SOCIAL SCIENCES

Challenges to capture the big five personality traits in non-WEIRD populations

Rachid Laajaj[1]*†, Karen Macours[2,3†], Daniel Alejandro Pinzon Hernandez[1†], Omar Arias[4], Samuel D. Gosling[5,6], Jeff Potter[7], Marta Rubio-Codina[8], Renos Vakis[4]

Can personality traits be measured and interpreted reliably across the world? While the use of Big Five personality measures is increasingly common across social sciences, their validity outside of western, educated, industrialized, rich, and democratic (WEIRD) populations is unclear. Adopting a comprehensive psychometric approach to analyze 29 face-to-face surveys from 94,751 respondents in 23 low- and middle-income countries, we show that commonly used personality questions generally fail to measure the intended personality traits and show low validity. These findings contrast with the much higher validity of these measures attained in internet surveys of 198,356 self-selected respondents from the same countries. We discuss how systematic response patterns, enumerator interactions, and low education levels can collectively distort personality measures when assessed in large-scale surveys. Our results highlight the risk of misinterpreting Big Five survey data and provide a warning against naïve interpretations of personality traits without evidence of their validity.

INTRODUCTION

While there is a large body of evidence on the importance of cognitive ability for predicting social and economic success, personality traits (PTs) are often emphasized to be equally important for many aspects of life (*1*, *2*). The most influential taxonomy of PTs is the Big Five personality inventory (*3*, *4*). Ample empirical evidence from the United States and other high-income countries shows that the Big Five PTs correlate with earnings, employment, and other labor market outcomes. Recent reviews conclude that Conscientiousness and Emotional Stability in particular are strong predictors of job performance and wages (*5*, *6*). PTs are found to be particularly important for people with lower levels of job complexity or education level, whereas cognitive ability is more important at higher levels of job complexity (*1*). One could hence hypothesize that PTs could be even more important in low- and middle-income countries, where large shares of the population participate in lower-complexity jobs.

Increasingly, the Big Five PTs are measured in developing country settings. A growing literature in economics [reviewed in (*7*)] analyzes not only whether different PTs can predict selection (*8*, *9*) or performance (*10*) in public sector jobs but also whether the effectiveness of interventions aimed at improving public service delivery (for instance, through increased payments or monitoring) depends on the PTs of the targeted personnel (*11*) and whether these interventions can change the quality of people being recruited into the public sector (*12*). Measurement of PTs are also used for predicting economic performance (*13*, *14*), for screening purposes in the job market (*15*), for credit eligibility (*16*, *17*), or for analyzing treatment heterogeneity of nudge interventions (*18*). Last, PTs are sometimes considered as outcomes that can be affected by skill-enhancing or behavioral interventions (*19*, *20*). In a few cases (*12*, *15*), studies use versions of Big Five measures that have been specifically validated for the countries studied.

This rapidly growing literature builds on the wide support for the universality of the Big Five across cultures established in the psychology literature (*21*–*25*). The motivation for the Big Five taxonomy originates in research observing that the same five factors broadly emerged each time a factor analysis was conducted to classify a set of questions describing personality (*4*), hence referred to as the five-factor model (FFM). However, these studies have mostly focused on highly educated populations (often college students) in high-income countries, often referred to in the literature as WEIRD (western, educated, industrialized, rich, and democratic) (*26*). Some notable exceptions are (*27*, *28*), who do not find robust support for the FFM using data from orally administered surveys on rural populations with low levels of education in Bolivia, Colombia, and Kenya. Related, Ludeke and Larson (*29*) flag concerns with the use of the BFI-10 (*30*), a short 10-item Big Five instrument used in the World Values Survey, showing low correlations between items meant to measure the same PT. While they interpret this as possible evidence against universality across cultures, Gosling *et al.* (*31*) explain that interitem correlation per se is not necessarily a good indicator of validity for short scales. More generally, Church (*32*) reviews evidence on cultural differences in personality and concludes that the FFM of personality continues to find cross-cultural support but may be difficult to replicate in less educated populations. Soto *et al.* (*33*) show that, between ages 10 and 18, there is a strong relationship between cognitive ability and between-domain differentiation of Big Five questions, while internal consistency of Big Five scales also increased with age. This suggests that variations in cognitive ability may explain variations in differentiation of the Big Five structure.

This paper compiles and analyzes Big Five data collected through a very diverse set of surveys from low- and middle-income countries in different parts of the world and covering all types of education levels, including large nationally representative surveys and surveys collected on targeted samples for impact evaluation purposes. It shows that commonly used personality questions fail to measure the intended PTs in these settings and do not pass standard validity tests. The analysis of the psychometric properties of Big Five measures raises serious concerns about their validity and hence about their use and subsequent interpretation in many studies in low- and middle-income countries when collected through surveys. In contrast, data collected through

[1]Universidad de Los Andes, Bogota, Colombia. [2]Paris School of Economics, Paris, France. [3]INRA, Paris, France. [4]World Bank, Washington, DC 20433, USA. [5]University of Texas at Austin, Austin, TX, USA. [6]University of Melbourne, Parkville, Victoria, Australia. [7]Atof Inc., Cambridge, MA 02139, USA. [8]Inter-American Development Bank, Washington, DC 20577, USA.
*Corresponding author. Email: r.laajaj@uniandes.edu.co
†These authors contributed equally to this work.

the internet from the same set of countries reflect the Big Five factor structure, suggesting that low validity is not primarily driven by cultural or contextual differences. We explore the possible reasons for the measurement challenges and derive recommendations for the use of personality questions collected through surveys among non-WEIRD populations.

As motivation for the analysis, Fig. 1 shows that the relationship between the Big Five PT measures and income in a large database covering 14 low- and middle-income countries appears much less stable than that found in the U.S. literature (see Supplementary Materials for regression specifications and variable definitions). Most notably, Conscientiousness (often emphasized in the U.S. literature as being predictive of earnings) is not a significant predictor of income in 10 of the 14 countries analyzed (with the point estimate being negative in three), while Extraversion and Agreeableness also show some negative coefficients. Emotional Stability and, to a lesser extent, Openness show a more consistent positive relationship. The bottom panel of Fig. 1 compares an index averaging the five PTs to cognitive ability for the nine countries for which a good measure of cognition was available. Cognitive ability shows a stronger and more significant relationship with income than the Big Five index in almost every country. Moreover, when pooling the nine countries together, the size of the coefficient of the Big Five index is less than half of that of cognitive ability. Taken at face value, these results would suggest that personality matters much less than cognition and potentially lead to puzzling conclusions such as the observation that Conscientiousness (being achievement oriented and organized) does not correlate with earnings in many low- and middle-income countries. In psychometrics, predictive validity assesses the extent to which the measure is a good predictor of outcomes with which it can be expected to be related. It is used as one of the criteria for assessing whether a measure captures what it intends to capture. Hence, the fact that the Big Five measures are less predictive of income than cognition and that the traits that appear to best predict income are quite different from what has been found in other contexts can raise doubts about the measures themselves. Of course, it could also mean that PTs actually matter differently for economic success in low- and middle-income countries, where most individuals face many more external constraints than do WEIRD populations. However, these conclusions

