 and time 2. This finding is important in several ways. It clearly shows that the self-report of personality characteristics also reflects situational influences, particularly in social situations such as a face-to-face interview. Thus, the stability coefficients that result in the method change assessment condition point to a transsituational consistency of the short Big Five inventory (cf. Mischel & Shoda, 1995). Considering the different social situations of face-to-face interviews versus self-administrated questionnaires, we submit that such stability coefficients may reflect a reliable estimation of

valid rank order consistency of personality traits across 5 years. By contrast, assessing personality stability in self-administrative assessment procedures may well overestimate the true stability of such self-report ratings. The finding serves to emphasize our conclusion that there is reliable robustness of self-report personality measures in survey research with face-to-face interviewing methods and self-administration methods. However, retest stability coefficients of the five personality constructs also reflect possible reliability costs of face-to-face interviewing. More precisely, when a questionnaire was completed in a self-administered context, we generally observed a stronger rank-order consistency across 5 years, as compared with other method conditions.

Third, these findings demonstrate strong invariance of the robust factor structure across young, middle, and old adulthood, with the exception of telephone-interviewing procedures. In the last case, findings suggest that unreliability may be associated with a higher mental workload and situational ambiguity of the telephone context among older adults.

In addition, we cannot preclude that the higher openness-to-experience scores among participants who are willing to respond to a telephone survey reflects a trait-specific sampling bias (e.g., related to willingness to answer questions on the phone). This suggestion is somewhat corroborated by the finding that mean levels of openness to experience were higher among highly educated older participants than among younger respondents in the telephone condition. Valuing new ideas and seeing oneself as having an active imagination may imply a greater capacity to generate reliable self-ratings on the phone. The findings of this research suggest that the use of self-report measures in telephone interviews allows for general robust personality assessment in early and middle adulthood. However, the common use of the telephone-interviewing method is associated with an increased risk of generating inconsistent response patterns among older adults. Participating in an interview on the phone implies that one attends and responds to the interviewer's questions without being able to read the items that are rated. This task appears to be particularly challenging for less well-educated older adults, who may not be as well trained with such self-descriptive ratings of personality characteristics. Again, this interpretation of our finding is supported by the finding that the convergent construct variability was higher among older participants in the telephone study, as compared with the other assessment conditions. This finding suggests that a lack of consistency and reliability may be a result of the cognitive challenges associated with the telephone context. Clearly, the present evidence is not sufficient and warrants further studies to elucidate in greater depth the cognitive demands and constraints of telephone interviewing.

There are also a couple of caveats that ought to be considered when interpreting the findings of this research, particularly with respect to the issues of sample selectivity and self-report bias.

Selectivity Within the SOEP, assignment to either self-administered questionnaire completion or face-to-face personal-interviewing procedures (FACE: PAPI or CAPI) was dependent upon the availability of respondents at the time the interviewer visited the household. Consequently, the characteristics of respondents in the different assessment conditions differed as a function of employment status, education, or traditional roles in household sharing. However, differences were relatively small, except for the SELF and the CATI contexts. where low-educated participants were underrepresented. We submit that the selectivity across the different assessment conditions in this study hampered the likelihood of identifying factorial invariance. We argue, therefore, that our comparison of method conditions within a real-life survey provides a reliable test. For example, there is some evidence suggesting that lower education is associated with greater acquiescence, resulting in a less reliable and robust structure of the Big Five constructs (Rammstedt et al., 2010). However, we did not find evidence that there were substantial structural differences in the Big Five personality factor loadings and variances in the FACE condition. For this reason, we decided against correcting for acquiescence in our data.

Self-report bias The present research and analyses were based exclusively on self-report data, which is known to be generally susceptible to distortion related to the wording, format, and context of research instruments (Schwarz, 1999). However, issues related to the features of the 15-item Big Five Inventory instrument were not addressed in this research. Therefore, our findings do not allow the conclusions that self-report measures of Big Five personality constructs are robust against the formatting, style, or wording of items or scales. For example, in this research, the same 7-point scale for self-report ratings of agreement to items was used across all method groups. Thus, we cannot preclude that in the CATI assessment condition, for example, a 5-point rating scale may have allowed us to obtain a more robust pattern of the five-factor model, even among less-educated older adults.

