Sociodemographic Effects on the Test-Retest Reliability of the Big Five Inventory

Timo Gnambs

Osnabrück University

Author Note

Timo Gnambs, Institute of Psychology, Osnabrück University, Germany.

Correspondence concerning this article should be addressed to Timo Gnambs, Institute of Psychology, Osnabrück University, Seminarstr. 20, 49069 Osnabrück, Germany. E-mail: timo.gnambs@uni-osnabrueck.de

Article type: Brief report

Word count: 2,388

Accepted for publication in the *European Journal of Psychological Assessment*.

Abstract

Psychometric properties of self-report scales can be affected by sociodemographic differences among respondents. For example, factor structures of established personality scales tend to be distorted in samples with less formal education. Whether test-retest reliabilities are comparably affected is of yet not well known. Therefore, this study examined the measurement precision of a short Big Five instrument in a diverse sample of the German population. 1,696 (50% women) participants reported on their personality twice within an interval of 10 weeks. The median test-retest reliability for the five traits, $r_{\text{tt}} = .66$, was notably smaller than previously reported coefficients from college students, median $\underline{r}_{\text{tt}} = .78$. Moderator analyses identified modest effects of age and educational attainment on these reliability estimates, whereas sex showed no differential impact. These results highlight that test-retest reliabilities derived from student samples should not be generalized to sociodemographically diverse groups of respondents.

Keywords: test-retest reliability, Big Five, measurement error, education, age

Sociodemographic Effects on the Test-Retest Reliability of the Big Five Inventory

An increasing number of large-scale population surveys include measures of the Big Five of personality (cf. Lang, John, Lüdtke, Schupp, & Wagner, 2011; Rammstedt, Kemper, & Borg, 2013). To examine personality effects across diverse sociodemographic groups using these scales they need to reliably assess the constructs in the focal population. However, most personality inventories are constructed using convenient samples that are typically dominated by college students; their psychometric properties are rarely investigated in sociodemographically diverse populations (for notable exceptions see Rammstedt, 2007, or Sutin, Costa, Evans, & Zonderman, 2013). This is rather unfortunate since respondent characteristics might affect the psychometric properties of self-report scales. For example, recent research highlighted that differences in respondents' literacy and educational attainment distort the factor structure of seemingly well-validated Big Five instruments (cf. Rammstedt et al., 2013; Sutin et al., 2013). So far, respective analyses for other key properties of measurement scales are scarce.

Observed scores are typically confounded with measurement error and do not represent pure indicators of the construct of interest. Therefore, their measurement precision is routinely evaluated using various coefficients of reliability. In most cases, reliability is quantified by indices of internal consistency (e.g., coefficient alpha) that reflects the consistency between different item responses (Schmidt, Le, & Ilies, 2003). Previous research suggested that internal consistencies of Big Five measures seem to be rather invariant across age, sex, and education levels (e.g., Löckenhoff et al., 2008). However, measurement error also includes other facets that are not reflected by internal consistency measures (see Gnambs, 2014a, or Schmidt et al., 2003, for overviews). Particularly, test-retest reliabilities that quantify occasion-specific, transient measurement error received increased attention in recent years because of their relevance for criterion validities, more so than internal consistencies (McCrae, Kurtz, Yamagata, &

Terracciano, 2011). Whereas internal consistency is sometimes also evaluated in norm samples for different subgroups of respondents, for example stratified by sex or age, comparably analyses are seldom reported for test-retest reliabilities. Rather, the latter are typically examined (if at all) in small convenience samples. A recent meta-analyses of test-retest reliabilities for measures of the Big Five (Gnambs, 2014b) indicated that the average retest sample is young, primarily female, and highly educated (see Table 1). Thus, if differences in test-retest reliabilities between sociodemographic groups exist, generalizations of instruments' retest reliabilities based on college students seem suspect at best. Therefore, the present study examines the invariance of test-retest reliabilities across sex, age, and educational levels for a short measure of the Big Five in a diverse sample of the German population.

