11 TOWARDS MEANINGFUL COMPARISONS OF PERSONALITY IN LARGE-SCALE CROSS-CULTURAL STUDIES

GWEN GARDINER, KYLE SAUERBERGER,
MEMBERS OF THE INTERNATIONAL SITUATIONS PROJECT,
AND DAVID FUNDER

One of Jüri Allik's major, pioneering contributions to psychology is the assessment of personality across numerous cultures. His contributions have inspired many other large collaborations of international researchers to move beyond early work confirming the Five Factor Model cross-culturally to assessing the reliability and validity of a broad range of personality traits. Cross-cultural comparisons of personality traits may be problematic if measures have unique meanings in different cultural contexts that influence how individuals respond to items. In this chapter we present a new and relatively simple method for assessing the comparability of measures in large-scale cross-cultural studies, and illustrate the method using responses to the Big Five Inventory-2 (BFI-2) from 15,368 participants in 63 countries participating in the International Situations Project.

Introduction

The recent growth in cross-cultural research has brought with it an expansion of the study of personality across cultures, particularly with large collaborations of researchers accumulating data across numerous cultural groups. Jüri Allik has been a pioneering participant and leader in this effort, and his contributions to the understanding of personality across cultures is one—just one—of his significant career accomplishments (e.g., Allik & McCrae, 2004; Mõttus, Allik, & Realo, 2010; Schmitt et al., 2007).

Initially, most cross-cultural research on personality focused on testing if the Five Factor Model (FFM) of personality was reproducible in samples outside of the

The present research was supported by the US National Science Foundation under Grant BCS-1528131. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the individual researchers and do not necessarily reflect the views of the National Science Foundation. Data gathering in the Czech Republic was supported by grant 17-14387S by the Czech Science Foundation and by institutional research funding RVO: 68081740 from the Institute of Psychology, Czech Academy of Sciences.

Members of the International Situations Project contributed the data used in the analyses reported in this chapter. Their names and affiliations are listed in Appendix 11A.

E-mail: ggard001@ucr.edu, kylesauerberger@gmail.com, and funder@ucr.edu

Western world (with the answer being, "generally, yes"). Later, researchers expanded the research question to include the reliability or accuracy of personality profiles of cultures. This has led to the issue of whether measures can be compared across cultural groups, who may have unique interpretations of the items in the measures. Unique cultural interpretations of items could bias results and limit the conclusions that can be drawn from the data (see, e.g., Allik & Realo, 2017).

Various methodological approaches have been suggested and used in an attempt to detect and (perhaps) correct for cultural biases in responses to measurement instruments and the study of "measurement invariance" has become a complex and daunting statistical issue (e.g., van de Schoot, Lugtig, & Hox, 2012). In the present chapter we suggest and demonstrate a new and relatively simple approach to assessing the comparability of measures in large-scale cross-cultural studies.

The Problem of Cross-Cultural Comparability

Early work on assessing personality around the world typically tested the generalizability of the FFM in one or two non-Western societies (e.g., Gurven, von Rueden, Massenkoff, Kaplan, & Lero Vie, 2013) or compared personality trait relationships and behavioral expressions among a handful of diverse nations (e.g., Ching et al., 2014). While each individual study provides unique contributions, the most informative studies are those that assess a wide range of cultures (Allik & Realo, 2017). A large sample of cultures is more informative in the same way a large sample of individuals is more informative. The large sample of cultures will exhibit a wider range of traits and be more representative of the larger population. Additionally, researchers interested in the reliability of country trait profiles need large, independent samples with enough overlapping cultures to test the replicability of previous findings (Allik & Realo, 2017). The number of large-scale cross-cultural research projects will continue to grow in the coming years as more researchers form international collaborations and technological access expands around the world allowing for easier data collection in more diverse nations.

One crucial aspect to cross-cultural research is assessing the comparability of the measures used across a range of diverse cultural groups. Typical questionnaires used to measure how a specific construct varies across cultures may inadvertently assess other cultural characteristics in relation to responding to the questionnaire itself. For example, a tendency to always choose the most extreme responses on a Likert scale biases the overall score on that measure. Response styles to questionnaires have been linked to cultural dimensions, implying that any cultural differences found in questionnaire results are partially due to cultural differences in responding to surveys (Harzing, 2006). Additionally, in cross-cultural studies researchers typically translate existing measures into the native language of the assessment group. Items that are mistranslated or represent a distinct cultural construct that is not universal will also bias the overall results from measures (Chen, 2008).

Given the range of potential sources of bias in the data, researchers have developed methods for testing the comparability of measures across groups. Typically, the factor model of a measure is compared between a reference group and a comparison group (Byrne & van de Vijver, 2010). The comparability of a measure is determined by the fit statistics of the model, often using seemingly arbitrary thresholds for determining "good fit." This method is problematic for researchers interested in understanding the nuances in potential cultural biases in the data, because it provides only a single overall measure of fit, without indicating clearly which items on measures are the source of convergence and difference without further testing. Additionally, the traditional method of comparing each new cultural group to a reference group, often the United States (US), becomes exponentially difficult as the number of countries grows and more comparisons are needed (Byrne & van de Vijver, 2010).