In sum, findings are relevant to psychological research that relies on national and cross-national surveys with age-heterogeneous samples. In recent years, interdisciplinary efforts have increased the implementation of psychological constructs in large-scale survey studies. The current brief measure of the Big Five personality domains is implemented in widely used international ongoing longitudinal

panel studies in Great Britain (BHPS), Australia (HILDA), and Germany (SOEP) and has already generated a wealth of new findings on the interplay of individual differences in personality trait domains with other variables (e.g., Lucas & Donnellan, 2009; Marsh et al., 2010; Marsh et al., 2011). The present study, for the first time, reports evidence for measurement equivalence and 5-year retest stability of this instrument across a range of different assessment contexts. We conclude that the short Big Five Inventory (BFI–S) delivers quick and rough, but robust and reliable estimations of the Big Five personality constructs. However, our findings also suggest that this brief assessment may not be well suited for phone surveys that include older adults.

Author Note Oliver Lüdtke is now at Humboldt University of Berlin, Germany. We are most thankful to Jane Thompson for invaluable comments and proofreading.

Open Access This article is distributed under the terms of the Creative Commons Attribution Noncommercial License which permits any noncommercial use, distribution, and reproduction in any medium, provided the original author(s) and source are credited.

References

Allemand, M., Zimprich, D., & Hendriks, A. A. J. (2008). Age differences in five personality domains across the life span. *Developmental Psychology, 44*, 758–770. doi:10.1037/0012-1649.44.3.758

Allemand, M., Zimprich, D., & Hertzog, C. (2007). Cross-sectional age differences and longitudinal age changes of personality in middle adulthood and old age. *Journal of Personality, 75*, 323–358. doi:10.1111/j.1467-6494.2006.00441.x

Allport, G. W., & Odbert, H. S. (1936). Trait names: A psycholexical study. *Psychological Monographs, 47*(1), Whole No. 211.

Aluja, A., García, O., García, L. F., & Seisdedos, N. (2005). Invariance of the "NEO-PI-R" factor structure across exploratory and confirmatory factor analyses. *Personality and Individual Differences, 38*, 1879–1889. doi:j.jrp.2006.02.001/j.paid.2004.11.014

Asparouhov, T., & Muthén, B. (2009). Exploratory structural equation modeling. *Structural Equation Modeling, 16*, 397–438. doi:10.1080/10705510903008204

Baltes, P. B., Reese, H. W., & Nesselroade, J. R. (1977). *Life-span developmental psychology: Introduction to research methods*. Monterey: Brooks Cole.

Bauer, T. N., Truxillo, D. M., Paronto, M. E., Weekley, J. A., & Campion, M. A. (2004). Applicant reactions to different selection technology: Face-to-face, interactive voice response, and computer-assisted telephone screening interviews. *International Journal of Selection and Assessment, 12*, 135–148. doi:10.1111/j.0965-075X.2004.00269.x

Benet-Martinez, V., & John, O. P. (1998). Los Cinco Grandes across cultures and ethnic groups: Multitrait–multimethod analyses of the Big Five in Spanish and English. *Journal of Personality and Social Psychology, 75*, 729–750. doi:10.1037/0022-3514.75.3.729

Bentler, P. M. (1990). Comparative fit indices in structural models. *Psychological Bulletin, 107*, 238–246. doi:10.1037/0033-2909.107.2.238

Blickle, G., Kramer, J., & Mierke, J. (2010). Telephone-administered intelligence testing for research in work and organizational psychology. *European Journal of Psychological Assessment, 26*, 154–161. doi:10.1027/1015-5759/a000022

Borghans, L., Duckworth, A. L., Heckman, J. J., & ter Weel, B. (2008). The economics and psychology of personality traits. *Journal of Human Resources, 43*, 972–1059.

Borkenau, P., & Ostendorf, F. (1990). Comparing exploratory and confirmatory factor analysis: A study on the 5-factor model of personality. *Personality and Individual Differences, 11*, 515–524. doi:j.jrp.2006.02.001/0191-8869(90)90065-Y

Brannick, M. T. (1995). Critical comments on applying covariance structure modeling. *Journal of Organizational Behavior, 16*, 201–213. doi:10.1002/job.4030160303

Cattell, R. B. (1946). *Description and measurement of personality*. Yonkers-on-Hudson: World Book.

Cercone, J. J., Danielson, C. K., Ruggiero, K. J., & Kilpatrick, D. G. (2009). Telephone surveys of traumatic experiences and other sensitive topics. In D. Buchanan, C. B. Fisher, & L. Gable (Eds.), *Research with high-risk populations: Balancing science, ethics, and law* (pp. 77–92). Washington: American Psychological Association.