Method

Participants and Procedure

The study draws on members of the German longitudinal election study (Rattinger, Roßteuscher, Schmitt-Beck, & Weßels, 2013) that observed political attitudes and behaviors of the German electorate in 2009. The study was implemented as a bi-weekly web-survey with seven waves. The present analyses are limited to the first and sixth waves that included the focal personality measures. The average interval between the two assessment occasions was about 10 weeks. Originally, a quota sample stratified by age, gender, and education was drawn from a non-probability online panel. Similar to most web-based samples (cf. Bosnjak et al., 2014) older individuals and people with lower education were slightly underrepresented among the respondents as compared to the German Microcensus (Federal Statistical Office, 2010). However, the sample was significantly more diverse than the typical retest sample in psychological research (see Table 1). The total sample size for the present analyses was *N* = 1,696 (50% women). The age ranged from 18 to 80 years (*M* = 42.06, *SD* = 14.39).

Measures

Big Five. The five basic traits of personality—openness, conscientiousness, extraversion, agreeableness, and neuroticism—were assessed with a short version of the Big Five Inventory (BFI-10; Rammstedt & John, 2007) that measures each trait with two items, one keyed in the positive direction and one in the negative direction. Each item was accompanied by a 5-point response scale from *fully disagree* (1) to *fully agree* (5). Trait scores for each respondent were calculated as item means, after inverting the negatively poled items. Means and standard deviations for each scale are summarized in Table 2.

Education. Educational attainment was assessed by splitting the sample into two groups including either individuals with higher secondary education (i.e. people with an entrance qualification for universities; $N = 509$) or those without ($N = 1,187$). Thus, higher education corresponded roughly to level 4 and higher of the International Standard Classification of Education (Schneider, 2008).

Results

The test-retest reliabilities for the five traits were $r_{tt} = .67$, 95% CI [.65, .70] for openness, $r_{tt} = .65$, 95% CI [.62, .67] for conscientiousness, $r_{tt} = .74$, 95% CI [.71, .76] for extraversion, $r_{tt} = .55$, 95% CI [.51, .58] for agreeableness, and $r_{tt} = .66$, 95% CI [.63, .68] for neuroticism. Overall, these reliabilities were notably smaller than those derived previously among students (Rammstedt & John, 2007), .78, .83, .87, .66, and .71, respectively. However, only the reliabilities for conscientiousness, $z = 3.02$, $p < .001$, and extraversion, $z = 2.82$, $p < .001$, were significantly different in the two samples. To examine the effects of sociodemographic differences on these reliability estimates, the retest correlations were reparameterized in form of linear regressions. Thus, for each trait the z-standardized trait score at the second measurement occasion was regressed on the respective z-standardized trait score at the first measurement occasion. The regression weights in these models are equivalent to the test-retest correlation. The impact of three moderators was investigated by adding interactions with sex (coded -1 for men and 1 for women), educational level (coded -1 for lower education and 1 for higher education), and linear as well as quadratic age trends (*z*-standardized) to the regression models. These analyses yielded three main findings (see Table 2): First, test-retest reliabilities for three traits were subject to a modest age trend. For extraversion and neuroticism scales reliabilities were somewhat lower for the youngest and oldest respondents, whereas openness scales yielded the largest reliabilities among young people. This effect is plotted in Figure 1 for different age groups: 18 to 30 years (*N* = 436), 31 to 40 years (*N* = 357), 41 to 50 years (*N* = 393), 51 to 60 years (*N* = 293), and 61 to 80 years (*N* = 217). Second, sex had no differential effect on test-retest reliabilities. Third, highly educated people generated more reliable scores than those with lower education (see Figure 2). This effect was most pronounced for openness, agreeableness, and neuroticism scales.

Discussion

Sociodemographice factors of respondents can distort key properties of measurement instruments (Rammstedt et al., 2013; Sutin et al., 2013).). Therefore, this study investigated the test-retest reliability of a short Big Five instrument in a diverse sample of the German population. In contrast to typical retest studies that, for the largest part, relied on college students (cf. Gnambs, 2014b) the current investigation also acknowledged respondents that are frequently neglected in psychological research, that is, individuals with lower formal education and people of higher ages (e.g., above 60 years). The presented results clearly showed that sample characteristics yielded notable effects on the reliability estimates. Whereas sex had a negligible impact on test-retest reliabilities, age and educational attainment showed more consistent effects:

Well-educated individuals (i.e. those with at least university-entrance qualifications) consistently generated more reliable scores for the five traits than those with less formal education. Albeit, the respective effect was somewhat modest and resulted in reliability differences between $\Delta r_{\mathrm{tt}} = .04$ for conscientiousness and $\Delta r_{\mathrm{tt}} = .14$ for agreeableness (see Figure 2). Thus, reliability estimates from typical student samples that are conventional used for scale construction tend to represent the upper bound of measurement precision; for more diverse samples that also include people without academic qualifications respective reliability estimates are expected to be somewhat smaller. Memory effects might serve as a potential explanation for these differences because larger retest reliabilities arise when respondents recall previous answers from memory without properly rereading the items. Thus, the higher reliabilities of individuals with higher education might simply be a consequence of their larger cognitive capacity. The respective age-related effect on test-retest reliabilities was more complex (see Figure 1). Whereas the openness scale showed a continuous decline with increasing age, extraversion and neuroticism scales followed an inverted U-shaped form. It is noteworthy that the results for the latter closely correspond to the four year stability estimates reported by Lucas and Donnellan (2011). Thus, it might be provocative to speculate that the previously observed long-term stability estimates for the Big Five more strongly reflect reliability differences rather than true trait changes. So far, there are no studies that examined personality development after correcting for transient error in their measures.

In conclusion, the presented results extend previous research on the psychometric properties of the BFI-10 (Rammstedt, 2007) and provided estimates of the test-retest reliabilities for the five traits of personality in a diverse sample of the general population. Moreover, the study also highlighted the danger of generalizing psychometric information such as test-retest reliabilities form student samples to more heterogeneous groups of respondents.

References

Bosnjak, M., Haas, I., Galesic, M., Kaczmirek, L., Bandilla, W., & Couper, M. P. (2013). Sample composition discrepancies in different stages of a probability-based online panel. *Field Methods, 25*, 339-360. doi:10.1177/1525822X12472951

Federal Statistical Office (2010). *Statistisches Jahrbuch 2010 für die Bundesrepublik Deutschland* [Statistical yearbook 2010 for the Federal Republic of Germany]. Wiesbaden, Germany: Federal Statistical Office.

Gnambs, T. (2014a). Facets of measurement error for scores of the Big Five: Three reliability generalizations. *Personality and Individual Differences*. Advance online publication. doi:10.1016/j.paid.2014.08.019

Gnambs, T. (2014b). A meta-analysis of dependability coefficients (test-retest reliabilities) for measures of the Big Five. *Journal of Research in Personality, 52*, 20-28. doi:10.1016/j.jrp.2014.06.003

Lang, F. R., John, D., Lüdtke, O., Schupp, J., & Wagner, G. G. (2011). Short assessment of the Big Five: Robust across survey methods except telephone interviewing. *Behavior Research Methods*, *43*, 548-567. doi:10.3758/s13428-011-0066-z

Löckenhoff, C. E., Terracciano, A., Bienvenu, O. J., Patriciu, N. S.,Nestadt, G., McCrae, R. R., ... Costa, P. T., Jr. (2008). Ethnicity, education, and the temporal stability of personality traits in the East Baltimore Epidemiologic Catchment Area study. *Journal of Research in Personality, 42*, 577-598. doi:10.1016/j.jrp.2007.09.004

Lucas, R. E., & Donnellan, M. B. (2011). Personality development across the life span: Longitudinal analyses with a national sample from Germany. *Journal of Personality and Social Psychology, 101*, 847-861. doi:10.1037/a0024298

McCrae, R. R., Kurtz, J. E., Yamagata, S., & Terracciano, A. (2011). Internal consistency, retest reliability, and their implications for personality scale validity. *Personality and Social Psychology Review*, *15*, 28-50. doi:10.1177/1088868310366253

Rammstedt, B. (2007). The 10-Item Big Five Inventory: Norm values and investigation of sociodemographic effects based on a German population representative sample. *European Journal of Psychological Assessment, 23*, 193-201. doi:10.1027/1015-5759.23.3.193

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:10.1016/j.jrp.2006.02.001

Rammstedt, B., Kemper, C. J., & Borg, I. (2013). Correcting Big Five personality measurements for acquiescence: An 18-country cross-cultural study. *European Journal of Personality*, *27*, 71-81. doi:10.1002/per.1894

Rattinger, H., Roßteutscher, S., Schmitt-Beck, R., & Weßels, B. (2013): Shortterm campaign panel (GLES 2009). GESIS Data Archive, Cologne. ZA5305 Data file Version 4.0.0. doi:10.4232/1.11766

Schmidt, F. L., Le, H., & Ilies, R. (2003). Beyond alpha: An empirical examination of the effects of different sources of measurement error on reliability estimates for measures of individual differences constructs. *Psychological Methods, 8*, 206-224. doi:10.1037/1082-989X.8.2.206

Schneider, S. L. (2008). Applying the ISCED-97 to the German educational qualifications. In S. L. Schneider (Ed.), *The International Standard Classification of Education* (pp. 77-102). Mannheim, Germany: MZES.