Large-scale Cross-Cultural Assessments of Personality

While the number of large-scale cross-cultural assessments of personality is growing but still small, the range of methods used to test if meaningful comparisons can be made across cultures is wide. Formal statistical models, although available, are also difficult to understand and use, and their application to actual cross-cultural data remains rare (although see Zecca et al., 2013 for an exception). Instead, the most common method is to compare the country level trait scores with previously collected country trait scores, and also with other country level data. Convergence across samples and associations with independently-measured country-level measures (such as demographic or economic development information) implies that the variation in personality scores across cultures is meaningful (Mõttus et al., 2010). While some external country level predictors of aggregated personality traits are surprising (e.g., Heine, Buchtel, & Norenzayan, 2008), it is probably still too soon to determine the validity of this method (Mõttus et al., 2010).

McCrae, Terracciano and colleagues (2005) were among of the first researchers to collect data on personality traits across a wide range of countries that had been previously assessed, allowing the replicability of results to be examined. Previous cross-cultural comparisons of personality traits had involved secondary data analysis accumulated from multiple independent research projects which, while maximizing the number of countries that could be compared, limited the degree to which the findings across cultures could be considered directly comparable. In an important advance over that approach, McCrae and colleagues (2005) assessed personality traits using the NEO PI-R in 50 cultures by asking college students to rate the personality of someone they knew well. Observer reports were used to limit biases inherent in selfreports and potentially expand the representativeness of the sample beyond traditional college students. Because McCrae and colleagues (2005) was one of the first large-scale assessments of personality, using the same measure in different cultures, the researchers were also one of the first to attempt to assess the comparability of their measure across numerous cultural groups. First, the researchers pooled all the data together and tested the Big Five factor structure using confirmatory factor analysis (CFA). Then, they used Procrustes rotation to compare the factor structure of each culture with the US as a reference group and found evidence for comparability across the groups, with some exceptions in the African countries.

Schmitt, Allik, and colleagues (2007) followed a similar method of testing the comparability of their country level traits score, this time assessed using the 44-item Big Five Inventory (BFI; John & Srivastava, 1999). The factor structure of the BFI was first examined in the total sample of the study and the authors found good fit to the data. To test for cultural differences in the factor structure, the countries were grouped into 10 regions that were then compared with the US as a reference group using Procrustes rotation. Overall, the researchers found evidence for good congruence. The large number of countries overlapping between Schmitt and colleagues (2007) and McCrae and colleagues (2005) allowed for the reliability of country level trait scores to be assessed using different measures. The correlations of personality traits between samples was positive for all traits but only statistically significant at $p < .05$ for extraversion, conscientiousness, and neuroticism. The moderate evidence found for the reliability of the country trait scores strengthened the argument that the cross-cultural variation in personality measures assesses something meaningful, rather than random noise.

Along the same lines as Schmitt et al. (2007), Bartram (2013) assessed the accuracy of personality trait measures of countries by correlating them with findings from previous studies and with other country-level variables. Once again, a different measure of personality was used, providing more evidence for convergent validity of the trait averages. The Occupational Personality Questionnaire (OPQ32) is a personality assessment questionnaire used for studies in the workforce and was tested in 31 countries. Items from the measure were selected to represent the Big Five traits. The OPQ32 is a forced choice assessment in which participants must choose from a list of 4 characteristics an item that is most like them and an item that is least like them. Forced-choice measures are especially useful in cross-cultural comparisons because they can decrease the effects of response styles, a tendency to bias results that are linked to some cultural aspects (Harzing, 2006). However, forced-choice measures can become problematic for traditional statistical tests of equivalence that assume item independence, which may be one reason no formal tests of equivalence were reported (Bartram, 2013).

Thus far, cross-cultural assessments of personality using a large number of cultural groups have largely focused on confirming the factor structure within each group and the convergence of scores with previous, independent assessments of the same construct. Thalmayer and Saucier (2014) assessed the QB6, a measure of the Big Six that can be reduced to the Big Five, across 26 countries. The countries were separated into three groups which were used to independently verify the factor model. Using "domain specific" fit statistics thresholds for multivariate measures derived from Hopwood and Donnellan (2010), the researchers found good model fit for the factor structure and item loadings for both the Big Five and the Big Six. However, even with the lower domain specific thresholds, removing problematic items, and excluding countries, the researchers still did not have enough evidence for equality in variable intercepts, a step usually considered necessarily for comparing means across groups. The researchers subsequently cautioned against group mean comparisons and did not report any trait scores for the countries assessed (Thalmayer & Saucier, 2014).

In sum, current methods for testing the accuracy of personality trait scores at the country level have been quite limited. The most common method is to compare newly assessed country trait scores with previously collected country trait scores to determine the reliability and validity of the findings. The few attempts at more formal methods have found evidence for the comparability of the measures across groups when using simplified methods for testing the factor structure (e.g., McCrae et al., 2005; Schmitt et al., 2007) and limited evidence when tested with more traditional psychometric methods (e.g., Thalmayer & Saucier, 2014). Recommendations for modifying existing methods for a large number of groups are labor intensive and lack the ability to compare numerous cultures to each other, rather than solely to one reference group. Therefore, a new, simpler approach might be worth trying, one that does not incorporate strict or arbitrary statistical thresholds for success while still allowing researchers flexibility for discovering potentially problematic items or cultural groups in their data.

