Don't Keep It Too Simple: Simplified Items Do Not Improve Measurement Quality

Rammstedt, Beatrice; Roemer, Lena; Danner, Daniel; Lechner, Clemens

Original Article

Don’t Keep It Too Simple

Simplified Items Do Not Improve Measurement Quality

Beatrice Rammstedt[1], Lena Roemer[1], Daniel Danner[2], and Clemens M. Lechner[1]

[1]GESIS – Leibniz Institute for the Social Sciences, Mannheim, Germany

[2]Psychological Diagnostics & Qualification, University of Applied Labour Studies, Mannheim, Germany

Abstract: When formulating questionnaire items, generally accepted rules include: Keeping the wording as simple as possible and avoiding double-barreled items. However, the empirical basis for these rules is sparse. The present study aimed to systematically investigate in an experimental design whether simplifying items of a personality scale and avoiding double-barreled items (i.e., items that contain multiple stimuli) markedly increases psychometric quality. Specifically, we compared the original items of the Big Five Inventory-2 – most of which are either double-barreled or can be regarded as complexly formulated – with simplified versions of the items. We tested the two versions using a large, heterogeneous sample (*N* = 2,234). The simplified versions did not possess better psychometric quality than their original counterparts; rather, they showed weaker factorial validity. Regarding item characteristics, reliability, and criterion validity, no substantial differences were identified between the original and simplified versions. These findings were also replicated for the subsample of lower-educated respondents, who are considered more sensitive to complex item formulations. Our study thus suggests that simplifying item wording and avoiding double-barreled items in a personality inventory does not improve the quality of a questionnaire; rather, using simpler (and consequently more vague) item formulations may even decrease factorial validity.

Keywords: BFI-2, item simplification, Big Five, personality assessment, low educated

OPEN MATERIALS

When formulating questionnaire items, one generally accepted rule is that the wording should be kept as simple as possible (Lenzner, 2012). In particular, the use of double-barreled items, including two separate stimuli in the same item while allowing for only one answer (Elson, 2016; Gehlbach, 2015; Porst, 2000) – is considered a “sin” in questionnaire development. Unduly high reading comprehension requirements in item formulations are assumed to result in “construct-irrelevant difficulty,” thereby leading to inaccurate scores, especially for weaker readers (Messick, 1995). In addition, it is assumed that the multiple stimuli in double-barreled items confuse respondents, as they are unclear about which elements in the item they should respond to. Thus, they might mentally modify the item by ignoring one stimuli while responding to the other (Krosnick, 1991) or substitute a difficult question with a simpler one (Kahneman, 2011).

Although these postulations regarding item formulation have existed for decades, very little research has been conducted to test whether they indeed hold – whether simpler and single-barreled items perform better than more complex and double-barreled ones. Efforts to reduce the linguistic complexity of items have primarily been undertaken in the context of cognitive assessments, for example, of mathematical skills. Here, several meta-analyses have shown that item simplification substantially reduces the language demands of the assessment for students with low English proficiency (e.g., Kieffer et al., 2009). In the field of personality assessment, evidence of the effects of item simplification is scarce and inconclusive. Research suggests that the comprehensibility of items is related to item nonresponse and unreliable responses (e.g., Lenzner, 2012) and results in higher cognitive burden and lower retest stabilities (Lenzner et al., 2010). Based on such findings, several efforts have been made to increase the comprehensibility of items. Examples include the modification of the Revised NEO Personality Inventory (NEO-PI-R), which resulted in the NEO-PI-3 (McCrae et al., 2005), and of the Multidimensional Personality Questionnaire (MPQ), which resulted in the MPQ-SF (Patrick et al., 2012). All this was done on the assumption that simplified items are better understood and result in better psychometric properties of the scale. However, studies directly investigating this assumption are rare and often unsupportive. Indeed, Pargent and colleagues (2019) found that neither a simplification nor a “deterioration” of the NEO Five-Factor Inventory (FFI) items markedly affected their reliability coefficients and factorial structure.

Although the theoretical critique of double-barreled items has also existed for over half a century, supporting evidence is scarce. Only a few studies have investigated

https://doi.org/10.1027/1015-5759/a000741

whether double-barreled items do indeed confuse respondents, thereby lowering the quality of the items. Some found evidence supporting the validity-reducing effect of double-barreled items (Grant Levy, 2018; Menold, 2020; Menold & Raykov, 2022). In addition, these items were found to cause more difficulties in responding (Bassili & Scott, 1996; Herzberg & Brähler, 2006). However, other studies did not support a general tendency of double-barreled items to show weaker psychometric quality (Schult et al., 2019) or increased item nonresponse (Hox & Borgers, 2001). One reason for these inconclusive results might be because there is a huge variety of multi-stimuli, thus double-barreled items differ, especially in the similarity or contradiction of the used stimuli.

The aim of the present study was to systematically investigate whether reducing the linguistic complexity of personality items and avoiding double-barreled items markedly increases the psychometric quality of a personality inventory. For this purpose, we used the Big Five Inventory-2 (BFI-2; Soto & John, 2017), of which about half of the items contain separate, albeit similar, stimuli and thus can be regarded as double-barreled (see Schult et al., 2019). In an experimental design, we compared the original psychometric properties of the original BFI-2 with those of a simplified version of the inventory whose items were optimized for readability – that is, used linguistically simplified phrases and/or phrases reduced to a single stimulus only. We tested these two versions based on a large, heterogeneous sample rather than on the typically used samples of psychology students, who are usually very familiar with responding to questionnaires and highly literate, and, thus, less prone to difficulties resulting from linguistically complex or double-barreled items.

Method

Sample and Design

Data were collected as part of the Programme for the Assessment of Adult Competencies (PIAAC) Pilot (Organisation for Economic Co-operation and Development [OECD], 2018), in which various personality scales in their original and modified forms were tested for potential inclusion in PIAAC. Data were collected via the Survey Monkey platform. Participants aged 16–65 years were selected in the USA and the UK according to a quota scheme based on sex, age, and region, broadly representative of US and UK census data. Participants received a personal URL to access the survey platform. Each access could be used only once. Data collection took place between June and July 2016. For the present analyses, we merged data from the two sites, resulting in a total sample of N = 5,910 respondents.

From this total sample, we excluded around 20% of respondents because they failed at least one of eight quality checks included in the dataset (e.g., agreement with the item "I fly to the International Space Station"; low response times; no correct answers on an ability test; same responses to at least four pairs of positively and negatively keyed items of the same factor), resulting in a quality-controlled sample of N = 4,711 respondents (57.6% female).

