Hunting Game: Targeting the Big Five

renebekkers.wordpress.com/2017/03/21/hunting-game-targeting-the-big-five/

3/21/2017

Do not use the personality items included in the World Values Survey. That is the recommendation of Steven Ludeke and Erik Gahner Larsen in a recent paper published in the journal *Personality and Individual Differences*. The journal is owned by Elsevier so the official publication is paywalled. Still I am writing about it because the message of the paper is extremely important. Ludeke and Gahner Larsen formulate their recommendation a little more subtle: “we suggest it is thus hard to justify the use of this data in future research.”

What went wrong here? Join me in a hunting game, targeting the Big Five.

The World Values Survey (WVS) is the largest, non-commercial survey in the world. It is frequently used in social science research. The most recent edition contained a short, 10 item measure of personality characteristics (BFI-10), validated in a well-cited paper by Rammstedt and John in the Journal of Research in Personality. The inclusion of the BFI-10 enables researchers to study how the Big Five personality traits is related to political participation, happiness, education, and health, among many other things.

So what is wrong with the personality data in the WVS? Ludeke and Gahner Larsen found that the pairs of adjectives designed to measure the five personality traits Openness, Conscientiousness, Extraversion, Agreeableness, Neuroticism are not correlated as expected. To measure openness, for instance, the survey asked participants to indicate agreement with the statement “I see myself as someone who: has few artistic interests” and “I see myself as someone who: has an active imagination”. One would expect a negative relation between the responses to the two statements. However, the correlation between the two items across all countries is positive, r = .164. This correlation is not strong, but in the wrong direction. Similar discrepancies were found between items designed to measure the four other dimensions of personality.

The BFI-10 included in the WVS is this set of statements (an r indicates a reverse-scored item):

I see myself as someone who:

- is reserved (E1r)
- is generally trusting (A1)
- tends to be lazy (C1r)
- is relaxed, handles stress well (N1r)
- has few artistic interests (O1r)
- is outgoing, sociable (E2)
- tends to find fault with others (A2r)
- does a thorough job (C2)
- gets nervous easily (N2)
- has an active imagination (O2)

In a factor analysis of the 1o items, we would expect to find the five dimensions. However, that is not the result of an exploratory factor analysis applying the conventional criterion of an Eigen value > 1. In this analysis and all following analyses negative items are reverse scored. Including all countries, a three factor solution emerges that is very difficult to interpret. Multiple items show high loadings on multiple factors. Removing these one by one, as is usually done in inventories with large numbers of items, we are left with a two-factor solution. If a five-factor solution is forced, we obtain the following component matrix. This is a mess.

	Component				
	1	2	3	4	5
O1 not artistic (r)	-.116	-.054	.105	-.049	**.961**
O2 active imagination	**.687**	.162	-.031	.197	-.140
C1 lazy (r)	.249	-.004	**.836**	-.045	.159
C2 thorough	**.640**	**.425**	.231	.078	.071
E1 reserved (r)	-.110	**-.825**	-.022	-.183	-.047
E2 outgoing	**.781**	.097	-.004	-.105	-.068
A1 trusting	.210	**.722**	.003	-.160	-.137
A2 fault with others (r)	**-.430**	.079	.614	-.259	-.051
N1 relaxed (r)	**-.461**	**-.377**	.235	**.534**	.144
N2 nervous	.188	.133	-.291	**.770**	-.112

So what is wrong with these data?

Upon closer inspection, Ludeke and Gahner Larsen found that the correlations were markedly different across countries. Bahrain is a clear outlier. The weakly positive correlation between O1 and O2r is due in part to the inclusion of data from Bahrain. Without this country, the correlation is only .135. Still positive, but not as strongly. The data for Bahrain are not only strange for openness, but also for other factors. In the table below I have computed the correlations among recoded items for the five dimensions.

Without Bahrain, the correlations are still strange, but a little less strange.

	O	C	E	A	N
With Bahrain	-.164	.238	-.207	-.036	.008
Without Bahrain	-.135	.275	-.181	-.009	.044

What is wrong with the data for Bahrain? The patterns of responses for cases from Bahrain, it turns out, are surprisingly often a series of ten exactly the same values, such as 1111111111 or 555555555555. I routinely check data from surveys for such patterns. While it is impossible to prove this, serial response patterns suggest fabrication of data. Participants and/or interviewers skipping questions may follow such patterns. Almost half of all the cases from Bahrain follow such a pattern. Other countries with a relatively high proportion of serial pattern responses are South Africa, Singapore, and China. The two countries for which the BFI-10 behaves close to what previous research has reported, the Netherlands and Germany, have a very low occurrence of serial pattern responses.