The Comparability of Measures Using an Inter-item Correlation Matrix

A critical concern for researchers interested in knowing whether or not a measure has comparable meaning across cultural groups is the degree to which the items on the measure are understood the same way. Only to the degree that items on the measure have the same meaning to the individuals who respond to them can we infer that different responses to the items reflect differences in the construct the researcher is trying to assess. The key idea underlying the method proposed in this chapter is simply this: The meaning of each item on a psychological measurement instrument can be conceptualized in terms of its relationships with the other items in the measure.

This approach to item meaning is analogous to how words are defined in a dictionary—each word is defined using other words in the dictionary, which in turn are defined by using still other words in the dictionary. The underlying assumption is that the meaning of a word is fully contained in, and reflected by, its relation to other words. Analogously, the meaning of an item on a self-report scale, especially one with a large number of items, could be assumed to be reflected in its relationships to the other items in the scale. A complete item-by-item correlation matrix, then, could be taken to reflect the meaning of each item in terms of its relationships with all of the others, and the overall pattern of correlations to reflect the meaning of the measure as whole.[1]

Therefore, one possible method for assessing the degree to which participants from different countries infer similar meaning from the items on a scale is by calculating the relationships between each item and every other item within each country. Each country will have its own resulting matrix of inter-item correlations that can then be correlated with the inter-item correlation matrix of every other country. The resulting correlation between any two countries (which is simply the vector correlation between the two sets of non-redundant inter-item correlations) represents how similarly participants in the two countries interpret each item in relation to every other item. In a study of many countries, this approach can be expanded to produce a country-by-country matrix that reveals how similar the pattern of inter-item correlations is between any two countries in the sample, how similar the pattern within any given country is to the average pattern of other countries, and how similar the pattern of item meaning is overall, across the world.

Comparing inter-item correlation matrices has several possible benefits over traditional methods of testing for the comparability of measures across cultures. First, it is simple and transparent. Compare this method to the one illustrated by Davidov, Schmidt, and Schwartz (2008). Their sophisticated approach began with computing a confirmatory factor analysis (CFA) within each country in their sample (20 countries), attempting to derive a factor structure adequate to describe all of the countries' response patterns, then following up with a multigroup confirmatory factor analysis (MGCFA) to assess the degree to which this attempt was successful. We suggest that our method is a much simpler and more transparent way to assess configural invariance. A second advantage is that our method clearly shows the degree to which each country is similar (or dissimilar) in its configural structure to each other country, and also, the degree which it is similar and dissimilar to other countries overall—information which the conventional MGCFA does not so readily provide.

Thus, researchers can compare all countries with each other, rather than every country with a single reference country, such as (most often) the US. This capability allows researchers to see if countries that have a lower correlation with the US also have a lower correlation with many other countries, indicating random error in the data, or if they are more similar to other culturally comparable countries,

[1] Conventional factor analytic methods are rooted in this item-by-item matrix and derive all of their information from it, but focus on latent factors or other multi-variate constructs that emerge, rather than the matrix itself.

implying a cultural bias in the data. Inter-item correlation matrices also can work well for multivariate measures, such as Big Five personality measures, which can be problematic for traditional methods of comparison (Hopwood & Donnellan, 2010). The simple method of matrix comparison is also useful even if a measure does not have any strong latent variables or has excess items that do not correspond to specific constructs, because it is the meaning of each item that is assessed, rather than latent variables that may or may not be culturally relevant for all groups tested. Lastly, this method is easier to conduct than traditional methods that require expensive software or advanced statistical knowledge to perform and understand, which can and we suspect does often limit the use of these methods in the field.[2] Here we present an example of this new method using personality data collected as part of a large-scale international research project.

Method

Participants

The International Situations Project (ISP) is a large, international collaboration involving over 130 researchers representing 63 countries and 40 languages (see Table 11.1). Participants (*N* = 15,368) were recruited by collaborators at their local university to answer a survey online that included several measures of personality, values, and situational experience. All measures were first translated into the local language and then back-translated by an independent source. The back-translation and original English were compared, and any discrepancies resolved.

Measures

Personality was assessed using the Big Five Inventory-2 (BFI-2; Soto & John, 2017). The BFI-2 consists of 60 items that measure the Big Five traits—Extraversion, Agreeableness, Conscientiousness, Negative Emotionality, and Open-Mindedness—and 15 facets (three facets nested within each Big Five trait). In the present analyses, we shall focus on the 60 items rather than their subsuming traits or facets.

Results and Discussion

An inter-item correlation matrix was first created for each country by correlating every BFI-2 item with every other BFI-2 item, resulting in 60 x 60 item matrix for each of 63 countries.[3] Then, each country's inter-item correlation matrix was correlated with

[2] The analyses reported in this paper were conducted using the open-source program R (R Core Team, 2017) and required no specialized or proprietary software.