In addition to item simplification, the PIAAC Pilot aimed to test the effect of the response scale format – namely, a 5-point scale versus a 4-point scale without a midpoint. Both designs were concurrently tested, and respondents were randomly assigned to one of four experimental conditions. In our analyses, we focused on contrasting the results of the two conditions – original and simplified BFI-2 – based on the 5-point response scale, as it is usually and originally used for the BFI-2 (see Soto & John, 2017). Our final analysis sample thus comprised 2,234 respondents (57.9% female) participating in the two conditions – original versus simplified items – using a 5-point response scale. The corresponding results for the two conditions with a 4-point response format are displayed on the project's OSF page (https://osf.io/atfsv/).

Instruments

Big Five Inventory-2 (BFI-2)

The BFI-2 (Soto & John, 2017) assesses the Big Five personality domains and three central facets per domain. It comprises 60 phrase-like items to be answered on a 5-point rating scale ranging from 1 = *fully disagree* to 5 = *fully agree*.

Simplified Big Five Inventory-2

As most of the 60 BFI-2 items can be regarded as linguistically complex or include multiple stimuli, the expert group responsible for developing the PIAAC Pilot[1] aimed to make "the original scales more appropriate for use with the general adult population. In many cases, the original items were perceived as potentially too complex and abstract for the less literate members of the general population." (OECD, 2018, p. 1). Therefore, the expert group developed simplified versions for 55 of the 60 items. In particular, they reduced the number of stimuli in all items to one (n = 29 items). When selecting one of the two stimuli in an item, the more content-valid and semantically simpler stimulus was selected. Further, the experts tried to simplify

[1] The members of the expert group were Daniel Danner, Beatrice Rammstedt, Brent Roberts, Richard Roberts, Manfred Schmitt, Fons van de Vijver, and Susanne Weiß.

the language of the items to enhance readability, especially for low-literate adults (n = 26 items). All simplified item versions were reviewed – especially with regard to (keeping its) content validity and signed off by all members of the expert group. The original and simplified items of the BFI-2 are displayed in Figure 1.

To assess the success of this simplification of the BFI-2 items, we computed Flesch Reading Ease scores (Flesch, 1948) as a numeric indicator of the readability of both the original and simplified items. This score is a commonly used measure of readability and relates the average sentence length in words and the average word length in syllables. Higher scores indicate higher reading ease. The median Flesch Reading Ease score was *Mdn* = 59.75 (*SD* = 41.94) for the original items and *Mdn* = 76.89 (*SD* = 41.20) for the simplified items (see the project's OSF page: https://osf.io/atfsv/, for the scores for all items), demonstrating that the simplified inventory version indeed had higher readability. According to the classification by Flesch, the original BFI-2 version can be regarded as fairly difficult, and the simplified version is fairly easy to read (see Table 5 in Flesch, 1948).

Criterion Variables

Based on relevant outcome measures investigated in the context of PIAAC (see Lechner et al. 2019; OECD, 2017; Rammstedt et al., in press), the following correlates were used to investigate the criterion validity of the two BFI-2 versions:

(a) satisfaction with life measured with the well-established single item (see Nießen et al., 2020) "How satisfied are you with your life in general?" rated on a scale from 1 = *not satisfied at all* to 10 = *completely satisfied*,
(b) self-rated health based on the single item "How would you describe your health status in general?" rated on a scale from 1 = *excellent* to 5 = *poor*, which we recoded such that higher values represented better health,
(c) current household income based on nine categories ranging from 1 = *less than $10,000* to 9 = *more than $150,000*.

In addition, we assessed sex, age (in years), and educational attainment (six categories ranging from 1 = *primary school* to 6 = *doctoral or professional degree*).

Results and Discussion

Does simplifying item wording and avoiding double-barreled items enhance the psychometric properties of the BFI-2? To investigate this question, we directly compared the descriptive statistics, reliability, validity, and measurement invariance of the original BFI-2 with those of the simplified version.

Because previous studies suggest that persons with lower cognitive ability and a lower level of education are more sensitive to complex item formulations (e.g., Knauper et al., 1997; Smith, 1982), we repeated in a second step each of our analyses for the lower-educated subsample (n = 564; high school diploma or lower) and investigated whether differences in psychometric quality between the two versions were indeed more pronounced in this subgroup.

Descriptive Statistics and Reliability Estimates for the Scales

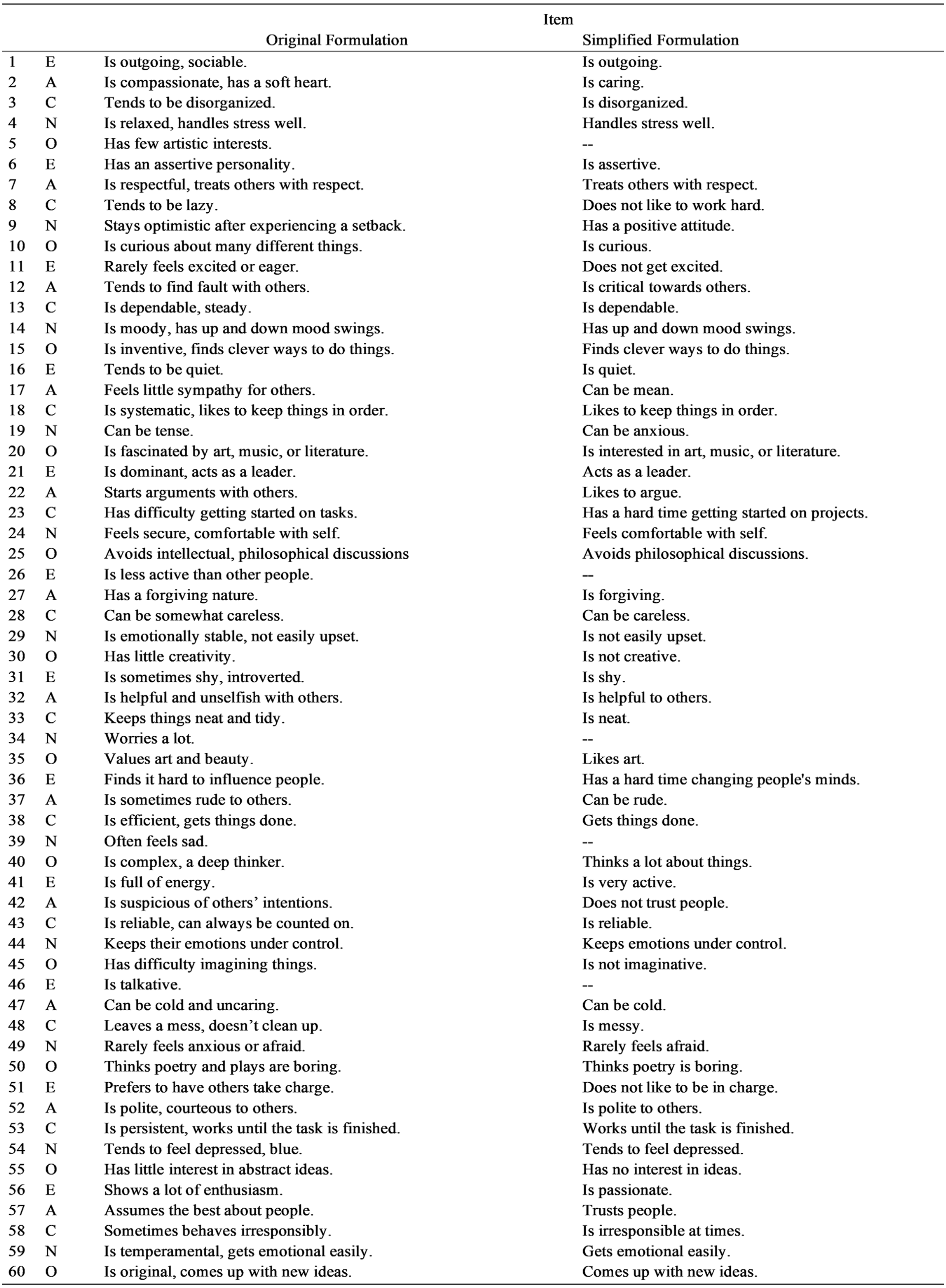
Table 1 shows means, standard deviations, standardized mean level differences, and internal consistency estimates for the domains and facets of the original and the simplified BFI-2 versions. Unsurprisingly, given the changed wording of the items, the means of some of the domain and facet scores differed slightly between the two versions. In all six cases in which Cohen's d exceeded or was equal to .20, means were higher for the simplified version, indicating higher levels of agreement with the simplified items.