	Number of serial pattern responses	%
Bahrain	598	49.83%
South Africa	250	7.08%
Singapore	108	5.48%

China	52		2.26%
Netherlands	8		0.42%
Germany	2		0.10%

Even without the data for Bahrain and the serial responses from all other countries, however, the factor structure is err…not what one would expect. Still a mess.

	Component				
	1	2	3	4	5
O1 not artistic (r)	-.094	-.040	.086	-.031	**.968**
O2 active imagination	.**691**	.150	-.046	.158	-.130
C1 lazy (r)	.297	.023	**.815**	-.017	.146
C2 thorough	**.637**	**.410**	.241	.050	.088
E1 reserved (r)	-.098	**-.828**	-.033	-.158	-.058
E2 outgoing	**.771**	.070	-.001	-.140	-.052
A1 trusting	.192	**.710**	.022	-.190	-.133
A2 fault with others (r)	**-.405**	.080	**.628**	-.230	-.048
N1 relaxed (r)	**-.421**	**-.352**	.218	.**592**	.123
N2 nervous	.192	.133	**-.315**	**.750**	-.104

Only for Germany and the Netherlands the factor structure is somewhat in line with previous research. Here is the solution for the two countries combined. In both countries, the two statements for agreeableness do not correlate as expected. Also the second statement for conscientiousness (thorough) has a cross-loading with one of the agreeableness items (trusting).

	Component				
	1	2	3	4	5
O1 not artistic (r)	-.047	-.056	**.842**	.120	-.089
O2 active imagination	.208	.050	**.729**	-.140	.173
C1 lazy (r)	.061	-.083	-.040	**.865**	-.087
C2 thorough	-.064	.053	.057	**.627**	**.440**
E1 reserved (r)	**.715**	-.113	.130	.032	-.219
E2 outgoing	**.732**	-.166	.126	.166	.210
A1 trusting	-.008	-.100	.042	.049	**.853**
A2 fault with others (r)	**-.657**	-.272	.090	.177	-.001

N1 relaxed (r)	.012	**.804**	-.002	.116	-.259
N2 nervous	-.052	**.835**	-.006	-.160	.117

This leaves us with three possibilities.

One possibility was raised by Christopher Soto on Twitter: acquiescence bias could be driving the results. In a study using data from another multi-country survey in the International Social Survey Program (ISSP), Rammstedt, Kemper & Borg subtracted each respondent's mean response across all BFI-10 items from his or her score on each single item. Doing this, however, does not clear the sky. Looking again at the correlations for the pairs of items measuring the same constructs, we see that they are not 'better' in the second row. In contrast, they are less positive.

	O	C	E	A	N
Unadjusted	-.122	.286	-.166	.001	.053
Attenuated	-.310	.078	-.235	-.107	.049

Also the factor structure of the attenuated scores is not anything like the 'regular' five-factor structure. Still a mess.

	Component				
	1	2	3	4	5
O1a	-.192	-.025	.096	-.117	**-.957**
O2a	**.509**	.190	**-.319**	.034	.269
C1a	-.133	**.469**	**.617**	**-.460**	.174
C2a	**.351**	**.681**	-.005	.050	.071
E1a	-.043	**-.846**	.080	-.250	.017
E2a	**.823**	.029	.034	.045	.114
A1a	.086	.285	.026	**.821**	.148
A2a	**-.497**	-.246	**.555**	.274	.067
N1a	**-.598**	**-.345**	-.223	**-.345**	-.047
N2a	-.123	.043	**-.854**	-.031	.178

The second possibility is that things went wrong in the translation of the questionnaire. The same adjectives or statements may mean different things in different countries or languages, which makes them useless as operationalizations of the same underlying construct. It will require a detailed study of the translations to see if anything went wrong. The questionnaires are available at the World Values Survey website. The Dutch questionnaire is good. I looked at a few other languages. The Spanish questionnaire for Ecuador also seems right. "Me veo como alguien que…… es confiable" is quite close to "I see myself as someone who is… generally trusting". My Spanish is not very good though. Rene Gempp wrote on Twitter that the BFI-10 is a Likert-type scale, but the Spanish translation asks about the frequency, and one of the options, "para nada frecuentemente" is *very* confusing in Spanish.