[3] The number of non-redundant correlations in this matrix is (60 x 59)/2, or 1,770, and these are the correlations that enter into the vector correlations that compare each pair of countries.

Table 11.1 Demographic information and sample size by country.

Country	Mean age	Total *N*	% female	Country	Mean age	Total *N*	% female
Argentina	24.83	140	78.85	Mexico	23.88	247	58.37
Australia	19.84	196	76.02	Netherlands	20.13	301	81.33
Austria	21.26	113	81.42	New Zealand	19.19	129	86.05
Belgium	19.14	50	84.00	Nigeria	24.75	135	33.58
Bolivia	21.01	135	57.78	Norway	23.89	159	74.21
Brazil	23.68	310	72.17	Pakistan	20.61	114	50.00
Bulgaria	25.05	152	70.67	Palestine	22.17	295	83.39
Canada	21.86	304	79.14	Peru	28.21	74	58.26
Chile	21.45	386	66.41	Philippines	19.71	337	69.18
China	25.31	432	46.01	Poland	22.35	234	83.33
Colombia	21.68	181	74.03	Portugal	21.66	157	87.82
Croatia	21.46	218	64.68	Romania	22.84	177	57.06
Czech Republic	22.65	193	80.83	Russia	21.92	159	78.48
Denmark	22.94	246	79.92	Senegal	23.32	635	47.48
Estonia	25.88	293	83.96	Serbia	23.57	185	75.85
France	22.60	231	85.53	Singapore	20.93	136	77.94
Georgia	20.29	140	80.00	Slovakia	22.41	148	69.59
Germany	24.49	458	75.70	Slovenia	20.43	123	57.38
Greece	24.09	225	79.22	South Africa	22.21	256	66.67
Hong Kong	19.00	144	59.15	South Korea	22.35	281	58.36
Hungary	25.33	178	66.67	Spain	19.73	419	85.20
India	24.99	221	57.04	Sweden	†	130	72.22
Indonesia	21.85	131	52.71	Switzerland	22.45	755	84.30
Israel	25.35	173	61.40	Taiwan	19.71	162	76.54
Italy	21.86	717	64.57	Thailand	19.24	196	80.32
Japan	22.58	243	61.98	Turkey	21.09	329	68.29
Jordan	19.87	141	80.85	Uganda	22.63	93	64.52
Kenya	21.17	139	65.47	Ukraine	23.91	244	75.79
Latvia	24.87	169	82.84	United Kingdom	25.61	136	88.41
Lithuania	20.26	145	78.47	United States	19.85	1366	67.72
Macedonia	21.22	54	74.07	Vietnam	19.05	168	77.25
Malaysia	21.53	230	71.05	*World sample*	*22.34*	*15,368*	*70.13*

Note: † = Data not available.

every other country's correlation matrix, resulting in a 63 x 63 correlation matrix (please see Table 11.S1)[4]. The resulting correlations between countries represent the degree of similarity in how items are responded to in terms of other items, with higher numbers indicating more similarity between countries. Additionally, the average correlation between each country's inter-item matrix and the matrices of the other countries was calculated, to determine which countries have the greatest overall similarity with other countries (see Table 11.2). Among the participants in the ISP, the country that is most similar to every other country is, unsurprisingly, the US[5] ($r = .80$), followed by Switzerland ($r = .77$), Canada ($r = .77$), Estonia ($r = .77$), and the Philippines ($r = .77$). The least similar countries, meaning countries in which participants interpreted items the most distinctively compared to the other countries, were Macedonia ($r = .46$), Pakistan ($r = .50$), Uganda ($r = .51$), Vietnam ($r = .53$), and Indonesia ($r = .55$). Overall, the average inter-item matrix correlation among countries was $r = .69$.

The overall matrix in Table 11.S1 also allows researchers to easily compare countries that are the most similar and the least similar. In these data, the countries that are most similar to each other are the US and Canada ($r = .91$), the US and the Philippines ($r = .91$), and Germany and Switzerland ($r = .91$). The countries that are least similar to each other are Uganda and Macedonia ($r = .34$), Vietnam and Macedonia ($r = .35$), and Belgium and Uganda ($r = .35$). The low comparability correlations for Macedonia and Belgium might reflect the smaller sample size for those countries.

Researchers can also test if low average correlations for countries are consistent across all countries or vary according to cultural differences. For example, it is possible to test if countries with lower overall correlations have equally low correlations with other culturally similar countries with lower overall countries. This shows whether low correlations are the result of random error in the data or if it reflects some underlying cultural bias. For example, one of the least similar countries overall is Pakistan ($r = .50$). However, the inter-item matrix correlation between Pakistan and India, a geographically and culturally close country, is one of the highest country correlations for Pakistan ($r = .59$). Uganda, another country with a low average correlation but less culturally similar to Pakistan than India, has a relatively low matrix correlation with Pakistan ($r = .42$). However, Uganda has a higher matrix correlation with Kenya ($r = .62$) and Nigeria ($r = .60$), other African countries in the dataset, and a lower matrix correlation with Vietnam ($r = .45$), a country culturally distinct from Uganda. Thus, while overall Pakistan, India, Uganda, Kenya, and Nigeria are all dissimilar to others, Pakistan and India are more similar in their dissimilarity compared with other countries and Uganda, Kenya, and Nigeria are also more similar to each other in their dissimilarities than with other countries.