We also compared the means on the level of the 60 individual items. The mean scores of 25 items differed slightly ($d \geq |.20|$). Here, too, agreement with the simplified items was, in most cases, higher than with the original items (17 items).

The reliability coefficients (Cronbach's α and McDonald's ω total; ω total was calculated with the R package MBESS; Kelley, 2018) for the five domain and 15 facet scales were highly similar for both the original and the simplified items (e.g., average Cronbach's α for the domain scale scores was .86 and .85, respectively). Only in one case (Intellectual Curiosity) did the difference between Cronbach's α coefficients of the two versions exceed |.10|, with a higher reliability coefficient for the original compared with the simplified version. Results for McDonald's ω coefficients were similar: Only in two cases (Intellectual Curiosity and Compassion) did differences in the coefficients between the versions exceed |.10|.

Domain scale intercorrelations were also highly similar for the original and simplified versions, with (absolute z- and back-transformed) averages of r = .34 and .33, respectively (see Table 2). Tucker's phi congruence coefficient of the values in the correlation matrix was .99, indicating near-perfect congruence.

In sum, these results indicate that the simplified items did not outperform the original ones with regard to the reliability estimates or major differences in the descriptive statistics of the scales. On the contrary, where differences in reliability estimates were found, these indicated better performance of the original compared with the simplified



		Item	
		Original Formulation	Simplified Formulation
1	E	Is outgoing, sociable.	Is outgoing.
2	A	Is compassionate, has a soft heart.	Is caring.
3	C	Tends to be disorganized.	Is disorganized.
4	N	Is relaxed, handles stress well.	Handles stress well.
5	O	Has few artistic interests.	--
6	E	Has an assertive personality.	Is assertive.
7	A	Is respectful, treats others with respect.	Treats others with respect.
8	C	Tends to be lazy.	Does not like to work hard.
9	N	Stays optimistic after experiencing a setback.	Has a positive attitude.
10	O	Is curious about many different things.	Is curious.
11	E	Rarely feels excited or eager.	Does not get excited.
12	A	Tends to find fault with others.	Is critical towards others.
13	C	Is dependable, steady.	Is dependable.
14	N	Is moody, has up and down mood swings.	Has up and down mood swings.
15	O	Is inventive, finds clever ways to do things.	Finds clever ways to do things.
16	E	Tends to be quiet.	Is quiet.
17	A	Feels little sympathy for others.	Can be mean.
18	C	Is systematic, likes to keep things in order.	Likes to keep things in order.
19	N	Can be tense.	Can be anxious.
20	O	Is fascinated by art, music, or literature.	Is interested in art, music, or literature.
21	E	Is dominant, acts as a leader.	Acts as a leader.
22	A	Starts arguments with others.	Likes to argue.
23	C	Has difficulty getting started on tasks.	Has a hard time getting started on projects.
24	N	Feels secure, comfortable with self.	Feels comfortable with self.
25	O	Avoids intellectual, philosophical discussions	Avoids philosophical discussions.
26	E	Is less active than other people.	--
27	A	Has a forgiving nature.	Is forgiving.
28	C	Can be somewhat careless.	Can be careless.
29	N	Is emotionally stable, not easily upset.	Is not easily upset.
30	O	Has little creativity.	Is not creative.
31	E	Is sometimes shy, introverted.	Is shy.
32	A	Is helpful and unselfish with others.	Is helpful to others.
33	C	Keeps things neat and tidy.	Is neat.
34	N	Worries a lot.	--
35	O	Values art and beauty.	Likes art.
36	E	Finds it hard to influence people.	Has a hard time changing people's minds.
37	A	Is sometimes rude to others.	Can be rude.
38	C	Is efficient, gets things done.	Gets things done.
39	N	Often feels sad.	--
40	O	Is complex, a deep thinker.	Thinks a lot about things.
41	E	Is full of energy.	Is very active.
42	A	Is suspicious of others' intentions.	Does not trust people.
43	C	Is reliable, can always be counted on.	Is reliable.
44	N	Keeps their emotions under control.	Keeps emotions under control.
45	O	Has difficulty imagining things.	Is not imaginative.
46	E	Is talkative.	--
47	A	Can be cold and uncaring.	Can be cold.
48	C	Leaves a mess, doesn't clean up.	Is messy.
49	N	Rarely feels anxious or afraid.	Rarely feels afraid.
50	O	Thinks poetry and plays are boring.	Thinks poetry is boring.
51	E	Prefers to have others take charge.	Does not like to be in charge.
52	A	Is polite, courteous to others.	Is polite to others.
53	C	Is persistent, works until the task is finished.	Works until the task is finished.
54	N	Tends to feel depressed, blue.	Tends to feel depressed.
55	O	Has little interest in abstract ideas.	Has no interest in ideas.
56	E	Shows a lot of enthusiasm.	Is passionate.
57	A	Assumes the best about people.	Trusts people.
58	C	Sometimes behaves irresponsibly.	Is irresponsible at times.
59	N	Is temperamental, gets emotional easily.	Gets emotional easily.
60	O	Is original, comes up with new ideas.	Comes up with new ideas.