Chen, F. F. (2007). Sensitivity of goodness of fit indexes to lack of measurement invariance. *Structural Equation Modeling, 14*, 464–504. doi:10.1080/10705510701301834

Cheung, G. W., & Rensvold, R. B. (2001). The effects of model parsimony and sampling error on the fit of structural equation models. *Organizational Research Methods, 4*, 236–264. doi:10.1177/109442810143004

Church, A. T., & Burke, P. J. (1994). Exploratory and confirmatory tests of the Big Five and Tellegen's three- and four-dimensional models. *Journal of Personality and Social Psychology, 66*, 93–114. doi:10.1037/0022-3514.66.1.93

Costa, P. T., Jr., & McCrae, R. R. (1995). Domains and facets: Hierarchical personality assessment using the Revised NEO Personality Inventory. *Journal of Personality Assessment, 64*, 21–50. doi:10.1207/s15327752jpa6401_2

Davis, J. A., & Smith, T. W. (1992). *The NORC General Social Survey: A user's guide*. Newbury Park: Sage.

Donnellan, M. B., & Lucas, R. E. (2008). Age differences in the Big Five across the life span: Evidence from two national samples. *Psychology and Aging, 23*, 558–566. doi:10.1037/a0012897

Donnellan, M. B., Oswald, F. L., Baird, B. M., & Lucas, R. E. (2006). The Mini-IPIP Scales: Tiny-yet-effective measures of the *Big Five* factors of personality. *Psychological Assessment, 18*, 192–203. doi:10.1037/1040-3590.18.2.192

Eysenck, H. (1947). *Dimensions of personality*. London: Routledge & Kegan Paul.

Gagné, P., & Hancock, G. R. (2006). Measurement model quality, sample size, and solution propriety in confirmatory factor models. *Multivariate Behavioral Research, 41*, 65–83. doi:10.1207/s15327906mbr4101_5

Gerlitz, Y., & Schupp, J. (2005). *Zur Erhebung der Big-Five-basierten Persönlichkeitsmerkmale im SOEP [Assessment of Big Five personality characteristics in the SOEP]. German Institute of Economic Research (Research Notes 4)*. Berlin: DIW.

Goldberg, L. R. (1990). An alternative "description of personality": The Big-Five factor structure. *Journal of Personality and Social Psychology, 59*, 1216–1229. doi:10.1037/0022-3514.59.6.1216

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. doi:j.jrp.2006.02.001/S0092-6566(03)00046-1

Groves, R. M., Fowler, F. J., Jr., Coupler, M. P., Lepkowski, J. M., Singer, E., & Tourangeau, R. (2009). *Survey methodology* (2nd ed.). Hoboken: Wiley.

Groves, R. M., & Mathiowetz, N. A. (1984). Computer assisted telephone interviewing: Effects on interviewers and respondents. *Public Opinion Quarterly, 48*, 356–369. doi:10.1086/268831.

Hansell, S., Sparacino, J., Ronchi, D., & Strodtbeck, F. L. (1984). Ego-development responses in written questionnaires and telephone interviews. *Journal of Personality and Social Psychology, 47*, 1118–1128. doi:10.1037/0022-3514.47.5.1118

Headey, B., Muffels, R., & Wagner, G. G. (2010). Long-running German panel survey shows that personal and economic choices, not just genes, matter for happiness. *Proceedings of the National Academy of Sciences, 107*, 17922–17926. doi:10.1073/pnas.1008612107

Heineck, G., & Anger, S. (2010). The returns to cognitive abilities and personality traits in Germany. *Labour Economics, 17*, 535–546. doi:j.jrp.2006.02.001/j.labeco.2009.06.001

Herman, J. B. (1977). Mixed-mode data collection: Telephone and personal interviewing. *Journal of Applied Psychology, 62*, 399–404. doi:10.1037/0021-9010.62.4.399

Holbrook, A. L., Green, M. C., & Krosnick, J. A. (2003). Telephone versus face-to-face interviewing of national probability samples with long questionnaires: Comparisons of respondent satisficing and social desirability response bias. *Public Opinion Quarterly, 67*, 79–125. doi:10.1086/346010

Holden, R. R., & Troister, T. (2009). Developments in the self-report assessment of personality and psychopathology in adults. *Canadian Psychology, 50*, 120–130. doi:10.1037/a0015959

Hopwood, C. J., & Donnellan, M. B. (2010). How should the internal structure of personality inventories be evaluated? *Personality and Social Psychology Review, 14*, 332–346. doi:10.1177/1088868310361240

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* (Vol. 2, pp. 102–138). New York: Guilford.

Klages, L. (1926). *Die Grundlagen der Charakterkunde [Foundations of characterology]*. Leipzig: Bart.