[4] Table 11.S1 is too large to appear in print, but can be accessed via a Google Sheet at https://goo.gl/rNynoq

[5] The BFI-2 was originally developed in the US.

Table 11.2 Average similarity of inter-item matrix correlations of the BFI-2 items, by country.

Country	Average	Country	Average
United States	0.80	Slovakia	0.71
Canada	0.77	Sweden	0.71
Estonia	0.77	Argentina	0.70
Philippines	0.77	Austria	0.70
Switzerland	0.77	Denmark	0.70
Chile	0.76	France	0.70
Germany	0.76	Hong Kong	0.70
Turkey	0.76	Israel	0.70
Croatia	0.75	New Zealand	0.70
South Africa	0.75	Peru	0.70
Hungary	0.74	Portugal	0.70
Mexico	0.74	Russia	0.70
Serbia	0.74	Jordan	0.69
Spain	0.74	Lithuania	0.69
United Kingdom	0.74	Palestine	0.69
China	0.73	Slovenia	0.69
Italy	0.73	Latvia	0.68
Netherlands	0.73	Thailand	0.68
Romania	0.73	Bulgaria	0.67
Australia	0.72	Georgia	0.65
Brazil	0.72	India	0.65
Czech Republic	0.72	Nigeria	0.64
Japan	0.72	Kenya	0.62
Norway	0.72	Belgium	0.60
Poland	0.72	Malaysia	0.60
Singapore	0.72	Senegal	0.60
South Korea	0.72	Indonesia	0.55
Taiwan	0.72	Vietnam	0.53
Ukraine	0.72	Uganda	0.51
Bolivia	0.71	Pakistan	0.50
Colombia	0.71	Macedonia	0.46
Greece	0.71	*World average*	*0.69*

One potential problem associated with this method is the lack of a metric to judge the resulting correlations among countries. This difficulty is not unique to this method; for more complex methods in the literature, various thresholds for acceptable degrees of "measurement invariance" have been proposed without clear justification. In the present case, as well, it is not obvious what a "good" correlation between two countries' inter-item matrices is, that implies sufficient comparability of the measure across these countries. One way to generate a reference point is by comparing the actual results with randomized correlations among arbitrary groups. In other words, what if it truly did not matter, at all, what country a participant was from? To test this hypothetical possibility, we removed the country identification from each of our more than 15,000 participants, and then re-assigned them to pseudo-"countries," randomly.

Specifically, a randomization program assigned each of the more than 15,000 participants to one of 63 groups, weighted to have equal sample sizes with the countries in the original dataset. Then, new inter-item correlation matrices were calculated for each of the 63 randomized groups. These group correlation matrices were then correlated with every other group to form a new inter-item correlation similarity matrix among randomized groups. The resulting average correlation among all the randomized cultural groups was $r = .80$, which is higher than the average correlation among the actual countries ($r = .69$). The correlation coefficient generated from randomized groups represents the upper limit of the best inter-item correlation matrix that can be expected from the data, given no cultural biases in item responses.

Once again, however, it is difficult to determine a metric for what is considered a high enough or too low of a matrix correlation for researchers to conclude enough equality in item interpretation for measures to be reliably compared across cultural groups. One method for assessing the amount of discrepancy expected is to assess the similarity of inter-item matrices within subgroups of one culture assumed to have very little, if any, discrepancies among groups. Six different sites within the US collected data for the ISP, representing Alabama, California, Connecticut, Idaho, Illinois, and Texas. While personality traits vary across the states (see Rentfrow, Gosling, & Potter, 2008), it is generally assumed that the comparability of measures across states is not an issue. Therefore, the US states represent a baseline metric for expected discrepancies between randomized group inter-item matrices and actual group inter-item matrices.

Following the same method as before, an inter-item correlation matrix was calculated for each of the six US sites and then correlated with each of the other US sites. The average inter-item correlation matrix for US sites was $r = .83$. For the randomized US sites, each US participant was randomly assigned to one of six groups, weighted to match the sample size of the original US sites. An inter-item correlation matrix was calculated for each of the randomized groups and then correlated with each of the other randomized groups. The average inter-item correlation matrix for the randomized US sites was $r = .84$, implying US states do not impact the comparability of measures across groups any more than randomly assigned groups.

Conclusion

Jüri Allik's pioneering contributions to cross-cultural research opened up new possibilities and new methodological challenges for psychology. One of those challenges is how to separate real differences between cultures from those that are products of response bias, shifts in meaning, or other measurement artifacts. The present chapter presents a new approach to that challenge and the current demonstration of the approach provides some interesting and important new information but also, as always, leaves us wanting to learn more.