Figure 1. Formulations of the 60 BFI-2 items in their original and simplified forms. E = Extraversion; A = Agreeableness; C = Conscientiousness; N = Negative Emotionality; O = Open-Mindedness.

Table 1. Means (*M*), standard deviations (*SD*), Cronbach's α, and McDonald's ω for the Original and Simplified BFI-2

	M		*SD*			Cronbach's α		McDonald's ω	
	Original	Simplified	Original	Simplified	*d*	Original	Simplified	Original	Simplified
Extraversion	3.13	3.27	0.71	0.69	0.20	.86	.85	.87	.86
Sociability	2.92	3.00	0.99	0.97	0.08	.85	.82	.85	.82
Assertiveness	3.13	3.32	0.88	0.84	0.22	.80	.76	.81	.80
Energy	3.34	3.50	0.77	0.77	0.20	.69	.66	.72	.73
Agreeableness	3.79	3.71	0.55	0.56	−0.13	.81	.82	.81	.83
Compassion	3.92	3.82	0.68	0.65	−0.15	.58	.64	.59	.71
Respectfulness	4.11	4.02	0.60	0.63	−0.14	.67	.63	.66	.61
Trust	3.33	3.30	0.74	0.76	−0.03	.71	.69	.71	.72
Conscientiousness	3.83	3.89	0.67	0.63	0.10	.89	.87	.89	.87
Organization	3.81	3.79	0.85	0.90	−0.02	.83	.87	.83	.88
Productiveness	3.80	3.97	0.79	0.69	0.23	.78	.72	.79	.71
Responsibility	3.87	3.92	0.68	0.67	0.06	.70	.65	.67	.63
Negative Emotionality	2.76	2.80	0.84	0.80	0.05	.92	.91	.93	.91
Anxiety	3.10	3.08	0.91	0.89	−0.03	.80	.77	.80	.78
Depression	2.56	2.46	0.95	0.92	−0.11	.85	.83	.86	.85
Volatility	2.63	2.87	0.93	0.90	0.27	.85	.79	.85	.79
Open-Mindedness	3.66	3.75	0.63	0.59	0.14	.84	.82	.84	.82
Intellectual curiosity	3.77	3.96	0.69	0.58	0.29	.65	.54	.66	.55
Aesthetic sensitivity	3.47	3.42	0.85	0.85	−0.06	.71	.69	.74	.72
Creative imagination	3.76	3.87	0.73	0.74	0.16	.73	.80	.73	.81
M (Domain)	3.43	3.48	0.68	0.65	0.07	.86	.85	.87	.86
M (Facet)	3.43	3.49	0.80	0.78	0.06	.75	.72	.75	.74

Table 2. Manifest Intercorrelations of the Original BFI-2 Scale (below the main diagonal) and the Simplified BFI-2 Scale (above the main diagonal)

	E	A	C	N	O
Extraversion		.26	.38	−.44	.39
Agreeableness	.22		.46	−.44	.18
Conscientiousness	.39	.40		−.43	.19
Negative Emotionality	−.46	−.37	−.48		−.11
Open-Mindedness	.36	.29	.18	−.17	

Note. E = Extraversion; A = Agreeableness; C = Conscientiousness; N = Negative Emotionality; O = Open-Mindedness. All correlations are significant with $p < .001$.

BFI-2 version. Likewise, the scale intercorrelations suggest similar associations of the domains for both versions. The slight differences found in scale and item means were not generally surprising given that the items had been reformulated: There seems to be a tendency toward a stronger agreement with the simplified items, thus indicating that the reformulations often reduced item difficulty. As we discuss below, this might be because simplified items became more ambiguous and easier to agree with.

Criterion Validity

To compare the criterion validity of the two BFI-2 versions, we correlated the domain and facet scales with a set of typical criterion variables (Danner et al., 2019; Lechner et al., 2019), namely, life satisfaction, self-rated health, and income (see Table 3). Both the original and the simplified BFI-2 scales showed the correlational pattern with all three criterion variables that was expected based on previous research: The strongest effects were found for life satisfaction and health; all three criteria were negatively associated with Negative Emotionality; life satisfaction was additionally positively associated with Conscientiousness (see Danner et al., 2019; Lechner et al., 2019).

Moreover, for these criterion correlations, the patterns were highly similar for the original and the simplified versions. Among the 60 corresponding coefficients, the largest difference between the correlations was a (*z*- and back-transformed) Cohen's *d* of .11; Tucker's phi congruence

Table 3. Correlations of the original and the simplified BFI-2 domain and facet scales with criterion variables

	Life satisfaction		Health		Income	
	Original	Simplified	Original	Simplified	Original	Simplified
Extraversion	.43	.37	.31	.32	.19	.18
Sociability	.31	.26	.16	.14	.13	.06
Assertiveness	.29	.27	.20	.22	.16	.19
Energy	.46	.38	.43	.44	.19	.19
Agreeableness	.27	.26	.10	.18	.05	.06
Compassion	.13	.19	.01	.12	.04	.04
Respectfulness	.20	.11	.06	.08	.03	.00
Trust	.32	.32	.16	.23	.05	.10
Conscientiousness	.32	.25	.24	.26	.23	.16
Organization	.24	.18	.19	.23	.16	.11
Productiveness	.33	.24	.24	.27	.22	.16
Responsibility	.28	.20	.20	.14	.21	.13
Negative Emotionality	−.54	−.51	−.37	−.38	−.21	−.22
Anxiety	−.44	−.39	−.30	−.28	−.16	−.18
Depression	−.62	−.62	−.40	−.43	−.21	−.24
Volatility	−.41	−.34	−.31	−.31	−.20	−.16
Open-Mindedness	.16	.08	.14	.12	.10	.04
Intellectual curiosity	.10	.05	.12	.07	.12	.06
Aesthetic sensitivity	.10	.02	.08	.05	.05	−.01
Creative imagination	.20	.13	.15	.16	.09	.05
M (*z*- and back-transformed)	.32	.27	.21	.23	.14	.12

Note. All correlations $\geq |.10|$ are $p < .001$ (slight imprecision due to rounding).

coefficients for the correlations with life satisfaction, health, and income were .99, .98, and .96, respectively.

Results thus indicate that the scales of both BFI-2 versions show similar external validities, with neither version outperforming the other.