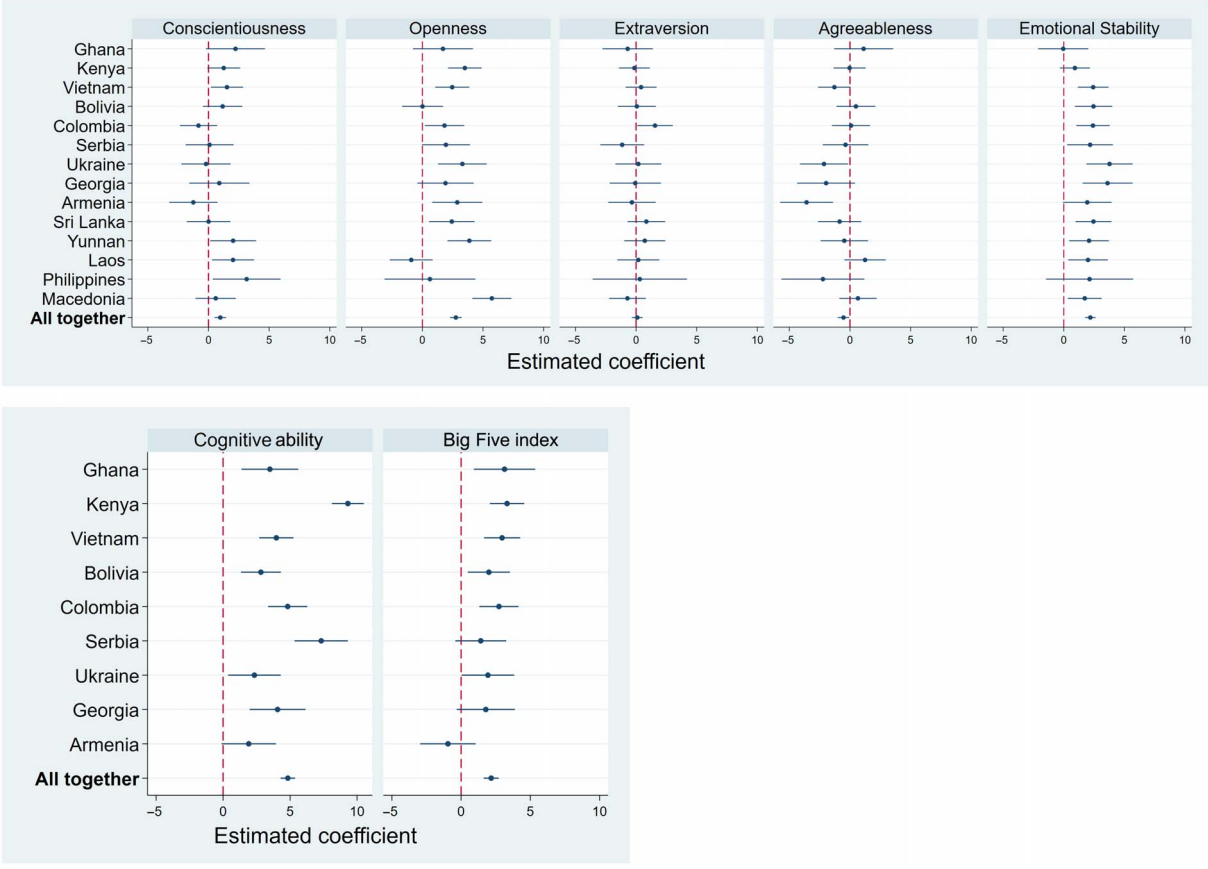

Fig. 1. Prediction of income with the Big Five PTs and cognitive ability. The top panel presents the relationship between income and the different PTs. We estimated the coefficients and their 95th confidence intervals running the following regression separately for each country: $y_i = \alpha_0 + \beta_0 cog_i + \sum_{PT=5}^{PT=1} \beta_{PT} PT_i + \epsilon_i$, where y_i is the income of person i (transformed into the rank of income and scaled from 0 to 100). For cog_i, the full literacy test was used when available, and the partial literacy test was used otherwise (see the Supplementary Materials). The last line pools all observations from all countries, controlling for the best measure of cognitive ability available in each country (country-fixed effects are not needed since relative rankings were calculated by country). All regressors are standardized, hence coefficients can be interpreted as the effect of 1 SD in the regressor on the percentile of the rank of income. The bottom panel is based on similar regressions, but replacing the five PTs by one index averaging the five values (for each observation). It only includes the nine countries with full literacy test included in the STEP database. The last line pools all observations from the nine countries. The coefficient of cognitive ability (literacy) is significantly higher than the coefficient of the Big Five index (at the 90% level) in four out of nine countries and in the pooled regression.

would be premature if, as we show in this paper, the Big Five questions collected in large-scale surveys may not only be noisy but also measure latent traits other than the PTs they are intended to measure.

RESULTS

Data

We use four types of databases, all of which include self-reported measures of the Big Five drawn from the BFI 44-item scale (*21*, *34*, *35*). First, the World Bank's Skills towards Employment and Productivity (STEP) database contains data on the same subset of 15 Big Five items, which we will refer to as the STEP items. They were collected in large representative samples in 12 countries, and in two additional countries, a larger set of 35 BFI items were collected (for a total of 20,584 individuals; see table S1 for descriptive statistics and Supplementary Materials section 1.1 for the list of STEP items). The STEP data were collected through face-to-face surveys, often undertaken in local languages (*36*) and mostly representative of urban populations with diverse levels of education. A critical advantage of these data is that the same data collection instrument and standardized methods were applied across different countries from Africa, Asia, Latin America, Europe, and Central Asia. The data also contain measures of functional literacy, incorporated in the STEP surveys to provide a proxy for the respondents' cognitive ability, and are therefore referred to as such in this paper. See Supplementary Materials for an explanation on the choice of literacy as a proxy for a broader set of foundational cognitive skills.

The second database consists of 15 other datasets, from 12 developing countries, including a total of 54,167 households, which all contain a (partial or complete) version of the BFI. These datasets, described in table S2, were collected for varied purposes and by different institutions and researchers in local languages. They were either graciously shared by the researchers who collected them or are in the public domain (and some were collected by authors of this article). These include datasets used in top-level publications (*9*, *12*, *20*, *37*–*43*). The database contains both face-to-face and self-administered surveys and covers different populations: farmers, entrepreneurs, civil servants, and adolescents. Most surveys are representative of specific subpopulations targeted for particular interventions and randomized controlled trials. While the number of items varies across datasets, we use the same 15 items as in the STEP data for most of the analysis for comparability. These datasets are identified with a code because the objective is not to provide diagnostics for individual studies but rather to show the generalizability of the findings.

The third database contains data obtained from volunteers who visited a noncommercial website (www.outofservice.com), provided sociodemographic data, and filled the 44 BFI items. This database has been widely used in the psychology literature (*44*, *45*). The website provides respondents immediate scoring and feedback regarding their personality, which is the main driver of the sample recruitment (*46*), resulting in a population of young and highly-educated respondents. To facilitate comparisons, we keep only the data from the 198,356 individuals who live in any of the 14 countries included in the STEP database and restrict most of the analysis to the same 15 items (see table S3 for descriptive statistics). The BFI could be completed in English, Spanish, German, or Dutch so not necessarily in the local languages.

A fourth database used, for comparison and reference, contains 44 BFI items self-administrated to a community sample in the United States, containing 642 adults. It was chosen because of its high degree of validity, shown by Soto and John (*35*), and includes a wide age range (18 to 85 years of age) and is balanced by gender (58% women). The factor structure in this data is similar to the one found in many other datasets on WEIRD populations in the United States and has been used in influential methodological work (*47*), making it an appropriate benchmark.

Overall, we combine 31 datasets from 24 countries amalgamating data on about 300,000 individuals.

We corrected all data for acquiescence response style, i.e., the tendency of an individual to consistently agree (yea saying) or disagree (nay saying). Not correcting the acquiescence before factor analysis often results in the emergence of a factor representing the response pattern (*48*). See Materials and Methods for an explanation of the correction. Table S4 shows that, without correction, the internal validity of the survey data is even lower. Most notably, before the correction, we find 17 cases (in 10 different countries) in which an item has a negative correlation with other items intended to measure the same PT construct (table S5A). This number goes down to only one case after correcting for acquiescence bias (table S5B).

The absence of the Big Five factor structure in survey data

We start by analyzing the extent to which the FFM can be found in the data. Separately for each dataset, we examine the factorial structure. Following the common practice in the literature (*25*, *49*), we use principal components analysis (PCA) with Procrustes rotation to align the factor loadings with those in the U.S. data (described in Materials and Methods). The matrix of factor loadings, which provides the estimation of how much each item contributes to each component, indicates the extent to which the factor structure of the data matches the expected one.

Table 1 provides a visual representation of the (lack of) congruence. We assigned each item to the PT for which it has the highest factor loading and colored in red bold font every cell where an item is associated to a factor that is different than the one it is intended to measure. When the data behave according to the FFM, we expect items within the same PT to correlate among themselves more than with items from a different PT so that they would be pulled together into the same factor, as can be seen for the U.S. data in Table 1. By contrast, of the 23 survey datasets, only two have all items sorted according to the FFM. For the other countries, of the 15 items, between one and nine items have their highest factor loadings on PTs other than those they intend to measure. The number of wrongly assigned items is high for Conscientiousness, Openness, and Agreeableness, whereas in most countries, Emotional Stability more clearly distinguishes itself from other PTs.

The congruence coefficient in the first column of Table 1 provides a quantitative indicator of the observed mismatch, as it can be interpreted as an index of similarity between two-factor structures. It is the correlation between the factor loadings in a given dataset compared to the factor loadings of the United States (see Materials and Methods for computational details). It is clear from the data that higher congruence coefficients are associated with more differentiation of the FFM. The two datasets with the lowest congruence coefficient (0.59 and 0.60) are the ones with the most items wrongly sorted (seven and nine items respectively), while the two datasets with the highest congruence coefficients (0.84 and 0.92) are the only ones for which all items were sorted according to the FFM. The average congruence coefficient across all survey data is 0.73. Although any threshold is somewhat arbitrary, Lorenzo-Seva and ten Berge (*50*) argue that a congruence coefficient in the range of 0.85 to 0.94 corresponds to a fair similarity. Across all databases used in this paper, we find that a proper differentiation of the FFM appears to emerge with congruence coefficients around 0.85.

Table 1. Congruence, factor structures obtained by PCA, and comparison with theoretical scale. Congruence coefficient (first column) is a proxy for the similarity of the factor structure, obtained from the correlation between factor loadings between two samples (in this case, the sample of the corresponding line is compared to factor loadings of the U.S. data). A detailed description of the calculation of the congruence coefficient is provided in Materials and Methods. In the rest of the table, each column represents one item (the same across datasets), sorted by PTs, and an "R" in its name means that it is a reverse-coded item (see Supplementary Materials section 1.1 for the phrasing of each item). The number that appears in each cell indicates for which factor the item has the highest factor loading in the PCA (with Procrustes rotation on U.S. data). Cell entries are in red bold font when the factor with highest loading differs from the one that the item aims to measure. For other surveys, only identifiers are provided to preserve anonymity. All data were corrected for acquiescence bias before the analysis.