Lang, F. R. (2005). *Erfassung des kognitiven Leistungspotenzials und der "Big Five" mit Computer-Assisted-Personal-Interviewing (CAPI): Zur Reliabilität und Validität zweier ultrakurzer Tests und des BFI-S* [Assessment of cognitive capabilities and the Big Five with Computer-Assisted Personal Interviewing (CAPI): Reliability and validity]. German Institute of Economic Research (Research Notes 9/2005), Berlin: DIW.

Lang, F. R., Baltes, P. B., & Wagner, G. G. (2007). Desired lifetime and end-of-life desires across adulthood from 20 to 90: A dual-source information model. *Journal of Gerontology: Psychological Science, 62*, 268–276.

Lang, F. R., Lüdtke, O., & Asendorpf, J. (2001). Testgüte und psychometrische Äquivalenz der deutschen Version des Big Five Inventory (BFI) bei jungen, mittelalten und alten Erwachsenen [Validity and psychometric equivalence of the German version of the Big Five Inventory in young, middle-aged and old adults]. *Diagnostica, 47*, 111–121. doi:10.1026//0012-1924.47.3.111

Lucas, R. E., & Donnellan, M. B. (2009). Age differences in personality: Evidence from a nationally representative Australian sample. *Developmental Psychology, 45*, 1353–1363. doi:10.1037/a0013914

Marsh, H. W., Hau, K.-T., Balla, J. R., & Grayson, D. (1998). Is more ever too much? The number of indicators per factor in confirmatory factor analysis. *Multivariate Behavioral Research, 33*, 181–220. doi:10.1207/s15327906mbr3302_1

Marsh, H. W., Hau, K.-T., & Wen, Z. (2004). In search of golden rules: Comment on hypothesis testing approaches to setting cutoff values for fit indexes and dangers in overgeneralising Hu & Bentler's (1999) findings. *Structural Equation Modeling, 11*, 320–341. doi:10.1207/s15328007sem1103_2

Marsh, H. W., Lüdtke, O., Muthén, B., Asparouhov, T., Morin, A. J. S., Trautwein, U., et al. (2010). A new look at the big-five factor structure through exploratory structural equation modeling. *Psychological Assessment, 22*, 471–491. doi:10.1037/a0019227

Marsh, H. W., Muthén, B., Asparouhov, T., Lüdtke, O., Robitzsch, A., Morin, A. J. S., et al. (2009). Exploratory structural equation modeling, integrating CFA and EFA: Applications to students' evaluations of university teaching. *Structural Equation Modeling, 16*, 439–476. doi:10.1080/10705510903008220

Marsh, H. W., Nagengast, B., & Morin, A. J. S. (2011). *Measurement invariance of big-five factors over the lifespan: ESEM tests of gender, age, plasticity, maturity, and La Dolce Vita effects*. Manuscript under review.

Marsh, H. W., Trautwein, U., Lüdtke, O., Köller, O., & Baumert, J. (2006). Integration of multidimensional self-concept and core personality constructs: Construct validation and relations to well-being and achievement. *Journal of Personality, 74*, 403–455. doi:10.1111/j.1467-6494.2005.00380.x

Marshall, M. B., De Fruyt, F., Rolland, J.-P., & Bagby, R. M. (2005). Socially desirable responding and the factorial stability of the NEO PI-R. *Psychological Assessment, 17*, 379–384. doi:10.1037/1040-3590.17.3.379

McCrae, R. R., & Costa, P. T., Jr. (1997). Personality trait structure as a human universal. *American Psychologist, 52*, 509–516. doi:10.1037/0003-066X.52.5.509

McCrae, R. R., Zonderman, A. B., Costa, P. T., Jr., Bond, M. H., & Paunonen, S. V. (1996). Evaluating replicability of factors in the Revised NEO Personality Inventory: Confirmatory factor analysis versus Procrustes rotation. *Journal of Personality and Social Psychology, 70*, 552–566. doi:10.1037/0022-3514.70.3.552

McHorney, C. A., Kosinski, M., & Ware, J. E., Jr. (1994). Comparisons of the costs and quality of norms for the SF-36 health survey collected by mail versus telephone interview: Results from a national survey. *Medical Care, 32*, 551–567. doi:10.1097/00005650-199406000-00002

Meredith, W. (1993). Measurement invariance, factor analysis and factorial invariance. *Psychometrika, 58*, 525–543. doi:10.1007/BF02294825

Mischel, W., & Shoda, Y. (1995). A cognitive-affective system theory of personality: Reconceptualizing situations, dispositions dynamics, and invariance in personality structure. *Psychological Review, 102*, 246–268. doi:10.1037/0033-295X.102.2.246

Muthén, L. K., & Muthén, B. O. (2008). *Mplus user's guide* (5th ed.). Los Angeles: Muthén & Muthén.