Factorial Validity and Measurement Equivalence

To explore the factorial structure of the two BFI-2 versions, we first conducted a principal component analysis (PCA), as is typically done for Big Five instruments (see, e.g., Soto & John, 2017). Second, we tested the model fit and its invariance across the two versions more formally using multigroup analyses within a confirmatory factor analysis (CFA) framework.

In the first step, we ran a PCA of the 60 BFI-2 items in both experimental groups. The extracted five factors were rotated toward a simple structure (Varimax). The resulting factorial loadings of the items are displayed in Table 4. For both the original and the simplified versions, nearly all items loaded the highest of their corresponding factors. Exceptions for both versions were the Extraversion item 26 ("Is less active than other people"), which loaded primarily on Conscientiousness, and the Agreeableness item 42 ("Is suspicious of others' intentions"), which loaded highest on Negative Emotionality. In addition, in the simplified version, Extraversion item 41 (original: "Is full of energy"; simplified: "Is very active") loaded highest on Negative Emotionality. Overall, the loading patterns were highly similar across both versions, with an average primary loading of .59 for the original and .57 for the simplified version and secondary loadings of .12 and .13, respectively, as well as an average congruence (Tucker's phi) of .97.

In some cases, however, the resulting loading pattern tended to be less clear for the simplified version. For example, simplifying the Open-Mindedness item "Is complex, a deep thinker" to "Thinks a lot about things" increased the secondary loadings on Negative Emotionality.

In the second step, we tested and compared the multidimensional structure of the two BFI-2 versions using CFA. Similar to the procedure described by Soto & John (2017) in their original BFI-2 publication, we started by analyzing the hypothesized structure for each of the five domains separately. In these models, we allowed the 12 items per domain to load on three correlated factors representing the three facets per domain. Additionally, we included in all models an acquiescence factor (see Soto & John, 2017) that had unit loadings on all items.

Fit indices for the measurement models for both the original and the simplified BFI-2 items are displayed in Table 5. All models based on the original items showed an acceptable to good fit to the data and tended to fit better than

Table 4. Standardized loadings of varimax rotated principal component analyses for the original and simplified BFI-2 item versions

	Original					Simplified				
Item	E	A	C	N	O	E	A	C	N	O
1	***.74***	.20	.07	−.23	.07	***.69***	.20	.14	−.20	.16
2	.08	***.72***	.06	.14	.12	.27	***.56***	.19	.17	.21
3	−.01	.04	***−.72***	.13	.03	.05	−.02	***−.75***	.21	.04
4	.19	.04	.17	***−.69***	.11	.14	.04	.16	***−.67***	.12
5	.13	−.11	.04	.03	***−.42***	.10	−.01	.03	−.03	***−.36***
6	***.68***	−.21	.20	−.06	.17	***.60***	−.19	.17	−.21	.18
7	.05	***.56***	.28	−.02	.16	.07	***.56***	.22	.05	.24
8	−.11	−.11	***−.63***	.24	−.03	−.14	−.17	***−.47***	.12	−.16
9	.34	.18	.15	***−.52***	.2	.30	.34	.19	***−.45***	.19
10	.21	.14	.07	−.02	***.57***	.17	−.01	.04	−.04	***.53***
11	***−.30***	−.21	.02	.04	−.20	***−.37***	−.21	−.02	−.14	−.20
12	−.06	***−.48***	−.08	.38	−.07	.03	***−.56***	−.11	.24	.02
13	.07	.23	***.48***	−.14	.07	.08	.27	***.37***	.07	.02
14	.01	−.17	−.18	***.74***	−.03	.00	−.21	−.13	***.69***	.00
15	.23	−.02	.21	−.06	***.57***	.22	−.01	.23	−.04	***.57***
16	***−.71***	−.05	.06	.09	.04	***−.70***	−.04	.06	.09	.10
17	.12	***−.28***	.00	−.08	−.16	.14	***−.59***	−.16	.28	.00
18	.04	−.06	***.72***	−.01	−.01	.03	−.03	***.71***	−.06	.06
19	−.04	−.19	−.06	***.70***	.01	−.06	−.13	−.10	***.69***	.05
20	.03	.13	−.10	.04	***.72***	.00	.08	−.07	.04	***.69***
21	***.63***	−.22	.24	−.11	.25	***.65***	−.13	.20	−.23	.25
22	.32	***−.45***	−.19	.29	−.01	.32	***−.46***	−.19	.19	.07
23	−.12	−.06	***−.58***	.33	−.10	−.11	−.16	***−.51***	.27	−.03
24	.33	.04	.31	***−.54***	.13	.28	.05	.22	***−.55***	.13
25	−.08	.07	−.11	.12	***−.52***	−.10	.02	−.02	.04	***−.40***
26	*−.28*	−.01	**−.35**	.29	−.13	*−.20*	.03	−.34	**.37**	−.14
27	.11	***.61***	.00	−.10	.12	.10	***.56***	−.05	−.18	.17
28	.03	−.14	***−.59***	.24	−.05	.03	−.22	***−.58***	.21	.01
29	.13	.08	.25	***−.70***	.08	.05	.08	.07	***−.61***	.05
30	−.04	−.07	−.08	.06	***−.64***	−.11	−.07	−.10	.09	***−.71***
31	***−.68***	−.01	−.08	.29	.02	***−.66***	−.03	−.07	.27	.02
32	.11	***.56***	.26	.00	.18	.23	***.50***	.37	.09	.17
33	.10	.08	***.74***	−.09	−.02	.05	.10	***.72***	−.10	.01
34	−.10	.09	−.07	***.78***	−.02	−.08	−.05	−.02	***.77***	−.05
35	.05	.22	−.03	.05	***.69***	.02	.06	−.06	.00	***.74***
36	***−.37***	.03	−.19	.30	−.28	***−.28***	−.06	−.16	.18	−.16
37	.09	***−.55***	−.21	.31	.01	.13	***−.61***	−.28	.22	.00
38	.21	.05	***.73***	−.16	.13	.33	.04	***.62***	−.13	.16
39	−.21	−.04	−.18	***.77***	.02	−.17	−.15	−.13	***.74***	.02
40	.09	−.02	.08	.24	***.55***	.05	.04	.07	**.39**	*.37*
41	***.49***	.12	.29	−.32	.12	*.32*	.03	.34	**−.37**	.18
42	−.09	*−.33*	.06	**.48**	−.03	−.17	*−.43*	.11	**.46**	−.03
43	.10	.31	***.56***	−.13	.12	.21	.31	***.52***	.00	.13
44	.03	.06	.31	***−.60***	.11	−.05	.08	.24	***−.64***	.10
45	−.05	−.12	−.08	.13	***−.50***	−.06	−.03	−.10	.04	***−.68***
46	***.76***	.13	−.05	.07	.03	***.73***	.12	−.07	.01	.08
47	−.01	***−.68***	−.14	.19	−.02	−.03	***−.65***	−.13	.21	.04
48	−.02	−.14	***−.67***	.08	−.01	.04	−.06	***−.76***	.21	.03
49	.10	−.05	.10	***−.60***	.01	.19	−.09	.13	***−.49***	.01