U.S. data

	Congruence coefficient	Openness			Conscientiousness			Extraversion			Agreeableness			Emot. stability		
		O1	O2	O3	C1	C2R	C3	E1	E2R	E3	A1	A2	A3	ES1R	ES2	ES3R
U.S. data	1	1	1	1	2	2	2	3	3	3	4	4	4	5	5	5

STEP survey data

	Congruence coefficient	Openness			Conscientiousness			Extraversion			Agreeableness			Emot. stability		
		O1	O2	O3	C1	C2R	C3	E1	E2R	E3	A1	A2	A3	ES1R	ES2	ES3R
Ghana	0.68	**4**	1	1	**1**	2	**4**	3	3	3	**1**	**1**	**1**	5	5	5
Kenya	0.71	**2**	1	**4**	2	2	**1**	3	3	3	4	4	4	5	5	5
Sri Lanka	0.71	**2**	1	1	**1**	2	**1**	3	3	**1**	4	4	4	5	5	5
Yunnan	0.75	1	1	**4**	2	2	2	3	3	3	**1**	4	4	5	5	5
Laos	0.70	1	1	**3**	**4**	2	**1**	**1**	3	**4**	4	4	4	5	5	5
Vietnam	0.69	1	1	1	2	2	2	**4**	3	3	4	**3**	**1**	5	5	5
Philippines	0.59	**3**	1	1	**4**	2	**4**	3	3	**1**	**3**	**2**	4	**4**	5	5
Bolivia	0.84	1	1	1	2	2	2	3	3	3	4	4	4	5	5	5
Colombia	0.72	**3**	**2**	**4**	2	**1**	2	3	3	3	4	4	4	5	5	5
Macedonia	0.67	1	1	**4**	**1**	**1**	**1**	3	3	3	4	4	4	5	**2**	5
Serbia	0.79	1	1	1	2	2	**4**	3	3	3	4	4	4	5	5	5
Ukraine	0.81	**2**	1	1	2	**4**	2	3	3	3	4	**1**	**1**	5	5	5
Georgia	0.77	**3**	1	1	**1**	2	2	3	**1**	3	4	4	**1**	5	5	5
Armenia	0.82	1	1	1	**1**	2	2	3	3	3	4	4	**1**	5	5	5
Average	**0.73**															

Other survey data

	Congruence coefficient	Openness			Conscientiousness			Extraversion			Agreeableness			Emot. stability		
		O1	O2	O3	C1	C2R	C3	E1	E2R	E3	A1	A2	A3	ES1R	ES2	ES3R
D1	0.78	**2**	1	1	2	2	2	3	3	**2**	4	**5**	4	5	5	5
D2	0.69	**2**	**3**	**5**	2	**1**	2	3	3	3	4	4	4	5	**1**	5
D3	0.92	1	1	1	2	2	2	3	3	3	4	4	4	5	5	5
D4	0.60	**3**	**4**	**4**	**3**	2	**4**	3	3	**4**	**1**	**1**	4	5	**1**	5
D5	0.67	**2**	1	1	**1**	2	**4**	3	3	**1**	4	4	4	5	5	5
D6	0.68	**2**	1	1	2	2	**3**	**2**	3	**5**	**1**	4	4	5	5	5

continued on next page

D7	0.66	1	1	1	2	4	4	2	3	3	4	5	4	5	5	2
D8	0.71	2	1	1	5	2	5	3	3	3	2	5	4	5	4	2
D9	0.73	1	1	2	2	4	2	3	3	3	1	4	4	5	5	5
Average	**0.71**															

Internet data

	Congruence coefficient	Openness			Conscientiousness			Extraversion			Agreeableness			Emot. stability		
		O1	O2	O3	C1	C2R	C3	E1	E2R	E3	A1	A2	A3	ES1R	ES2	ES3R
Ghana	0.97	1	1	1	2	2	2	3	3	3	4	4	4	5	5	5
Kenya	0.86	1	1	1	2	2	2	3	3	3	4	4	4	5	5	5
Sri Lanka	0.64	3	1	2	2	2	2	3	3	3	2	1	2	5	5	5
China	0.98	1	1	1	2	2	2	3	3	3	4	4	4	5	5	5
Laos	0.81	1	1	5	2	2	2	3	3	3	4	4	5	5	5	5
Vietnam	0.98	1	1	1	2	2	2	3	3	3	4	4	4	5	5	5
Philippines	0.85	1	1	1	2	4	2	3	3	3	4	4	4	5	5	5
Bolivia	0.94	1	1	1	2	2	2	3	3	3	4	4	4	5	5	5
Colombia	0.97	1	1	1	2	2	2	3	3	3	4	4	4	5	5	5
Macedonia	0.76	1	1	1	2	2	2	3	4	3	4	4	4	5	5	5
Serbia	0.98	1	1	1	2	2	2	3	3	3	4	4	4	5	5	5
Ukraine	0.98	1	1	1	2	2	2	3	3	3	4	4	4	5	5	5
Georgia	0.90	1	1	1	2	2	2	3	3	3	4	4	4	5	5	5
Armenia	0.97	1	1	1	2	2	2	3	3	3	4	4	4	5	5	5
Average	**0.90**															

For comparison, the top panel of Table 1 shows that the variables in the U.S. data perfectly sort into the FFM, and in the bottom panel of Table 1, we present the same analysis with the internet dataset, restricted to the same countries and 15 items as STEPs. In stark contrast with the survey data, the average congruence coefficient is 0.9, and in 10 out of the 14 countries, the PCA matched all the 15 items with the intended PT. The internet data are from the same countries as the STEP database, hence the results suggest that cultural differences are unlikely to be the main driver of the low validity found in survey data. In many cases, the language of administration differed between the survey and the internet; however, differences are about as stark when comparing survey and internet data from countries in which the languages used in the survey are one of the four languages in which the internet questionnaire was available (e.g., Colombia or Kenya).

Other reliability and validity statistics

To more comprehensively document the lack of reliability and validity of the Big Five measures in survey data, Table 2 shows the within correlation (average correlation between items that are within the same PT), the between correlation (average correlation between items of different PTs), Cronbach's alpha, and the congruence coefficients. These indicators were first calculated by dataset and then aggregated by database (table S6 shows statistics by dataset, and table S7 splits up the results by PT).

The top panel of Table 2 shows within and between correlations for all available datasets but restricts the analysis to the 15 items available in the STEP. Strictly positive correlations between items indicate that they are capturing something in common rather than just noise. The expected factor structure stands out when there is sufficient difference between the within correlation and the between correlations (i.e., the three items belonging to a same PT should have a higher correlation among them than with the other 12 items).

We find that the survey data have an average within correlation of 0.22 and a between correlation of 0.09. This compares to a within and between correlation of 0.45 and 0.10, respectively, in the United States and 0.32 and 0.09 for the internet data. The fact that the difference between the between and within correlation is greater with the internet data is consistent with its more discernible factor structure. This can further be inferred from table S8, which shows the item-by-item correlation coefficients for the survey database (combining STEP and other surveys) versus the correlation coefficients of the internet data and the United States. For internet data and the United States, correlations between

Table 2. Psychometric indicators by database. Datasets, psychometric measures, and different datasets and subsamples are described in detail in the main text and Supplementary Materials. All calculations were done after correcting for acquiescence bias. Within correlation, between correlation, Cronbach's alpha, and congruence coefficient are first calculated by dataset before calculating a nonweighted average across all datasets. See table S6 for calculations for each dataset separately. In the case of within correlation, Cronbach's alpha, and congruence coefficient, for each dataset, we first calculate it by PT and then average it across PT (before averaging across datasets).

	No. of items	No. of datasets	No. of observations	Within correlation	Between correlation	Cronbach's alpha (avg. of 5 PTs)	Congruence coeff. (avg. of 5 PTs)
Restricting each dataset to the 15 items available in step surveys							
STEP surveys	15	14	40,584	0.26	0.09	0.49	0.73
Other survey data	15	9	14,051	0.17	0.08	0.35	0.71
All survey data (that have the 15 STEP items)	15	23	54,635	0.22	0.09	0.44	0.73
Internet data	15	14	198,356	0.32	0.09	0.57	0.90
U.S. data	15	1	642	0.45	0.10	0.70	-
Using all items available in each dataset							
All survey data	10 to 44	29	94,751	0.23	0.10	0.51	0.72
Survey data with 44 items	44	6	6,017	0.17	0.11	0.62	0.67
Internet data	44	14	198,356	0.30	0.09	0.77	0.89
U.S. data	44	1	642	0.40	0.09	0.85	-
Subsample of respondents							
STEP surveys (tertiary education)	15	14	9,747	0.26	0.09	0.49	0.73
Internet data (number of observations $\leq$ STEP)	15	14	29,528	0.32	0.09	0.57	0.87
Internet data (age 26 to 48)	15	14	50,622	0.33	0.10	0.57	0.89

items meant to capture the same PT are consistently much higher than correlations with any other items (with the only exception of the first Openness item). By contrast, for the survey database, there are several items that show higher correlations with some items meant to measure other PTs than with items meant to measure the same PT (including two items of Conscientiousness). For the internet and U.S. data, the 10 highest correlations are all within correlations. But for the survey data, despite averaging over a large number of datasets, of the 10 highest correlations, 4 are between correlations and 6 are within. The fact that many questions correlate more with items intended to measure a different PT than with items intended to measure the same PT makes it arguably hard to interpret items as capturing the intended PT.