I am not sure about your fluency in Kinyarwanda, the language spoken in Rwanda, but the backtranslation of the questionnaire in English does not give me much confidence. Apparently, "…wizera muri rusange" is the translation of "is generally trusting". The backtranslation is "…believe in congregation".

V160A-J. I am going to read for you some things and later tell me how you can show similarity or resemblance
[READ HARD AND PUT A CODE TO ONE ANSWER ON EVERYTHING]:

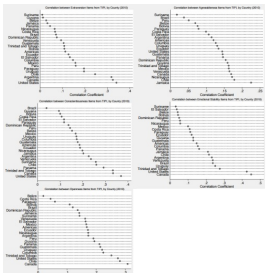
I see myself as a person …

		I do not agree at all	Disagree	I am holding to	Disagree	Totally agree	I do not know
V160A.	Queite	1	2	3	4	5	9
V160B.	…believe in congregation	1	2	3	4	5	9
V160C.	…do you want to be weak	1	2	3	4	5	9
V160D.	…meek, patient in problems	1	2	3	4	5	9
V160E.	…not mixed up	1	2	3	4	5	9
V160F.	…has friends, associates	1	2	3	4	5	9
V160G.	…who sees mistakes on others	1	2	3	4	5	9
V160H.	…does a finished job	1	2	3	4	5	9
V160I.	…who waries alot	1	2	3	4	5	9
V160J.	…thinks alot	1	2	3	4	5	9

The third possibility is that personality structure may indeed be different in different countries. This would be the most problematic one.

Data from the 2010 AmericasBarometer Study, conducted by the Latin American Public Opinion Project (LAPOP) support this interpretation. The survey included a different short form of the Big Five, the TIPI, developed by Gosling, Rentfrow, and Swann. A recent study by Weinschenk published in *Social Science Quarterly* shows that personality scores based on the TIPI are hardly related to turnout in elections in the Americas. This result may be logical in countries where voting is mandatory, such as Brazil. But the more disconcerting methodological problem is that the Big Five are not reliably measured with pairs of statements in most of the countries included in the survey. Here are the correlations between the pairs of items for each of the five dimensions, taken from the supplementary online materials of the Weinschenk paper.

The graphs show that the TIPI items only work well in the US and Canada – the two 'WEIRD' countries in the study. In Brazil, to take one example, the correlations are <.10 for extraversion, agreeableness and conscientiousness, and lower than .25 for emotional stability and openness.



Back to the WVS case, which raises important questions about the peer review process. Two journal articles based on the WVS (here and here) were able to pass peer review because neither the reviewers nor the editors asked questions about the reliability of the items being used. Neither did the authors check, apparently. Obviously, researchers should check the reliability of measures they use in an analysis. In case authors fail to check this, reviewers and editors should ask. Weinschenk reported the low correlations in the online supplementary materials, but did not report reliability coefficients in the paper.

The good thing is that because the WVS is in the public domain, these problems came to light relatively quickly. Of course, they could have been avoided if the WVS had scrutinized the reliability of the measure before putting the data online, if the authors of the papers using the data had checked the reliability of the items or if the reviewers and editors had asked the right questions. Another good thing is that the people at the WVS (volunteers?) at the WVS twitter account have been frank in tweeting about the problems found in the data.

Summing up:

1. We still do not know why the BFI-10 measure of the Big Five personality does not perform as in previous research.

2. It is probably not due to acquiescence bias. Translations may be problematic for some countries.
3. Do not use the WVS BFI-10 data from countries other than Germany and the Netherlands.
4. Treat the WVS data from Bahrain and with great caution, and to be on the safe side, just exclude it from your analyses.
5. The reliability of short Big Five measures is very low in non-WEIRD countries.

The code for the analyses reported in this blog is posted at the Open Science Framework.

Update 22 March 2017. The factor loadings in the table with the results of the analysis of attenuated scores has been updated. The table displayed previously was based on a division of the original scores by the total agreement scores. Rammstedt et al. subtracted the original scores from the total agreement scores. The results of the new analysis are close to the previous one and still confusing. The code on the OSF has been updated. Also a clarification was added that the negative items used in the factor analyses were all recoded such that they scored positively (HT to Christopher Soto).