(Continued on next page)

Table 4. (Continued)

	Original					Simplified				
Item	E	A	C	N	O	E	A	C	N	O
50	−.05	−.23	.05	.03	−***.57***	.01	−.13	.00	.04	−***.55***
51	−***.43***	.22	−.28	.18	−.29	−***.59***	.12	−.20	.23	−.25
52	−.06	***.59***	.31	−.03	.14	.03	***.59***	.30	.10	.13
53	.15	.22	***.62***	−.10	.13	.15	.25	***.56***	−.10	.14
54	−.22	−.09	−.20	***.77***	.01	−.15	−.19	−.17	***.74***	−.02
55	−.03	−.05	.03	.18	−***.66***	−.14	−.12	−.11	.17	−***.51***
56	***.56***	.30	.21	−.21	.19	***.43***	.22	.22	−.02	.40
57	.28	***.59***	−.01	−.22	.09	.23	***.40***	−.10	−**.40**	.04
58	.07	−.21	−***.50***	.29	.06	.07	−.25	−***.51***	.25	−.01
59	.06	−.08	−.21	***.74***	−.03	.04	.00	−.18	***.70***	−.01
60	.36	.02	.19	−.09	***.61***	.32	.00	.19	−.12	***.62***

Note. Loadings on corresponding factors are italicized; highest loadings are set in bold. E = Extraversion; A = Agreeableness; C = Conscientiousness; N = Negative Emotionality; O = Open-Mindedness. Congruence of the factors (Tucker's phi) are E = .97, A = .95, C = .98, N = .96, O = .97.

Table 5. Fit measures for the domain measurement models for the original and simplified BFI-2

	χ^2	*df*	*p*	CFI (robust)	RMSEA (robust)	SRMR
Extraversion						
Original	477	50	< .001	.922	.086	.059
Simplified	634	50	< .001	.885	.102	.096
Simplified, with correlated residuals	429	49	< .001	.926	.083	.055
Agreeableness						
Original	233	50	< .001	.953	.055	.040
Simplified	683	50	< .001	.853	.107	.102
Simplified, with correlated residuals	384	49	< .001	.923	.078	.052
Conscientiousness						
Original	331	50	< .001	.953	.068	.044
Simplified	326	50	< .001	.949	.069	.056
Negative Emotionality						
Original	215	50	< .001	.980	.052	.027
Simplified	250	50	< .001	.970	.058	.035
Open-Mindedness						
Original	250	50	< .001	.955	.057	.047
Simplified	223	50	< .001	.958	.054	.039
Total model						
Original	5,241	1,604	< .001	.894	.043	.059
Original, with correlated residuals	4,923	1,601	< .001	.903	.041	.058
Simplified[a]	5,779	1,602	< .001	.868	.047	.067
Simplified, with modifications	6,091	1,639	< .001	.858	.048	.069

Note.[a]Estimation problems, covariance matrix of latent variables not positive definite. CFI = comparative fit index; RMSEA = root-mean-square error of approximation; SRMR = standardized root-mean-square residual.

(or at least as well as) the models based on simplified items. For the simplified items, the models for Extraversion and Agreeableness showed only an acceptable fit to the data after including correlated residuals (i.e., Item 26 "Is less active than other people" and Item 41 "Is very active", and Item 42 "Does not trust people" and Item 57 "Trusts people"), thus indicating that in some cases, simplification disguised the construct specificity of the items.

Next, to explore the overall factorial validity of the two BFI-2 versions, we combined the five measurement models into a total model. The total model for the original items had an acceptable fit to the data. The model for the simplified items revealed estimation problems, indicating a not positive definite covariance matrix of latent variables, which disappeared only after substantial modifications (e.g., correlated residuals across items from different

Table 6. Testing measurement invariance for the measurement models across the two versions of the BFI-2

	Fit indices							Fit comparison		
	χ^2	*df*	*p*	CFI (robust)	RMSEA (robust)	SRMR	BIC	ΔCFI	ΔRMSEA	ΔSRMR
Extraversion[a]										
Configural	906	99	< .001	.924	.084	.053	73,160			
Metric	988	108	< .001	.917	.084	.063	73,172	.007	< .001	–.010
Scalar	1,288	116	< .001	.889	.094	.069	73,410	.028	–.010	–.006
Agreeableness[a]										
Configural	617	99	< .001	.937	.067	.043	64,426			
Metric	776	108	< .001	.918	.073	.070	64,516	.019	–.006	–.027
Scalar	1,402	116	< .001	.839	.099	.081	65,080	.079	–.026	–.011
Conscientiousness										
Configural	657	100	< .001	.951	.069	.046	64,243			
Metric	780	109	< .001	.941	.072	.060	64,296	.010	–.003	–.014
Scalar	1,160	117	< .001	.906	.088	.068	64,615	.035	–.016	–.008
Negative Emotionality										
Configural	465	100	< .001	.976	.055	.029	69,929			
Metric	498	109	< .001	.974	.055	.040	69,893	.002	< .001	–.011
Scalar	732	117	< .001	.958	.067	.046	70,065	.016	–.012	–.006
Open-Mindedness										
Configural	472	100	< .001	.956	.056	.040	69,486			
Metric	611	109	< .001	.940	.062	.053	69,555	.016	–.006	–.013
Scalar	1,121	117	< .001	.876	.086	.068	70,004	.064	–.024	–.015

Note. [a]The necessary adjustments were included for the simplified items. CFI = comparative fit index; RMSEA = root-mean-square error of approximation; SRMR = standardized root-mean-square residual; BIC = Bayesian information criterion.

domains; no facetted structure for the items for Negative Emotionality; see our OSF page). Thus, the data for the simplified items did not adhere to the theoretical structure of 60 items loading on 15 correlated facets (and an acquiescence factor).