Sutin, A. R., Costa Jr, P. T., Evans, M. K., & Zonderman, A. B. (2013). Personality assessment in a diverse urban sample. *Psychological Assessment*, *25*, 1007-1012. doi:10.1037/a0032396

Table 1.

Sociodemographic Characteristics of Retest Samples as Compared to the German Microcensus

	Microcensus 2009 [1]	Meta-analytic averages [2]	BFI-10 scale development [3]	Present sample
Sample size		92	57	1,696
Percent female	51%	63%	66%	50%
Mean age (*SD*)	45 (16)	25 (7)	25 (-)	42 (14)
Higher education (ISCED ≥ 4)	21%	70% [4]	100%	30%

Note. [1] recalculated from Federal Statistical Office (2010); [2] from Gnambs (2014b); [3] from Rammstedt & John (2007); [4] Student (versus non-student) samples

Table 2.

Effects of Sociodemographic Factors on Test-Retest Reliabilities

Predictors		Openness		Conscientiousness		Extraversion		Agreeableness		Neuroticism	
		B (*SE*)	|*t*|	*B* (*SE*)	|*t*|	*B* (*SE*)	|*t*|	*B* (*SE*)	|*t*|	*B* (*SE*)	|*t*|
	Main effects										
1.	Trait score at T1	0.71 (0.03)	25.61^{*}	0.66 (0.03)	22.70^{*}	0.79 (0.03)	32.04^{*}	0.57 (0.03)	18.52^{*}	0.70 (0.03)	24.82^{*}
2a.	Age: linear trend	0.06 (0.02)	3.49^{*}	0.10 (0.02)	4.80^{*}	0.02 (0.02)	1.07	0.02 (0.02)	1.18	-0.08 (0.02)	4.13^{*}
2b.	quadratic trend	-0.01 (0.02)	0.66	0.01 (0.02)	0.37	0.03 (0.02)	1.90^{+}	-0.03 (0.02)	1.38	0.04 (0.02)	2.00^{*}
3.	Sex	0.01 (0.02)	0.74	0.02 (0.02)	1.15	0.03 (0.02)	1.60	0.00 (0.02)	0.11	0.08 (0.02)	4.20^{*}
4.	Educational level	0.03 (0.02)	1.59	0.00 (0.02)	0.22	-0.01 (0.02)	0.29	0.04 (0.02)	1.74	-0.03 (0.02)	1.25
	Moderator effects (= interactions with trait score at T1)										
5a.	Age: linear trend	-0.04 (0.02)	2.33^{*}	0.00 (0.02)	0.11	0.03 (0.02)	1.95^{+}	-0.02 (0.02)	1.04	0.02 (0.02)	0.93
5b.	quadratic trend	-0.02 (0.02)	1.11	-0.02 (0.02)	1.03	-0.04 (0.02)	-2.63^{*}	0.01 (0.02)	0.26	-0.04 (0.02)	1.98^{*}
6.	Sex	-0.01 (0.02)	0.49	0.00 (0.02)	0.25	-0.02 (0.02)	1.34	0.00 (0.02)	0.01	0.01 (0.02)	0.52
7.	Educational level	0.06 (0.02)	3.11^{*}	0.04 (0.02)	1.72^{+}	0.03 (0.02)	1.70^{+}	0.07 (0.02)	3.20^{*}	0.07 (0.02)	3.59^{*}
	Explained variance (R^2)	.46		.43		.55		.31		.45	
	M (*SD*) at T1	3.59 (0.88)		3.76 (0.74)		3.18 (0.89)		2.91 (0.73)		2.63 (0.87)	
	M (*SD*) at T2	3.44 (0.86)		3.69 (0.72)		3.14 (0.89)		2.97 (0.69)		2.67 (0.82)	

Note. Linear regressions of standardized trait scores at second measurement occasion (T2) on standardized trait scores at first measurement occasion (T1) and sociodemographic variables. Coding: *z*-standardization for age; -1 = men and 1 = women for sex; -1 = ISCED <= 3 and 1 = ISCED ≥ 4 for educational level.