The new information is the possibly-encouraging finding that the average similarity of patterns of item response to the BFI-2 across 63 countries is $r = .69$. However, we say "possibly encouraging" because we lack a clear benchmark of comparison. An r of .69 is generally regarded as large in most research contexts[6], but our randomization procedure described in this chapter suggests that if country really did not matter for item response, the r would be .80. A further analysis found that patterns of item response were consistent across several states of the US, with the actual and pseudo-groupings resulting in almost exactly equivalent patterns of response similarity. So our overall conclusion is this: For responding to the items of the BFI-2, it does not seem to matter which of the US states a participant is from. But internationally, it does seem to matter what country one is from. Beyond that conclusion, how, exactly, should we interpret this difference in response pattern similarity, between an empirically found average r of .69 across countries, and an r of .80 that would be obtained if countries did not make a difference? This is a matter yet to be resolved.

In any event, the ability of purely statistical methods to assess the comparability of measurements across cultures is fundamentally limited. While such methods as MGCFA and the much simpler approach used here provide interesting and useful information, the study of validity will, in the end, always require data from outside the measure being validated (Cronbach & Meehl, 1955). Two especially promising approaches are the use of anchoring vignettes to assess the comparability of the meaning of constructs across cultures (Mõttus et al., 2012), and using patterns of theoretically-predicted correlations with independently-assessed attributes of cultures (Mõttus et al., 2010). In other words, future research of the sort that Jüri Allik helped to pioneer and continues to conduct, will be needed as we continue to seek to understand the ways in which personality differs across cultures, and the ways in which it is the same.

[6] It is also highly statistically significant, given that the N for this correlation is 1,770, the number of non-redundant correlations being compared.

References

Allik, J., & McCrae, R. R. (2004). Toward a geography of personality traits: Patterns of profiles across 36 cultures. *Journal of Cross-Cultural Psychology*, *35*, 13–28.

Allik, J., & Realo, A. (2017). How valid are culture-level mean personality scores? In T. A. Church (Ed.), *The Praeger handbook of personality across cultures Vol 1: Trait psychology across cultures* (pp. 193–224). Santa Barbara, CA: Praeger.

Bartram, D. (2013). Scalar equivalence of OPQ32: Big Five profiles of 31 countries. *Journal of Cross-Cultural Psychology*, *44*, 61–83.

Byrne, B. M., & van de Vijver, F. J. R. (2010). Testing for measurement and structural equivalence in large-scale cross-cultural studies: Addressing the issue of nonequivalence. *International Journal of Testing*, *10*, 107–132.

Chen, F. F. (2008). What happens if we compare chopsticks with forks? The impact of making inappropriate comparisons in cross-cultural research. *Journal of Personality and Social Psychology*, *95*, 1005–1018.

Ching, C. M., Church, T. A., Katigbak, M. S., Reyes, J. A. S., Tanaka-Matsumi, J., Takaoka, S., … Ortiz, F. A. (2014). The manifestation of traits in everyday behavior and affect: A five-culture study. *Journal of Research in Personality*, *48*, 1–16.

Cronbach, L. J., & Meehl, P.E. (1955). Construct validity in psychological tests. *Psychological Bulletin, 52,* 281–302.

Davidov, E. Schmidt, P., & Schwartz, S.H. (2008). Bringing values back in: The adequacy of the European Social Survey to measure values in 20 countries. *Public Opinion Quarterly, 72,* 420–445.

Gurven, M., von Rueden, C., Massenkoff, M., Kaplan, H., & Lero Vie, M. (2013). How universal is the Big Five? Testing the five-factor model of personality variation among forager–farmers in the Bolivian Amazon. *Journal of Personality and Social Psychology*, *104*, 354–370.

Harzing, A.-W. (2006). Response styles in cross-national survey research a 26-country study. *International Journal of Cross Cultural Management*, *6*, 243–266.

Heine, S. J., Buchtel, E. E., & Norenzayan, A. (2008). What do cross-national comparisons of personality traits tell us? The case of conscientiousness. *Psychological Science, 19*, 309–313.

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.

John, O. P., & Srivastava, S. (1999). The Big Five trait taxonomy: History, measurement, and theoretical perspectives. In L. A. Pervin & O. John (Eds.), *Handbook of personality: Theory and research* (pp. 102–138). New York: Guilford Press.

McCrae, R. R., Terracciano, A., & 78 Members of the Personality Profiles of Cultures Project (2005). Universal features of personality traits from the observer's perspective: Data from 50 cultures. *Journal of Personality and Social Psychology*, *88,* 547–561.

Mõttus, R., Allik, J., & Realo, A. (2010). An attempt to validate national mean scores of Conscientiousness: No necessarily paradoxical findings. *Journal of Research in Personality, 44,* 630–640.

Mõttus, R., Allik, J., Realo, A., Pullmann, H., Rossier, J., Zecca, G., … Tseung, C. N. (2012). Comparability of self-reported conscientiousness across 21 countries. *European Journal of Personality, 26,* 303–317.

Rentfrow, P. J., Gosling, S. D., & Potter, J. (2008). A theory of the emergence, persistence, and expression of geographic variation in psychological characteristics. *Perspectives on Psychological Science*, *3*, 339–369.

Schmitt, D. P., Allik, J., McCrae, R. R., Benet-Martínez, V., Alcalay, L., Ault, L., ... Zupancic, A.. (2007). The geographic distribution of Big Five personality traits: Patterns and profiles of human self-description across 56 nations. *Journal of Cross-Cultural Psychology*, *38*, 173–212.