Finally, to formally test the psychometric comparability of the scores as assessed by the two BFI-2 versions, we tested the measurement invariance of the respective models across original and simplified items. Due to the estimation difficulties for the total model, we used the domain-level measurement models for the invariance analyses. Results are presented in Table 6. For each of the five domains, we analyzed the same measurement model simultaneously for each of the two item versions (configural invariance) and subsequently constrained the loadings (metric invariance) and intercepts (scalar invariance) to be equal across the original and simplified items. Using the criteria recommended by Chen (2007), configural invariance could be assumed for the models of Agreeableness and Open-Mindedness, whereas metric invariance could be assumed for the models of Extraversion, Conscientiousness, and Negative Emotionality. Thus, whereas for Extraversion, Conscientiousness, and Negative Emotionality, the latent facets represented the same meaning across the two versions, this was not the case for Agreeableness and Open-Mindedness. Unsurprisingly, all domains were scalar non-invariant, indicating that responses to the items were systematically higher or lower across the two versions.

Results regarding the factorial structure thus indicate that simplifying items can substantially change their meaning and the latent variables. Thus, the simplified items became more vague and ambiguous indicators for a specific facet (e.g. "Can be cold" vs. "Can be cold and uncaring") or showed overlap with intentionally independent facets, thereby rendering it necessary to modify the model by allowing secondary loadings. For example, the original BFI-2 Conscientiousness item "Is full of energy" was simplified to "Is very active." Consequently, the item showed a semantic and – unsurprisingly – an empirical overlap with the (inverted) Extraversion item "Is less active than other people."

Lower-Educated Subsample

To investigate whether differences in psychometric quality between the two versions were more pronounced for lower-educated persons, who are considered more sensitive to complex item formulations, we repeated our analyses for the subsample of lower-educated respondents. Overall, the results (displayed in detail on our OSF page) show a highly similar picture to that of the full sample: Reliability coefficients between the two versions did not differ markedly; criterion validity was broadly similar across the two versions, with slightly lower associations for the simplified version; and the factorial validity indicated a better replication of the intended multidimensional structure for the original BFI-2 version compared with the simplified version. Thus, even among lower-educated (and, on average less literate) respondents among whom the simplified item sets were especially intended to improve the psychometric quality of the measures, the simplified items did not outperform the original items.

Conclusion

Is it really a "sin" to use complexly worded and double-barreled items (e.g., Gehlbach, 2015)? The present study aimed to systematically investigate whether simplifying item wording and avoiding multiple stimuli markedly increases the psychometric quality of personality inventories. We directly compared the original BFI-2 version with a simplified one. Our results clearly show that in no case did the simplified versions outperform the original versions. In fact, our factorial results indicate that the intended multidimensional structure of the questionnaire was better replicated by the original items than by the simplified versions. These results also held for lower-educated persons, considered more sensitive to linguistically complex items and for whom the simplified items were primarily intended (e.g., Knauper et al., 1997; Smith, 1982).

Our findings thus question general rules of item formulation and suggest that established personality items do not benefit from linguistic simplification and reduction to single stimuli. Personality items are usually carefully developed with a focus on simple phrasing (see, e.g., Soto & John, 2017). Multiple stimuli – representing similar nuances of the same personality aspect – are intentionally included in these phrases to enhance the bandwidth and sharpen the item intention (see, e.g., Gosling et al., 2003). Our results indicate that purely linguistically focused simplifications do not enhance the quality of these items but rather may blur their unique meaning, thereby decreasing their validity. Further, it might be oversimplified to categorize each item containing multiple stimuli as double-barreled. Instead, one might regard the double-barreledness of items as a continuum from very similar stimuli (e.g., synonyms or paraphrases) to stimuli that can be interpreted as even contradictory. In the case of BFI-2, the items are clearly on the synonym pole of this dimension. They are thus probably less confusing to respondents than items containing contradictory stimuli and instead help to clarify the item's intention. Future studies should investigate the degree to which our results generalize at the other pole of this continuum of double-barreledness.

Additionally, the classical indicators of psychometric quality used in the present study may not have been sensitive enough to identify problems of complex items. Some previous studies have indicated that item nonresponse and response duration are affected by item complexity (e.g. Lenzner, 2012). Future studies should therefore investigate a parsimonious set of different quality indicators to compare different item formulations. This study focused on the scale level and used classical test theory (CTT) methods. Future studies may garner additional insights into the performance of individual items by using item response theory (IRT) models, in particular four-parametric-logistic models suggested by Waller and Reise (2010).

Overall, our results indicate that an established personality inventory does not benefit from item simplification. As such, the results may refine traditional item formulation guidelines and encourage future studies research to probe the generalizability of our findings.

References

Bassili, J. N., & Scott, B. S. (1996). Response latency as a signal to question problems in survey research. *Public Opinion Quarterly, 60*(3), 390–399. https://doi.org/10.1086/297760

Chen, F. F. (2007). Sensitivity of goodness of fit indexes to lack of measurement invariance. *Structural Equation Modeling: A Multidisciplinary Journal, 14*(3), 464–504. https://doi.org/10.1080/10705510701301834

Danner, D., Rammstedt, B., Bluemke, M., Lechner, C., Berres, S., Knopf, T., Soto, C., & John, O. (2019). Das Big Five Inventar 2: Validierung eines Persönlichkeitsinventars zur Erfassung von 5 Persönlichkeitsdomänen und 15 Facetten [The German Big Five Inventory 2: Measuring five personality domains and 15 facets]. *Diagnostica, 65*(3), 1–12. https://doi.org/10.1026/0012-1924/a000218

Elson, M. (2016). Question wording and item formulation. In J. Matthes, R. Potter, & C. S. Davis (Eds.), *International encyclopedia of communication research methods* (pp. 1–8). Wiley-Blackwell. https://doi.org/10.1002/9781118901731.iecrm0200

Flesch, R. (1948). A new readability yardstick. *Journal of Applied Psychology, 32*(3), 221–233. https://doi.org/10.1037/h0057532

Gehlbach, H. (2015). Seven survey sins. *The Journal of Early Adolescence, 35*(5–6), 883–897. https://doi.org/10.1177/0272431615578276

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*(6), 504–528. https://doi.org/10.1016/S0092-6566(03)00046-1

Grant Levy, S. (2018). Deconstructing a double-barreled alternative: Evolution and creationism. *Psychological Reports, 122*(5), 1995–2004. https://doi.org/10.1177/0033294118795145

Herzberg, P. Y., & Brähler, E. (2006). Assessing the Big-Five personality domains via short forms: A cautionary note and a proposal. *European Journal of Psychological Assessment, 22*(22), 139–148. https://doi.org/10.1027/1015-5759.22.3.139

Hox, J. J., & Borgers, N. (2001). Item nonresponse in questionnaire research with children. *Journal of Official Statistics, 17*(2), 321–335. https://hdl.handle.net/1874/23617

Kahneman, D. (2011). *Thinking, fast and slow*. Farrar, Straus and Giroux.