Table 2 also shows Cronbach's alpha, one of the most widely used measures of internal consistency of a test. For a given number of items, it increases when the correlation between items of the same PT increases. Hence, it is higher when the noise of each item is low and when they measure the same underlying factor. A minimum threshold of 0.7 is often applied, but as explained by Gosling *et al.* (*31*), a short measure should not aim at maximizing Cronbach's alpha, because each set of items of a PT needs to capture the breadth of the concept. Still, the fact that the survey data obtain substantially lower Cronbach's alpha than the internet data or the U.S. data when comparing the same items raises further concerns about the internal validity of the measure in the large survey databases. While all PTs measured in surveys show relatively low values, results indicate that internal consistency in the survey data is the lowest for Agreeableness and somewhat better for Emotional Stability.

This core set of results highlights that many Big Five survey data collected in low- and middle-income countries do not follow the FFM. Conceptually, this implies that some items correlate less with items that aim at measuring the same PT than with items that belong to different PTs. The evidence that led psychometricians to conclude that the Big Five are universal tends to hold for the internet data, but it is far from apparent in the survey data from developing countries. Moreover, this appears not to be driven by differences in average age of the respondents or sample size in the internet data, as limiting the internet sample to ages and sample sizes similar to those of the survey data yields broadly similar results as for the full sample of internet data (bottom panel of Table 2).

What can explain the lack of clear Big Five factor structure in survey data?

We investigate possible explanations for the low validity of the PT measures, including the number of items, the cognitive ability and education levels of the respondents, the administration method, and systematic response patterns. One might be concerned that it is difficult to recover the factor structure with only 15 items or that the 15 items selected for STEP were poorly chosen. Note, however, that validity with the same 15 items is much higher in the internet data. Moreover, the congruence

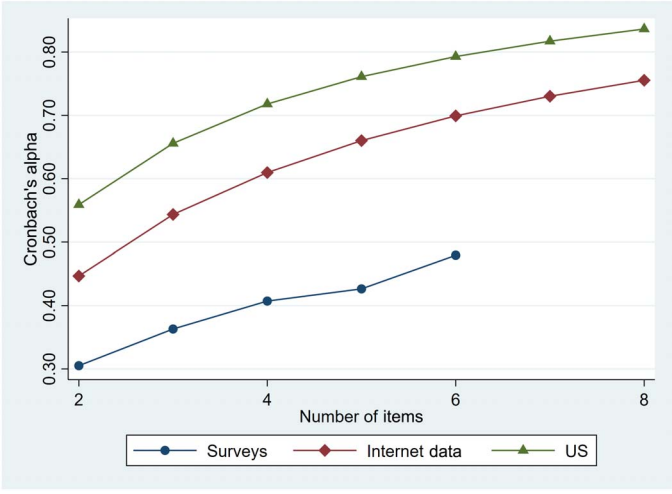

Fig. 2. Cronbach's alpha as a function of the number of items by type of data. The estimates for the survey data are based on surveys with at least six items per PT, while the estimates for the internet data are based on data from the 14 STEP countries, using all 44 items of the BFI. The U.S. data also use all items of the BFI. For each dataset and for each number of items *n*, we calculate Cronbach's alpha for every possible combination of *n* items before averaging it across all combinations and then averaging it across datasets (by type of data collection).

coefficient, if anything tends to decrease with the number of items because the overfitting that occurs when the number of components is high with respect to the number of items, is reduced. In addition, the between and within correlation are on average not affected by the number of items. Second, Fig. 2 shows that, even if the Cronbach's alpha is increasing in the number of items, following the Spearman-Brown prophecy formula, for any number of items, the survey data perform substantially worse than the internet or U.S. data. Last, the middle panel of Table 2 presents statistics using the full (*44*) set of items available in other surveys and still shows overall low validity, substantially lower than the levels observed for the internet data.

The educational level of the respondents is another potential driver of the differences in the reliability and internal validity of survey data compared to datasets used by psychologists in the United States or the internet data. On average, only 25% of respondents of the STEP surveys have college education, compared to 81% in the internet data. The bottom panel of Table 2 therefore presents a set of psychometric indicators when restricting the STEP data to respondents who have had some college education. This increases comparability between the STEP and internet data and brings the sample of STEP respondents closer to the convenience samples of university students often used in psychometric studies. Unexpectedly, we find no improvement in any of the indicators, suggesting that the level of education of respondents may not be a primary driver of the low validity. Similar results are obtained when restricting the STEP samples to individuals with white-collar jobs.

More generally, cognitive ability could play a role since it may affect people's understanding of the arguably abstract Big Five questions. We analyze the role of cognitive ability with the STEP database, using as a proxy for cognition the measure of functional literacy that is comparable between individuals and countries of the STEP surveys. Figure 3 presents the relationship between psychometric indicators and the cognitive ability of the respondents. In this figure, the unit of observation is the region, corresponding to the largest geographical division within each country as indicated in STEP, resulting in between 2 and 15 regions per country. Each figure depicts one of the indicators, separately calculated for each of the region, and their correlation with the regional-level average cognitive ability of the respondents. The analysis is limited to the nine countries for which good cognitive measures are available. Of course, since regions with low-average cognitive ability are likely to differ along many other dimensions, these correlations may not have any causal interpretation. However, the congruence coefficient and the Cronbach's alpha have a clear positive and significant relationship with the measure of cognitive ability. The variation is substantial with the congruence coefficient varying from about 0.5 in the lower end to about 0.7 in the upper end of the cognitive scores. Hence, survey data for regions with lower-average cognitive ability show factor structures that are less consistent with the FFM and less internally valid. Yet, this is not the entire story because even the regions with the highest average cognitive ability remain below acceptable psychometric standards. Moreover, once we account for average differences between countries by including country-fixed effects in the estimations, these relationships are no longer significant, making it unclear whether they capture differences in cognition or rather other cross-country contextual differences that could affect responses in face-to-face surveys.

Low validity of the PT measures could also be related to systematic response biases and answering patterns that are potentially more prevalent in survey data. Social desirability bias, for instance, could help explain why Conscientiousness and Agreeableness are the most problematic PTs and why Conscientiousness has little predictive power in the survey data. The databases do not allow us to quantify social desirability bias, but we consider two other indicators of undesirable response patterns that can contribute to blurring the measures of the PTs: (i) the absolute value of the acquiescence that indicates a respondent's tendency to agree with two contradictory statements and ii) the share of the variance in responses that can be explained by enumerator-fixed effects (i.e., dummy variables capturing which enumerator administrated the face-to-face survey), which we refer to as potential enumerator bias.

The bottom left panel of Fig. 3 shows a small negative relationship between acquiescence bias and cognition. Using regions as the level of observation, the relationship is not significant, but it is steeper and significant when using individual data rather than regional averages ($P < 0.001$). This result highlights that, within a country, respondents with lower cognitive skills are more likely to agree with statements that are mutually inconsistent. It is in line with Soto *et al.* (*33*) who found acquiescence bias to reduce between ages 10 and 18, when cognitive ability increases. Overall, Fig. 3 suggests that, even if low levels of respondents' cognitive ability alone cannot explain the low validity in survey questions, they may accentuate response biases and contribute to making the PTs difficult to identify.

The role of enumerators and administration method

The contrast between the findings with STEP and other survey measures of the Big Five and internet-based measures suggests that the latter is more valid in non-WEIRD countries. To better understand this contrast, we focus on the possible role of enumerators and of the administration method. Enumerator-fixed effects explain on average 5.3% of the variation and are jointly significant in all of the 14 cases. The right bottom panel of Fig. 3 shows that, in some regions, enumerators explain up to 25% of the variation, a worryingly high share, indicating that enumerators can affect many of the responses possibly through the way they ask questions. Unexpectedly, enumerator effects are, if anything, more prevalent in regions with higher-average levels of cognitive abilities.