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,* 117–143.

Thalmayer, A. G., & Saucier, G. (2014). The questionnaire Big Six in 26 nations: Developing cross-culturally applicable Big Six, Big Five and Big Two inventories. *European Journal of Personality*, *28*, 482–496.

van de Schoot, R., Lugtig, P., & Hox, J. (2012). A checklist for testing measurement invariance. *European Journal of Developmental Psychology*, *9*, 486–492.

Zecca, G., Verardi, S., Antonietti, J. P., Dahourou, D., Adjahouisso, M., Ah-Kion, J., ... Rossier, J. (2013). African cultures and the Five-Factor Model of Personality: Evidence for a specific Pan-African structure and profile? *Journal of Cross-Cultural Psychology, 44*, 684–700.

Appendix 11A Members of the International Situations Project.

Argentina:	Maite Beramendi, Universidad de Buenos Aires
Australia:	Brock Bastian, University of Melbourne
Austria:	Aljoscha Neubauer, University of Graz
Bolivia:	Diego Cortez, Universidad Católica Bolviana, La Paz
Bolivia:	Eric Roth, Universidad Católica Bolviana, La Paz
Brazil:	Ana Torres, Federal University of Paraíba
Brazil:	Daniela S. Zanini, Pontifical Catholic University of Goiás
Bulgaria:	Kristina Petkova, Bulgarian Academy of Sciences
Canada:	Jessica Tracy, University of British Columbia
Canada:	Catherine Amiot, Université du Québec à Montréal
Canada:	Mathieu Pelletier-Dumas, Université du Québec à Montréal
Chile:	Roberto González, Pontificia Universidad Católica de Chile
Chile:	Ana Rosenbluth, Universidad Adolfo Ibáñez
Chile:	Sergio Salgado, Universidad de La Frontera
China, Beijing:	Yanjun Guan, Durham University, United Kingdom
China, Shanghai:	Yu Yang, Shanghai Tech University
Colombia:	Diego Forero, Universidad Antonio Nariño, Bogotá and Universidad de Ciencias Aplicadas y Ambientales, Bogotá
Colombia:	Andrés Camargo, Universidad Antonio Nariño, Bogotá and Universidad de Ciencias Aplicadas y Ambientales, Bogotá
Crete:	Emmanouil Papastefanakis, University of Crete
Crete:	Georgios Kritsotakis, Technological Institute of Crete
Crete:	Irene Spyridaki, University of Crete
Crete:	Evangelia Fragkiadaki, Hellenic American University
Croatia:	Željko Jerneić, University of Zagreb
Czech Republic:	Martina Hřebíčková, Czech Academy of Sciences
Czech Republic:	Sylvie Graf, Czech Academy of Sciences
Denmark:	Pernille Strøbæk, University of Copenhagen
Estonia:	Anu Realo, University of Tartu

France:	Maja Becker, Université de Toulouse
France:	Christelle Maisonneuve, Univ Rennes, Rennes
Gaza (Palestine):	Sofian El-Astal, Al Azhar University-Gaza
Georgia:	Vladimer Gamsakhurdia, Ivane Javakhishvili Tbilisi State University
Germany:	Matthias Ziegler, Humboldt University
Germany:	Lars Penke, University of Goettingen & Leipniz Science Campus Primate Cognition
Germany:	John Rauthmann, Universität zu Lübeck
Hong Kong:	Emma E. Buchtel, The Education University of Hong Kong
Hong Kong:	Victoria Wai-Lan Yeung, Lingnan University
Hungary:	Ágota Kun, Budapest University of Technology and Economics
Hungary:	Peter Gadanecz, Budapest University of Technology and Economics
Hungary:	Zoltán Vass, Karoli Gaspar University of the Reformed Church in Hungary
Hungary:	Máté Smohai, Karoli Gaspar University of the Reformed Church in Hungary
India:	Abhijit Das, AMRI Institute of Neurosciences, Kolkata
India:	Anagha Lavalekar, Jnana Prabodihini's Institute of Psychology, Pune
Indonesia:	Meta Zahro Aurelia, Univeritas Ahmad Dahlan
Indonesia:	Dian Kinayung (translators), Univeritas Ahmad Dahlan
Indonesia:	Vanessa Gaffar, Universitas Pendidikan Indonesia
Indonesia:	Gavin Sullivan, Coventry University
Indonesia:	Christopher Day, Coventry University
Israel:	Eyal Rechter, Ono Academic College
Italy:	Augusto Gnisci, University of Campania
Italy:	Ida Sergi, University of Campania
Italy:	Paolo Senese, University of Campania
Italy:	Marco Perugini, University of Milan-Bicocca
Italy:	Giulio Costantini, University of Milan-Bicocca
Japan:	Asuka Komiya, Hiroshima University
Japan:	Tatsuya Sato, Ritsumeikan University
Japan:	Yuki Nakata, Ritsumeikan University
Japan:	Shizuka Kawamoto, Yamanashi University
Jordan:	Marwan Al-Zoubi, University of Jordan
Kenya:	Nicholas Owsley, Busara Center for Behavioral Economics
Kenya:	Chaning Jang, Busara Center for Behavioral Economics
Kenya:	Georgina Mburu, Busara Center for Behavioral Economics
Kenya:	Irene Ngina, Busara Center for Behavioral Economics
Latvia:	Girts Dimdins, University of Latvia
Lithuania:	Rasa Barkauskiene, Vilnius University
Lithuania:	Alfredas Laurinavicius, Vilnius University
Malaysia:	Khairul A. Mastor, Universiti Kebangsaan Malaysia
Mexico:	Elliott Kruse, EGADE Business School Monterrey
Mexico:	Nairán Ramírez-Esparza, Fundación Universidad de las Américas Puebla
Netherlands:	Jaap Denissen, Tilburg University
Netherlands:	Marcel Van Aken, University of Utrecht
New Zealand:	Ron Fischer, Victoria University of Wellington, Wellington