Kelley, K. (2018). *MBESS: The MBESS R package* (Version 4.4.3) [R package]. https://CRAN.R-project.org/package=MBESS

Kieffer, M. J., Lesaux, N. K., Rivera, M., & Francis, D. J. (2009). Accommodations for English language learners taking large-scale assessments: A meta-analysis on effectiveness and validity. *Review of Educational Research, 79*(3), 1168–1201. https://doi.org/10.3102/0034654309332490

Knauper, B., Belli, R. B., Hill, D. H., & Herzog, A. R. (1997). Question difficulty and respondents' cognitive ability: The impact on data quality. *Journal of Official Statistics, 13*(2), 181–199. https://www.psc.isr.umich.edu/pubs/abs/1046

Krosnick, J. A. (1991). Response strategies for coping with the cognitive demands of attitude measures in surveys. *Applied Cognitive Psychology, 5*(3), 213–236. https://doi.org/10.1002/acp.2350050305

Lechner, C. M., Anger, S., & Rammstedt, B. (2019). Socio-emotional skills in education and beyond: Recent evidence and future research avenues. In R. Becker (Ed.), *Research handbook on the sociology of education* (pp. 427–453). Edward Elgar Publishing. https://doi.org/10.4337/9781788110426.00034

Lenzner, T. (2012). Effects of survey question comprehensibility on response quality. *Field Methods, 24*(4), 409–428. https://doi.org/10.1177/1525822X12448166

Lenzner, T., Kaczmirek, L., & Lenzner, A. (2010). Cognitive burden of survey questions and response times: A psycholinguistic experiment. *Applied Cognitive Psychology, 24*(7), 1003–1020. https://doi.org/10.1002/acp.1602

McCrae, R. R., Costa, J. P. T., & Martin, T. A. (2005). The NEO-PI-3: A more readable revised NEO Personality Inventory. *Journal of Personality Assessment, 84*(3), 261–270. https://doi.org/10.1207/s15327752jpa8403_05

Menold, N. (2020). Double barreled questions: An analysis of the similarity of elements and effects on measurement quality. *Journal of Official Statistics, 36*(4), 855–886. https://doi.org/10.2478/jos-2020-0041

Menold, N., & Raykov, T. (2022). On the relationship between item stem formulation and criterion validity of multiple-component measuring instruments. *Educational and Psychological Measurement, 82*(2), 356–375. https://doi.org/10.1177/0013164420988169

Messick, S. (1995). Validity of psychological assessment: Validation of inferences from persons' responses and performances as scientific inquiry into score meaning. *American Psychologist, 50*(9), 741–749. https://doi.org/10.1037/0003-066X.50.9.741

Nießen, D., Groskurth, K., Rammstedt, B., & Lechner, C. (2020). *An English-language adaptation of the General Life Satisfaction Short Scale (L-1)*. GESIS: Zusammenstellung sozialwissenschaftlicher Items und Skalen (ZIS). https://doi.org/10.6102/zis284

Organisation for Economic Co-operation and Development (OECD). (2017, November 6). *A module on adults' social and emotional skills: Proposal for the 2nd cycle of PIAAC*. 19th meeting of the PIAAC Board of Participating Countries, 20–21 November 2017, Singapore (COM/DELSA/EDU/PIAAC(2017)11)

Organisation for Economic Co-operation and Development (OECD). (2018). *Programme for the International Assessment of Adult Competencies (PIAAC), English Pilot Study on Non-Cognitive Skills* (ZA6940; Version 1.0.0). GESIS Data Archive. https://doi.org/10.4232/1.13062

Pargent, F., Hilbert, S., Eichhorn, K., & Bühner, M. (2019). Can't make it better nor worse. *European Journal of Psychological Assessment, 35*(6), 891–899. https://doi.org/10.1027/1015-5759/a000471

Patrick, C. J., Kramer, M. D., Tellegen, A., Verona, E., & Kaemmer, B. A. (2012). Development and preliminary validation of a simplified-wording form of the Multidimensional Personality Questionnaire. *Assessment, 20*(4), 405–418. https://doi.org/10.1177/1073191112467051

Porst, R. (2014). Question Wording – Zur Formulierung von Fragebogen-Fragen [Question Wording – On the formulation of questionnaire items]. In R. Forst (Ed.), *Fragebogen* (pp. 99–118). Springer VS. https://doi.org/10.1007/978-3-658-02118-4_7

Rammstedt, B., Lechner, C., & Danner, D. (in press). *Beyond literacy: The incremental value of non-cognitive skills*. *OECD Working Paper*. OECD.

Roemer, L., Rammstedt, B., & Lechner, C. M. (2022). *Don't keep it too simple*. https://osf.io/atfsv

Schult, J., Schneider, R., & Sparfeldt, J. (2019). Assessing personality with multi-descriptor items: More harm than good? *European Journal of Psychological Assessment, 35*(1), 117–125. https://doi.org/10.1027/1015-5759/a000368

Smith, T. (1982). Educated don't knows: An analysis of the relationship between education and item nonresponse. *Political Methodology, 8*(3), 47–57.

Soto, C. J., & John, O. P. (2017). The next Big Five Inventory (BFI-2): Developing and assessing a hierarchical model with 15 facets to enhance bandwidth, fidelity, and predictive power. *Journal of Personality and Social Psychology, 113*(1), 117–143. https://doi.org/10.1037/pspp0000096

Waller, N. G., & Reise, S. P. (2010). Measuring psychopathology with nonstandard item response theory models: Fitting the four-parameter model to the Minnesota Multiphasic Personality Inventory. In S. E. Embretson (Ed.), *Measuring psychological constructs: Advances in model-based approaches* (pp. 147–173). American Psychological Association. https://doi.org/10.1037/12074-007

History
Received January 19, 2022
Revision received May 20, 2022
Accepted August 3, 2022
Published online December 16, 2022
EJPA Section / Category Methodological topics in assessment

Open Science

Open Data: We confirm that there is sufficient information for an independent researcher to reproduce all of the reported results, including codebook if relevant (OECD, 2018).

Open Materials: We confirm that there is sufficient information for an independent researcher to reproduce all of the reported methodology (https://osf.io/atfsv/, Roemer et al., 2022).

Preregistration of Studies and Analysis Plans: This study was not preregistered.

Funding
Open access publication enabled by GESIS Leibniz Institute for the Social Sciences.

ORCID
Beatrice Rammstedt
https://orcid.org/0000-0002-6941-8507

Beatrice Rammstedt
GESIS – Leibniz Institute for the Social Sciences
PO Box 12 21 55
68072 Mannheim
Germany
beatrice.rammstedt@gesis.org