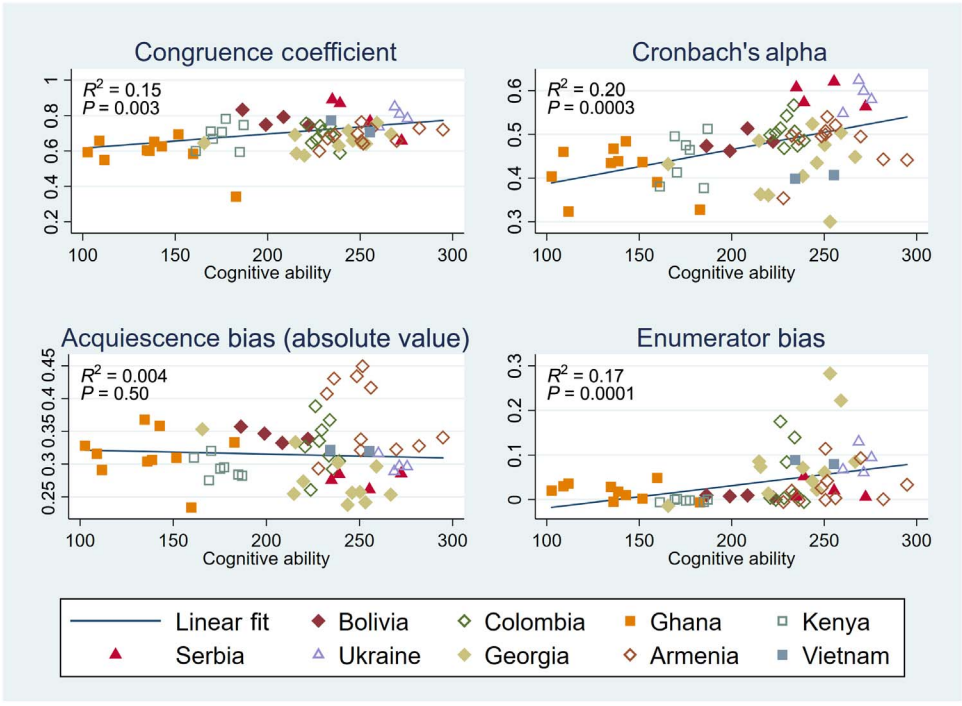

Fig. 3. Relationship between psychometric indicators and cognitive ability. In each figure, the level of observation is the largest possible geographical division in the country (regions, provinces, or district). We apply a weight that is the inverse of the number of geographical divisions to give the same weight to each country. The calculations of congruence coefficient, Cronbach's alpha, absolute value of acquiescence bias, and enumerator bias are described in Materials and Methods. Enumerator bias measures the share of the variation in responses (by PT) that can be explained by systematic biases due to which enumerator administrated each survey. Cognitive ability is measured by the full literacy test, also described in the Supplementary Materials. The nine countries in the regression are the nine countries with full literacy test included in the STEP surveys.

Because enumerators are rarely randomly allocated to subjects, the enumerator results may reflect other factors if assignment of respondents to enumerators captures other within-region variation, for instance, differences in gender, language, access, etc. We therefore use two of the datasets (non-STEP) analyzed (from Kenya and Colombia), where we randomly allocated enumerators to respondents, to isolate the enumerator effect from the selection effects previously mentioned. We find significant response biases of similar orders of magnitude as those found for most STEP regions. In Kenya, randomly assigned enumerators explain on average 9% of the variation in Big Five measures (significant for all 5 PTs) and in Colombia they explain 3% of such variation (significant for Extraversion and Agreeableness). The fact that the explanatory power of enumerator effects remains relatively strong even when enumerators are randomly assigned further points to the administration mode as a plausible explanation for the difference between face-to-face interviews and internet surveys. This is in line with other studies of Big Five data and other psychological data, using face-to-face interviews and internet data from Germany and Austria, which also conclude that response patterns such as acquiescence bias can depend on the characteristics of the particular measurement occasion (*51*, *52*). To directly test the importance of the mode of administration in developing country settings, we set up a survey experiment to test the effect of administrating a face-to-face survey compared to a paper-and-pencil survey. Our survey in Colombia included 330 farmers who completed primary education. Among them, about 40% were randomly selected to receive 22 BFI items administrated through face-to-face surveys and the other 60% answered the same items on paper on their own. As shown in table S9, Cronbach's alpha, within and between correlations, and congruence coefficient are all higher with self-administration compared to face-to-face surveys by enumerators, even if none of the differences are significant. That said, even with self-administration, these indicators remain below standards and below the ones obtained with internet data. By contrast, in a similar experiment, comparing self-administration and face-to-face interviews using the Big Five data in the German socioeconomic panel (*53*) found that the anticipated Big Five factor structure was present in the data, irrespective of the mode of data collection, but Rammstedt *et al.* (*54*) also show that response styles make the Big Five factor structure emerge in a blurry way at lower educational levels, whereas for highly educated persons, it emerged with textbook-like clarity. This then helps to reconcile the different findings: The survey method possibly affects response styles more in non-WEIRD populations with lower levels of education.

Overall, we cannot point to one single factor that explains the lower validity of PT questions in developing country surveys. The evidence presented and reviewed highlights several factors jointly at play. Previous research has emphasized the wording of the questions, the quality of the translations, and how those questions are interpreted in the culture (*25*). To these factors, the evidence presented in this paper adds as relevant the role of enumerators, stronger response biases in face-to-face interviews, and possibly respondents' self-selection and their ability to comprehend the items (cognition). More generally, our findings suggest that many of the known potential response pattern biases are accentuated in field survey data among less educated populations, making it hard to identify the intended latent traits. This association might be further explained by the lower interest in the survey

topic of typical survey respondents in these contexts, whose incentives and expectations are likely quite different from the ones of those filling in personality tests on the internet or in other WEIRD populations.

DISCUSSION

We show that although the BFI has been validated in specific countries and languages, one cannot assume that validity holds when administered in large-scale surveys amongst non-WEIRD populations. The lack of support for the FFM across a large set of surveys in diverse contexts indicates that the issues identified are not unique to a specific data collection exercise but point to a general problem in the measure of PTs through survey data in developing countries. Although the FFM may well be universal, it appears hard to uncover this factor structure with survey data in contexts other than those for which it was developed. The various psychometric indicators analyzed suggest that measurement error is correlated with cognition, which, in turn, is associated with income and other factors. Therefore, a set of items used as a proxy for a specific PT can also capture other factors and lead to incorrect inferences.

This psychometric evidence on the lack of a clear Big Five factor structure in survey data points to an interpretation of Fig. 1 that is illustrative of the potential consequences of ignoring the identified measurement error. Previous literature has found Conscientiousness and Emotional Stability to be the strongest predictors of income, and yet, we find Emotional Stability, but not Conscientiousness, to be a strong predictor. As illustrated in Table 1, Emotional Stability is the PT that best differentiates itself from the others and hence is likely to be the least affected by systematic response biases. Conscientiousness, on the other hand, is the PT that cumulates the highest number of misplaced items (with their highest factor loads on other PTs). Together, this can explain why Emotional Stability, but not Conscientiousness, stands out as a strong predictor. Besides this, and in contrast to most findings in WEIRD populations, Openness appears to be a strong predictor of income. However, given that Openness items poorly differentiate from the items of other PTs, one cannot exclude that this result is driven by systematic response biases and a combination of different factors captured by the measure. Hence, one should be very cautious before attributing the observed correlations to Openness.

Our results do not mean that there is no information content in the BFI questions of the surveys that we analyzed. They rather show that it also includes multiple forms of systematic response biases, which can be consequent enough to blur the distinctions between the different PTs. Several approaches can contribute to reduce these biases and should be applied when feasible, including: (i) balancing reversed and nonreversed items and using acquiescence bias corrections, (ii) self-administrated surveys, (iii) quality of translation and enumerator training with a particular effort to adapt to respondents with lower cognitive ability, and (iv) random assignment of enumerators. Assuring that enumerator assignment is balanced on the variables of interest (e.g., treatment variables in data collection for impact evaluation purposes) can also be important to assure that the enumerator effects do not bias the estimation of key relationships.

These findings have important implications for the interpretation of Big Five data collected through surveys in developing countries. Inferences about specific PTs will only be credible when preceded by a clear prior demonstration of the factor structure with the same data. When the FFM cannot be confirmed, as found in the vast majority of datasets analyzed in this paper, a more agnostic approach regarding the latent traits being measured by specific items is warranted. Along those lines, (*1*, *28*, *55*, *56*) use exploratory factor analysis to construct aggregate indices of latent noncognitive traits without necessarily assigning specific labels to the different factors. Using such an approach allows one to extract the pertinent latent information from the individual questions and to obtain an indicator of personality without interpreting them as being indicative of the importance of any specific PT. To the extent that these broader constructs are predictive of economic outcomes of interest, they can provide sufficient information for many research questions.

Last, our findings also call for the development of innovative methods that more accurately capture specific PTs in developing country surveys, which would need to focus on reducing the impact of response patterns and enumerator interactions, adapted to the context and mode of administration. This may involve task-based measures or indicators of behavior consistent with envisioned PTs, even if the context specificity of these measures may limit their general application. The results in this paper suggest that this improvement in measurement is a sine qua non for a better understanding of the role of PTs in developing countries.

MATERIALS AND METHODS

This section briefly explains the psychometric indicators used in the analysis.