Nigeria:	Ike E. Onyishi, University of Nigeria, Nsukka
Nigeria:	Kalu T. Ogba, University of Nigeria, Nsukka
Norway:	Siri Leknes, University of Oslo
Norway:	Vera Waldal Holen, University of Oslo
Norway:	Ingelin Hansen, University of Oslo
Norway:	Christian Krog Tamnes, University of Oslo
Norway:	Kaia Klæva, University of Oslo
Pakistan:	Muhammad Rizwan, Government of Pakistan
Pakistan:	Rukhsana Kausar, University of the Punjab, Lahore
Pakistan:	Nashi Khan, University of the Punjab, Lahore
Peru:	Agustín Espinosa, Pontificia Universidad Católica del Peru
Philippines:	Maria Cecilia Gastardo- Conaco, University of Philippines-Diliman
Philippines:	Diwa Malaya A. Quiñones, University of Philippines-Diliman
Poland:	Piotr Szarota, Institute of Psychology of The Polish Academy of Sciences
Poland:	Paweł Izdebski, Kazimierz Wielki University
Poland:	Martyna Kotyśko, University of Warmia and Mazury
Portugal:	Joana Henriques-Calado, Universidade de Lisboa, Faculdade de Psicologia, CICPSI, Alameda da Universidade, Lisboa
Romania:	Florin Alin Sava, West University of Timisoara
Russia:	Olga Lvova, St. Petersburg State University
Russia:	Victoria Pogrebitskaya, St. Petersburg State University
Russia:	Mikhail Allakhverdov, St. Petersburg State University
Russia:	Sergey Manichev, St. Petersburg State University
Senegal:	Oumar Barry, Université Cheikh Anta Diop de Dakar-Sénégal
Serbia:	Snežana Smederevac, University of Novi Sad
Serbia:	Petar Čolović, University of Novi Sad
Serbia:	Dušanka Mitrović, University of Novi Sad
Serbia:	Milan Oljača, University of Novi Sad
Singapore:	Ryan Hong, National University of Singapore
Slovakia:	Peter Halama, Slovak Academy of Sciences
Slovenia:	Janek Musek, University of Ljubljana
South Africa:	Francois De Kock, University of Capetown
South Korea:	Gyuseog Han, Chonnam National University
South Korea:	Eunkook M. Suh, Yonsei University
South Korea:	Soyeon Choi, Yonsei University
Spain:	Luis Oceja, Universidad Autónoma de Madrid
Spain:	Sergio Villar, Universidad Autónoma de Madrid
Spain:	David Gallardo-Pujol, University of Barcelona
Sweden:	Zoltan Kekecs, Lund University
Sweden:	Nils Arlinghaus, Lund University
Sweden:	Daniel P. Johnson, Lund University
Sweden:	Alice Kathryn O'Donnell, Lund University
Switzerland:	Janina Larissa Bühler, University of Basel
Switzerland:	Clara Kulich, Université de Genève
Switzerland:	Fabio Lorenzi-Cioldi, Université de Genève
Switzerland:	Mathias Allemand, University of Zurich
Taiwan:	Yenping Chang, University of North Carolina

Taiwan:	Weifang Lin, Chulalongkorn University
Thailand:	Watcharaporn Boonyasiriwat, Chulalongkorn University
Turkey:	Adil Saribay, Boğaziçi University
Turkey:	Oya Somer, Cyprus International University
Turkey:	Pelin Karakus Akalin, Istinye University, Istanbul
Uganda:	Peter Kakubeire Baguma, Makerere University
Ukraine:	Alexander Vinogradov, Taras Shevchenko National University of Kyiv
Ukraine:	Larisa Zhuravlova, Zhytomyr Ivan Franko State University
United Kingdom:	Jason Rentfrow, University of Cambridge
United Kingdom:	Mark Conner, University of Leeds
United States, AL:	Alexa Tullett, University of Alabama
United States, CA:	Erica Baranski, University of California, Riverside
United States, CT:	Nairán Ramírez-Esparza, University of Connecticut
United States, ID:	Douglas E. Colman, Idaho State University
United States, IL:	Joey T. Cheng, University of Illinois at Urbana-Champaign
United States, TX:	Eric Stocks, University of Texas, Tyler
Viet Nam:	Huyen Thi Thu Bui, Hanoi National University of Education