* $p < .05$, $^{+}$ $p < .10$

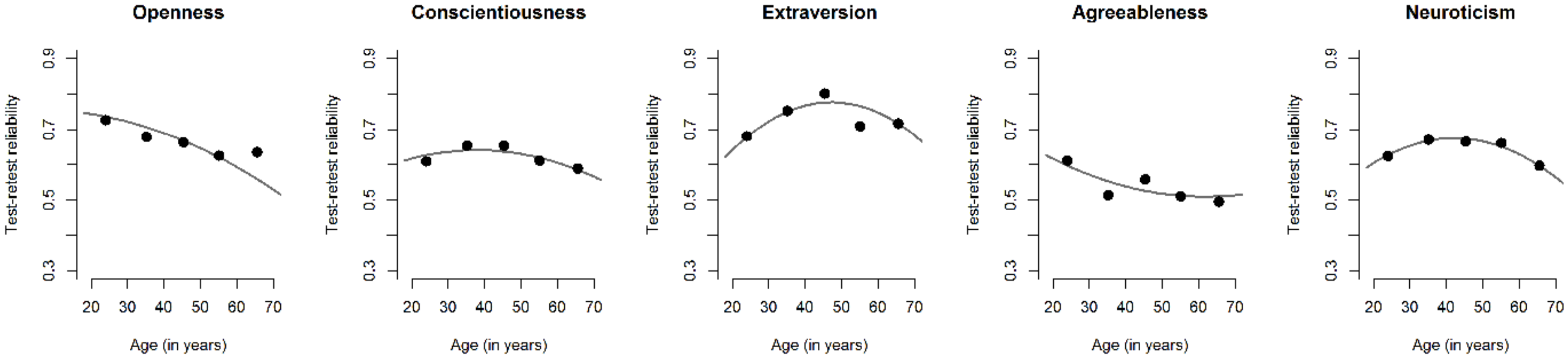

Figure 1. Test-retest reliabilities by age. Solid lines represent model implied change trajectories; dots indicate observed mean reliabilities of age groups.

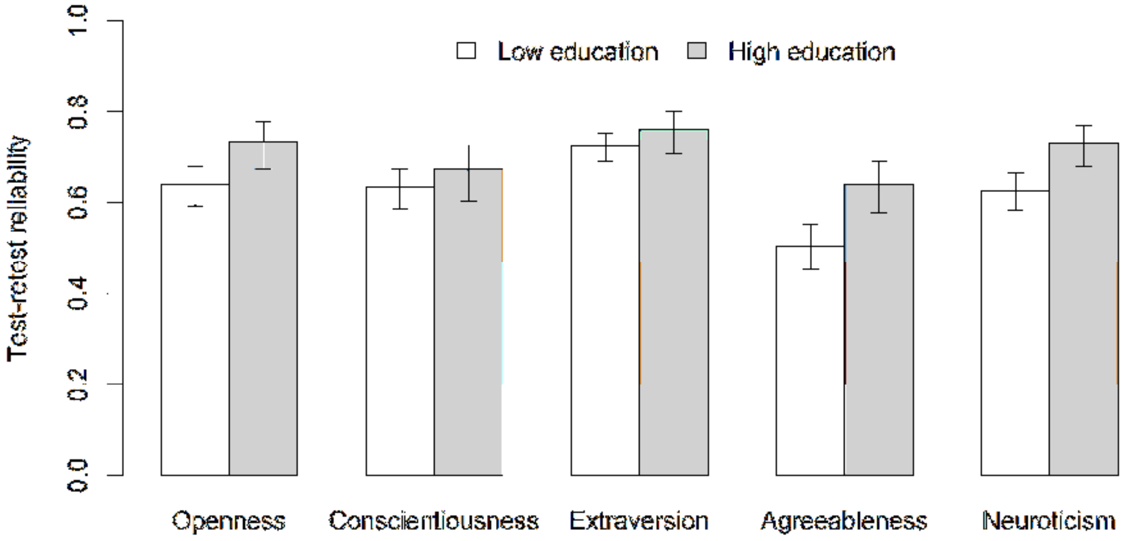

Figure 2. Test-retest reliabilities with 95% bias-corrected confidence intervals (based upon 10,000 bootstrap samples) by educational level.