Congruence coefficient and the alignment of factor loads

Tucker's congruent coefficient (or just congruence coefficient) is an index that assesses the similarity between factor structures of the same set of items applied to two different populations. PCA is a dimensionality reduction technique, which reduces a number of items into a smaller number of components (in our case five) that best explain the variation of the items. Each component is a weighted average of the items, and the vector of weights is also called the vector of loadings. The correlation between the vector of loadings in two different populations provides a measure of similarity between the two components.

After applying the PCA to two different populations, one calculates the correlation coefficient of the two vectors of loadings to assess the similarity between a component x and a component y

$$\varphi(\mathbf{x},\mathbf{y}) = \frac{\Sigma_{i,}\, x_i y_i}{\sqrt{(\Sigma_i\, x_i^2)(\Sigma_i\, y_i^2)}}$$

where $x_{i,\,j}$ and $y_{i,\,j}$ are the loadings of item i on factors x and y, respectively (each one extracted from applying the PCA of the same items to a different population). In our case, one population is the database of interest and the other population is one of the United States, which is used as the reference. Hence, a higher congruence coefficient indicates that the factor structure follows the one that has been found in the United States, where the Big Five PTs clearly stand out. To obtain the normative target, the Varimax orthogonal rotation was used for the U.S. data.

The congruence coefficient can be interpreted as a standardized measure of proportionality of elements in both vectors. A coefficient that is equal to 1 corresponds to a perfectly identical factor structure between the two populations, while a coefficient equal to 0 corresponds to a structure that is completely orthogonal.

The calculation of the congruence coefficient also requires a decision on the type of rotation to be applied to the survey data. The Procrustes rotation on target of the reference population is typically used in confirmatory factor analysis as a way to ease comparability. First, the factor

solution obtained from a replication sample is rotated orthogonally to conform to a predetermined factor structure (i.e., the target) as much as possible. Because the spatial orientation of factors in factor analysis is arbitrary, factor solutions obtained in different groups may be rotated in reference to each other to maximize their similarity. This is known as the Procrustes rotation or targeted rotation. Compared to other rotation choices, Procrustes tends to increase congruence coefficients. In the survey data, not applying the Procrustes rotation reduces the congruence coefficient, hence reinforcing the concerns raised about the factor structure (results available upon request). The low level of congruence in the survey data is especially notable given that we use the Procrustes rotation method, which specifically rotates the data to obtain the factor structure that most aligns with its target (the U.S. data).

The congruence coefficient is initially calculated by component or factor. We average it across the five factors to obtain a congruence coefficient that is an indicator of similar factor structure. To know which factor in the first population matches each factor of the second population, we calculate the average congruence coefficient for every possible combination and keep the one that maximizes the congruence coefficient. This explains why congruence coefficients tend to be higher when there are less items per factor (for reasons that are similar to overfitting when using a regression). To give an order of magnitude for the interpretation, Lorenzo-Seva and ten Berge (*50*) indicate that a congruence coefficient in a range of 0.85 to 0.94 shows fair similarity, whereas coefficients over 0.95 imply a high level of similarity (where the components can be considered equal).

In Table 1, we complement the congruence coefficient with a visual inspection of how the items sort themselves into the components for each database. To do this, each item was assigned to the component in which it has the highest loading. The red cells highlight items that are matched to the wrong component with respect to the FFM. One should take into account that, as explained above, each component was matched with a Big Five PT in the way that best aligns the factor structures, and despite this, we found on average of 4 (of 15) items per dataset that do not fit in the right component.

Within correlation, between correlation, and their interpretations

Within correlation refers to the average correlation of all combinations of two items that are within the same PT. By contrast, between correlation is obtained by calculating the average correlation of all combinations of two items of different PTs. Higher within correlation indicates that items of the same PT measure a common trait. However, one should not aim for a within correlation equal to 1 because items are noisy measures of the traits that they intend to capture and because even items of the same PT measure a slightly different dimension of the PT. In addition, the between correlation does not need to be 0 given that some degree of correlation is expected between the PT. That said, the extent to which the factor structure emerges depends on the difference between the within and between variations. When items of the same PT have a much higher correlation among them, they are more likely to be pooled together in the PCA (or factorial analysis).

Cronbach's alpha

Cronbach's alpha is one of the most widely used measures of internal consistency of a test. Cronbach's alpha is mathematically equivalent to the expected value of the split-half reliability. Split-half reliability is obtained by (i) randomly splitting the items into two sets of items of equal size, (ii) calculating the average of each set of items, and (iii) calculating the correlations between these two sets of items. Cronbach's alpha is equal to the average of the correlations obtained through all the possible combinations of split-half reliability. It provides an indicator of how well the items correlate among them (conditional on the number of items).

Assume that we have a measure X made of k items: $X = Y_1 + Y_2 + \ldots + Y_k$. Its Cronbach's alpha is given by

$$\alpha = \frac{K}{K-1}\left(1 - \frac{\sum_{i=1}^{K}\sigma^2_{Y_i}}{\sigma^2_X}\right)$$

where $\sigma^2_{Y_i}$ is the variance of item i and σ^2_X is the variance of the measure X.

Cronbach's alpha provides an assessment of both the construct's validity and its reliability. It decreases with measurement errors of the items and when the items tend to measure different latent constructs. A Cronbach's alpha of 0.9 tends to be required when individual decisions will be made based on a specific test (for example, for student's admissions) (*57*), but an alpha of 0.7 is often considered acceptable for the purpose of statistical analysis.

For a given within correlation c and number of items K, the Cronbach's alpha can be predicted by the Spearman-Brown prophecy formula, which is an increasing function of K

$$\alpha = \frac{c.K}{1 + c(K-1)}$$

This simply reflects the fact that a greater number of items reduces the noise of the aggregate index. Because the Big Five measure of the STEP data includes only three items per PT, one should expect a relatively low Cronbach's alpha. Hence, comparisons of Cronbach's alpha are arguably more meaningful when both datasets have the same number of items. Therefore, we show in Fig. 2 that Cronbach's alphas of the survey data are below the ones of the internet data (itself below the one of the U.S. data) for any number of items. Hence, Cronbach's alpha obtained from survey data is very low, even when taking into account the small number of items per construct.

Acquiescence bias estimation and correction

Acquiescent response style refers to the tendency of an individual to systematically agree (yea saying) or disagree (nay saying) with questionnaire items, regardless of their content. For example, consider the two following items, with answering scales from 1 (strongly agree) to 5 (strongly disagree):

- "Are you relaxed during stressful situations?"
- "Do you get nervous easily?"

They both aim at measuring the same PT (Emotional Stability), but only the second one is reverse-coded in the sense that a higher degree of agreement in the response is associated with a lower Emotional Stability. Hence, a respondent that strongly agrees with both statements shows a form of contradiction, indicative of a positive acquiescence bias. This response pattern ended up being a strong driver of the variation in the data. McCrae *et al.* (*48*) found that, when not corrected for, acquiescence bias came out as the first factor, highlighting the importance to correct it, which is now standard in psychometric analysis (*33*, *58*).

The calculation of the acquiescence bias (AB) requires questions that aim at measuring the same PT but of which at least one is reversed

and at least one is not reversed. The AB is calculated at the individual level. In the case of the 15 STEP items, because Agreeableness and Openness do not include any reverse question, AB can only be calculated using items of Conscientiousness, Extraversion, and Emotional Stability, but the correction is then applied to all items.

We calculate the acquiescence bias and apply the correction using the following steps:

1) Reverse the reverse-coded items. For example, if the possible answers range from 1 to 5, then answer 1 (fully disagreeing with a reverse-coded statement) is assigned a value of 5, answer 2 is assigned a value of 4, and so on.

2) For each PT that has at least one reverse-coded item and one nonreverse-coded item, calculate the average answer of reverse-coded items and the average answer of nonreverse-coded items.

3) For each PT, take the difference between the average of nonreverse-coded items and the average of reverse-coded items and divide it by 2.

4) The AB is the average of the differences obtained in (3) across all PTs

5) To correct for AB, add the AB obtained in (4) to every reverse-coded item and subtract the AB from every nonreverse-coded item.

The intuition of the AB correction is that in the absence of contradiction, in a scale from 1 to 5, the average raw answer between reversed items and nonreversed items should be 3 (the more one agrees with a statement, the more one should disagree with the opposite statement). Any average deviation from 3 is attributed to the AB and corrected by making the adjustment that will bring this average back to 3 after the correction.

In table S5, we show Cronbach's alpha, calculated by dataset and PT, using the items that were corrected for AB and without the correction for AB. We find that Cronbach's alphas are substantially higher after the AB correction, which indicates that the correction increased internal consistency. The within correlations, as well as the difference between the between and within correlations, are also lower without the correction. Furthermore, while the calculation of Cronbach's alpha uses the absolute correlation between the items, the dagger (†) in table S5 indicates that at least one of the correlations between two items belonging to the same PT is negative. Without acquiescence bias correction, we found a large number of cases of these negative correlations. This is driven by negative correlations between reverse and nonreverse items, suggesting that the AB response pattern is more influential than the PT that the items aim to measure.