Rammstedt, B., Goldberg, L. R., & Borg, I. (2010). The measurement equivalence of the Big-Five factor markers for persons with different levels of education. *Journal of Research in Personality, 44*, 53–61. doi:j.jrp.2006.02.001/j.jrp.2009.10.005

Rammstedt, B., Holzinger, B., & Rammsayer, T. H. (2004). Zur Äquivalenz der Papier-Bleistift- und einer computergestützten Version des NEO-Fünf- Faktoren-Inventars (NEO-FFI) [Equivalence of a paper-pencil- and a computer-based version of the NEO-FFI]. *Diagnostica, 50*, 88–97. doi:10.1026/0012-1924.50.2.88

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. doi:j.jrp.2006.02.001/j.jrp.2006.02.001

Rentfrow, P. J. (2010). Statewide differences in personality. *Toward a psychological geography of the United States. American Psychologist, 65*, 548–558. doi:10.1037/a0018194

Richman, W. L., Kiesler, S., Weisband, S., & Drasgow, F. (1999). A meta-analytic study of social desirability distortion in computer-administered questionnaires, traditional questionnaires, and interviews. *Journal of Applied Psychology, 84*, 754–775. doi:10.1037/0021-9010.84.5.754

Roberts, B. W., Walton, K. E., & Viechtbauer, W. (2006). Patterns of mean-level change in personality traits across the life course: A meta-analysis of longitudinal studies. *Psychological Bulletin, 132*, 1–25. doi:10.1037/0033-2909.132.1.1

Schräpler, J. P., Schupp, J., & Wagner, G. G. (2010). Changing from PAPI to CAPI: Introducing CAPI in a longitudinal study. *Journal of Official Statistics, 26*, 233–269.

Schwarz, N. (1999). Self-reports: How the questions shape the answers. *American Psychologist, 54*, 93–105. doi:10.1037/0003-066X.54.2.93

Siedler, T., Schupp, J., & Wagner, G. G. (2011). Innovative methods within in the context of archival data: Examples from household panel surveys. In K. H. Trzesniewski, M. B. Donnellan, & R. E. Lucas (Eds.), *Secondary data analysis* (pp. 103–118). Washington: American Psychological Association.

Smith, T. W., & Williams, P. G. (1992). Personality and health: Advantages and limitations of the five-factor model. *Journal of Personality, 60*, 395–423. doi:10.1111/j.1467-6494.1992.tb00978.x

Stoop, I., Billiet, J., Koch, A., & Fitzgerald, R. (2010). *Improving survey response: Lessons learned from the European Social Survey*. Chichester: Wiley.

Sutin, A. R., Costa, P. T., Miech, R., & Eaton, W. W. (2009). Personality and career success: Concurrent and longitudinal relations. *European Journal of Personality, 23*, 71–84. doi:10.1002/per.704

Swanberg, A. B., & Martinsen, O. L. (2010). Personality, approaches to learning and achievement. *Educational Psychology, 30*, 75–88. doi:10.1080/01443410903410474

Taylor, M. F., Brice, J., Buck, N., & Prentice-Lane, E. (2009). *British household panel survey user manual: Vol. A. Introduction, technical report and appendices*. Colchester: University of Essex.

Tucker, L. R., & Lewis, C. (1973). A reliability coefficient for maximum likelihood factor analysis. *Psychometrika, 38*, 1–10. doi:10.1007/BF02291170

Tun, P. A., & Lachman, M. E. (2008). Age differences in reaction time and attention in a National Telephone sample of adults: Education, sex, and task complexity matter. *Developmental Psychology, 44*, 1421–1429. doi:10.1037/a0012845

Vassend, O., & Skrondal, A. (1997). Validation of the NEO Personality Inventory and the five-factor model: Can findings from exploratory and confirmatory factor analysis be reconciled? *European Journal of Personality, 11*, 147–166. doi:10.1002/(SICI)1099-0984(199706)11:2<147::AID-PER278>3.0.CO;2-E

Wagner, G. G., Frick, J. R., & Schupp, J. (2007). The German Socio-Economic Panel Study (SOEP): Scope, evolution and enhancements. *Schmollers Jahrbuch, 1*, 139–170.

Winkelmann, L., & Winkelmann, R. (2008). Personality, work, and satisfaction: Evidence from the German Socio-Economic Panel. *Journal of Positive Psychology, 3*, 266–275. doi:10.1080/17439760802399232