Enumerator bias

Enumerator bias refers to any way in which enumerators systematically influence the answers of the respondents. It can result from differences in the way in which questions are asked or in the way in which responses are registered. For example, some enumerators may unwillingly react more positively to answers that reflect higher levels of skills, giving the impression to the respondent that these are the "right answers." In addition, while enumerators are expected to let the respondent pick one of the answers proposed on a Likert scale, enumerators may give their own interpretation to answers that do not fall in the standardized scale or provide subtle leads on what the answer of the respondent implies, all of which can accentuate the role of the enumerators.

If each enumerator has a tendency to systematically bias answers in a given direction, then enumerator identifiers will explain some of the variation in the answers. To quantify this bias, we analyzed how much of the variation in the Big Five indicators can be explained by a set of enumerator dummies (indicating which enumerator administrated each survey). We did so separately for each PT and took into account the fact that other factors may be correlated to the assignment of enumerators. One common limitation is that different enumerators work in different areas, in which case the explanatory power of enumerators would capture features of the respondent rather than enumerator biases. To filter out the effect of the geographical area, we first run a regression of the PT index on dummies for the smallest geographical division available and name R^2_{gt} the R^2 obtained for the PT t. We then ran the regression of the PT t on dummies of geographical division and enumerators and named its R^2 R^2_{get}. Thus, the additional explanatory power of PT t provided by the enumerator dummies is equal to $R^2_{get} - R^2_{gt}$, and the variance that remains to be explained after filtering out the variation by geographical unit is given by $1 - R^2_{gt}$. Hence, $R^2_{et} = \frac{R^2_{get} - R^2_{gt}}{1 - R^2_{gt}}$ is an estimation of the explanatory power of enumerators on the remaining variance, after filtering out the effect of the geographical division. Last, to obtain an average across PTs, we calculated $R^2_e = \frac{\sum_{t=1}^{t=5} R^2_{et}}{5}$, where $t = 1, \ldots, 5$ refers to the five PTs. R^2_e is our estimation of the enumerator bias estimated within each region and is represented in Fig. 3.

SUPPLEMENTARY MATERIALS

Supplementary material for this article is available at http://advances.sciencemag.org/cgi/content/full/5/7/eaaw5226/DC1

Additional data description

Fig. S1. Prediction of income with imperfect proxy of cognitive ability and the Big Five PTs.

Table S1. Descriptive statistics of STEP data.

Table S2. Descriptive statistics of other survey data and reference data.

Table S3. Descriptive statistics of internet data.

Table S4. Psychometric indicators by database, using data without correcting for acquiescence bias.

Table S5. Cronbach's alpha by PT, using STEP survey data, without versus with acquiescence bias correction.

Table S6. Psychometric indicators by dataset.

Table S7. Cronbach's alpha by PT and database.

Table S8. Average item-by-item correlation coefficients in different databases.

Table S9. Psychometric indicators for Colombia, comparing randomly assigned face-to-face versus self-administrated surveys.

REFERENCES AND NOTES

1. J. J. Heckman, J. Stixrud, S. Urzua, The effects of cognitive and noncognitive abilities on labor market outcomes and social behavior. *J. Labor Econ.* **24**, 411–482 (2006).
2. B. W. Roberts, N. R. Kuncel, R. Shiner, A. Caspi, L. R. Goldberg, The power of personality: The comparative validity of personality traits, socioeconomic status, and cognitive ability for predicting important life outcomes. *Perspect. Psychol. Sci.* **2**, 313–345 (2007).
3. R. R. McCrae, P. T. Costa Jr., Personality trait structure as a human universal. *Am. Psychol.* **52**, 509–516 (1997).
4. O. P. John, S. Srivastava, The Big Five Trait taxonomy: History, measurement, and theoretical perspectives, in *Handbook of Personality: Theory and Research*, L. A. Pervin, O. P. John, Eds. (Guilford Press, 1999), pp. 102–138.
5. M. Almund, A. Duckworth, J. Heckman, T. Kautz, Personality Psychology and Economics, in *Handbook of the Economics of Education*, E. Hanushek, S. Machin, L. Woessman, Eds. (Elsevier, Amsterdam, 2011), pp. 1–181.
6. L. Borghans, A. L. Duckworth, J. J. Heckman, B. ter Weel, The economics and psychology of personality traits. *J. Hum. Resour.* **43**, 972–1059 (2008).
7. F. Finan, B. A. Olken, R. Pande, The Personnel Economics of the State, in *Handbook of Economic Field Experiments*, A. V. Banerjee, E. Duflo, Eds. (Elsevier, 2017).
8. R. Hanna, S.-Y. Wang, Dishonesty and selection into public service: Evidence from India. *Am. Econ. J. Econ. Policy* **9**, 262–290 (2017).
9. E. Dal Bó, F. Finan, N. Li, L. Schechter, Government Decentralization Under Changing State Capacity: Experimental Evidence from Paraguay (NBER Working Paper no. 24879, 2018).

10. M. C. Araujo, P. Carneiro, Y. Cruz-Aguayo, N. Schady, Teacher quality and learning outcomes in kindergarten. *Q. J. Econ.* **131**, 1415–1453 (2016).
11. K. Donato, G. Miller, M. Mohanan, Y. Truskinovsky, M. Vera Herndández, Personality traits and performance contracts: Evidence from a field experiment among maternity care providers in India. *Am. Econ. Rev.* **107**, 506–510 (2017).
12. E. Dal Bó, F. Finan, M. A. Rossi, Strengthening state capabilities: The role of financial incentives in the call to public service. *Q. J. Econ.* **128**, 1169–1218 (2013).
13. M. Groh, D. McKenzie, T. Vishwanath, Reducing information asymmetries in the youth labor market of jordan with psychometrics and skill based tests. *World Bank Econ. Rev.* **29**, S106–S117 (2015).
14. G. Calderon, L. Iacovone, L. Juarez, Opportunity versus necessity: Understanding the heterogeneity of female micro-entrepreneurs. *World Bank Econ. Rev.* **30**, S86–S96 (2017).
15. M. Groh, D. McKenzie, N. Shammouth, T. Vishwanath, Testing the importance of search frictions and matching through a randomized experiment in Jordan. *IZA J. Lab. Econ.* **4**, 7 (2015).
16. B. Klinger, A. Khwaja, C. del Carpio, *Entreprising Pschychometrics and Poverty Reduction* (Springer, 2013), Springer Briefs in Psychology/Springer Briefs in Innovations in Poverty Reduction Series.
17. I. Arráiz, M. Bruhn, R. Stucchi, Psychometrics as a tool to improve credit information. *World Bank Econ. Rev.* **30**, S67–S76 (2017).
18. M. Ibanez, G. Riener, Sorting through affirmative action: Three field experiments in Colombia. *J. Labor Econ.* **36**, 437–478 (2018).
19. C. Blattman, J. C. Jamison, M. Sheridan, Reducing crime and violence: Experimental evidence from cognitive behavioral therapy in Liberia. *Am. Econ. Rev.* **107**, 1165–1206 (2017).
20. F. Campos, M. Frese, M. Goldstein, L. Iacovone, H. Johnson, D. McKenzie, M. Mensmann, Teaching personal initiative beats traditional training in boosting small business in West Africa. *Science* **357**, 1287–1290 (2017).
21. V. Benet-Martinez, O. P. John, *Los Cinco Grandes* across cultures and ethnic groups: Multitrait-multimethod analyses of the Big Five in Spanish and English. *J. Pers. Soc. Psychol.* **75**, 729–750 (1998).
22. O. P. John, L. P. Naumann, C. J. Soto, Paradigm Shift to the Integrative Big-Five Trait Taxonomy: History, Measurement, and Conceptual Issues, in *Handbook of Personality: Theory and Research*, O. P. John, R. W. Robins, L. A. Pervin, Eds. (Guilford Press, 2008), pp. 114–158.
23. R. R. McCrae, P. Costa, *NEO PI-R Professional Manual* (Psychological Assessment Resources, Inc., 1992).
24. R. R. McCrae, A. Terracciano, Universal features of personality traits from the observer's perspective: Data from 50 cultures. *J. Pers. Soc. Psychol.* **88**, 547–561 (2005).
25. R. L. Piedmont, E. Bain, R. R. McCrae, P. T. Costa, The Applicability of the Five-Factor Model in a Sub-Saharan Culture: The NEO-PI-R in Shona, in *International and Cultural Psychology Series. The Five-Factor Model of Personality Across Cultures*, R. R. McCrae, J. Allik, Eds. (Kluwer Academic/Plenum Publishers, 2002), pp. 155–173.
26. J. Henrich, S. J. Heine, A. Norenzayan, The weirdest people in the world? *Behav. Brain Sci.* **33**, 61–83 (2010).
27. M. Gurven, C. Von Rueden, M. Massenkoff, H. Kaplan, M. Lero Vie, How universal is the Big Five? Testing the five-factor model of personality variation among forager–farmers in the Bolivian Amazon. *J. Pers. Soc. Psychol.* **104**, 354–370 (2013).
28. R. Laajaj, K. Macours, Measuring Skills in Developing Countries (CEPR discussion paper no. 13271, 2018).
29. S. G. Ludeke, E. G. Larson, Problems with the Big Five assessment in the World Values Survey. *Pers. Indiv. Dif.* **112**, 103–105 (2017).
30. B. Rammstedt, O. P. John, Measuring personality in one minute or less: A 10-item short version of the Big Five inventory in English and German. *J. Res. Pers.* **41**, 203–212 (2007).
31. S. D. Gosling, P. J. Rentfrow, W. B. Swann Jr., A very brief measure of the Big-Five personality domains. *J. Res. Pers.* **37**, 504–528 (2003).
32. A. T. Church, in *The Praeger Handbook of Personality across Cultures*, vol. 1, *Trait psychology* (ABC-CLIO/Praeger, 2017).
33. C. J. Soto, O. P. John, S. D. Gosling, J. Potter, The developmental psychometrics of Big Five self-reports: Acquiescence, factor structure, coherence, and differentiation from ages 10 to 20. *J. Pers. Soc. Psychol.* **94**, 718–737 (2008).
34. O. P. John, E. M. Donahue, R. L. Kentle, *The Big Five Inventory – Versions 4a and 54* (University of California, Berkeley, Institute of Personality and Social Research, 1991).
35. C. J. Soto, O. P. John, Ten facet scales for the Big Five inventory: Convergence with NEO PI-R facets, self-peer agreement, and discriminant validity. *J. Res. Pers.* **43**, 84–90 (2009).
36. G. Pierre, M. L. Sanchez Puerta, A. Valerio, T. Rajadel, STEP Skills Measurement Surveys - Innovative Tools for Assessing Skills (Social protection and labor discussion paper no. 1421, World Bank Group, 2014).
37. A. Andrew, O. Attanasio, E. Fitzsimons, S. Grantham-McGregor, C. Meghir, M. Rubio-Codina, Impacts 2 years after a scalable early childhood development intervention to increase psychosocial stimulation in the home: A follow-up of a cluster randomised controlled trial in Colombia. *PLOS Med.* **15**, e1002556 (2018).
38. O. Attanasio, H. Baker-Henningham, R. Bernal, C. Meghir, D. M. Pineda, M. Rubio-Codina, Early Stimulation: The impacts of a scalable intervention (NBER Working Paper No. 25059, 2018).
39. S. De Mel, D. McKenzie, C. Woodruff, Returns to capital in microenterprises: Evidence from a field experiment. *Q. J. Econ.* **123**, 1329–1372 (2008).
40. S. De Mel, D. McKenzie, C. Woodruff, One-time transfers of cash or capital have long-lasting effects on microenterprises in Sri Lanka. *Science* **335**, 962–966 (2012).
41. S. de Mel, D. McKenzie, C. Woodruff, Business training and female enterprise start-up, growth, and dynamics: Experimental evidence from Sri Lanka. *J. Dev. Econ.* **106**, 199–210 (2014).
42. X. Giné, D. Yang, Insurance, credit, and technology adoption: Field experimental evidencefrom Malawi. *J. Dev. Econ* **89**, 1–11 (2009).
43. A. Pranita, A. Andrew, M. Das, A. Gautam, M. Huepe, S. Krutikova, S. Kumar, S. Sharma, R. Soni, H. Verma, R. Verma, *Promoting Adolescent Engagement, Knowledge and Health Evaluation of PAnKH : An adolescent girl intervention in Rajasthan* (India Baseline Report, IFS, 2016); www.ifs.org.uk/publications/8701.
44. W. Bleidorn, T. A. Klimstra, J. J. A. Denissen, P. J. Rentfrow, J. Potter, S. D. Gosling, Personality maturation around the world: A cross-cultural examination of social-investment theory. *Psychol. Sci.* **24**, 2530–2540 (2013).
45. S. D. Gosling, S. Vazire, S. Srivastava, O. P. John, Should we trust web-based studies? A comparative analysis of six preconceptions about Internet questionnaires. *Am. Psychol.* **59**, 93–104 (2004).
46. S. Srivastava, O. P. John, S. D. Gosling, J. Potter, Development of personality in early and middle adulthood: Set like plaster or persistent change? *J. Pers. Soc. Psychol.* **84**, 1041–1053 (2003).
47. L. R. Goldberg, A broad-bandwidth, public domain, personality inventory measuring the lower-level facets of several five-factor models, in *Personality Psychology in Europe*, I. Mervielde, I. Deary, F. De Fruyt, F. Ostendorf, Eds. (Tilburg University Press, 1999), vol. 7, pp. 7–28.
48. R. R. McCrae, J. H. Herbst, P. T. Costa Jr., Effects of acquiescence on personality factors structures, in *Personality and Temperament: Genetics, Evolution, and Structure*, R. Riemann, F. M. Spinath, F. Ostendorf, Eds. (Pabst Science Publishers, 2001), pp. 217–231.
49. D. P. Schmitt, J. Allik, R. R. McCrae, V. Benet-Martínez, The geographic distribution of Big Five personality traits: Patterns and profiles of human self-description across 56 nations. *J. Cross Cult. Psychol.* **38**, 173–212 (2007).
50. U. Lorenzo-Seva, J. M. F. ten Berge, Tucker's congruence coefficient as a meaningful index of factor similarity. *Methodology* **2**, 57–64 (2006).
51. J. Aichholzer, Intra-individual variation of extreme response style in mixed-mode panel studies. *Soc. Sci. Res.* **42**, 957–970 (2013).
52. D. Danner, J. Aichholzer, B. Rammstedt, Acquiescence in personality questionnaires: Relevance, domain specificity, and stability. *J. Res. Pers.* **57**, 119–130 (2015).
53. L. Hilgert, M. Kroh, D. Richter, The effect of face-to-face interviewing on personality measurement. *J. Res. Pers.* **63**, 133–136 (2016).
54. B. Rammstedt, L. R. Goldberg, I. Borg, The measurement equivalence of Big-Five factor markers for persons with different levels of education. *J. Res. Pers.* **44**, 53–61 (2010).
55. F. Cunha, J. J. Heckman, S. M. Schennach, Estimating the technology of cognitive and non-cognitive skill formation. *Econometrica* **78**, 883–931 (2010).
56. O. Attanasio, S. Cattan, E Fitzsimons, C. Meghir, M. Rubio-Codina, Estimating the Production Function for Human Capital: Results from a Randomized Control Trial in Colombia (NBER Working Paper No. 20965, 2017).
57. J. C. Nunnally, I. H. Bernstein, *Psychometric Theory* (McGraw-Hill, ed. 3, 1994).
58. B. Rammstedt, C. J. Kemper, I. Borg, Correcting Big Five personality measurements for acquiescence: An 18-country cross-cultural study. *Eur. J. Pers.* **27**, 71–81 (2013).

Acknowledgments: We thank the editor and two anonymous referees for helpful comments. We gratefully acknowledge A. Andrew, O. Attanasio, R. Bernal, F. Finan, E. Fitzsimons, S. Grantham-McGregor, S. Krutikova, D. McKenzie, C. Meghir, L. Schechter, C. Soto, C. Woodruff, and the World Bank's STEP team for providing access to different datasets used in this paper and for comments received on an earlier draft. The activities associated with this research received institutional review board approval from the Paris School of Economics; Universidad de Los Andes, Bogota; and the University of Texas. The findings, interpretations, and conclusions expressed are entirely those of the authors and do not necessarily reflect the views of the World Bank or the Inter-American Development Bank, their Board of Directors, or the countries they represent. **Funding**: This research was funded by the World Bank and the French National Research Agency (ANR) under grant ANR-17-EURE-0001. **Author contributions**: R.L., K.M., M.R.-C., R.V., and O.A. conceived the study; R.L., K.M., and

D.A.P.H. analyzed the data; S.D.G. and J.P. collected the internet data; R.L. and K.M. wrote the paper; and S.D.G., M.R.-C., and D.A.P.H. critically revised the manuscript.
Competing interests: The authors declare that they have no competing interests.
Data and materials availability: All data needed to evaluate the conclusions in the paper are present in the paper and/or the Supplementary Materials. Additional data related to this paper may be requested from the authors. All programs and details on how to access the different databases used are provided on www.laajaj.com/research for replication purposes.

Submitted 1 January 2019
Accepted 5 June 2019
Published 10 July 2019
10.1126/sciadv.aaw5226

Citation: R. Laajaj, K. Macours, D. A. Pinzon Hernandez, O. Arias, S. D. Gosling, J. Potter, M. Rubio-Codina, R. Vakis, Challenges to capture the big five personality traits in non-WEIRD populations. *Sci. Adv.* **5**, eaaw5226 (2019).

Science Advances

Challenges to capture the big five personality traits in non-WEIRD populations

Rachid Laajaj, Karen Macours, Daniel Alejandro Pinzon Hernandez, Omar Arias, Samuel D. Gosling, Jeff Potter, Marta Rubio-Codina and Renos Vakis

Sci Adv **5** (7), eaaw5226.
DOI: 10.1126/sciadv.aaw5226