Scalar Equivalence of OPQ32: Big Five Profiles of 31 Countries

Dave Bartram

Journal of Cross-Cultural Psychology 2013 44: 61 originally published online 28 December 2011
DOI: 10.1177/0022022111430258

Journal of Cross-Cultural Psychology
44(1) 61–83

DOI: 10.1177/0022022111430258
jccp.sagepub.com
SAGE

Scalar Equivalence of OPQ32: Big Five Profiles of 31 Countries

Dave Bartram[1]

Abstract

Scalar equivalence of Big Five scale scores, derived from OPQ32i data, for over one million people are reviewed in terms of differences between 31 countries involving over 20 different languages. Strong relationships are found between country average scale scores and country standard deviations (*SD*s), on the one hand, and two of Hofstede's cultural dimensions, on the other. Country *SD*s are also seen to vary with cultural "tightness" ratings. Country-level performance indicators are also examined (the World Economic Forum Global Competitiveness Index and UN Human Development indices). Strong correlations are found between these indicators and both country-level mean personality scores and *SD*s of personality scores. While Hofstede's dimensions also predict variation in global competitiveness ($R = 0.66$), adding OPQ32 Big Five personality scale data increases the level of prediction to $R = 0.84$. It is argued that the strength of these relationships with independent country-level metrics supports the view that between-country differences represent true score variance rather than systematic instrument-related biases.

Keywords

Big Five personality factors, OPQ32, global competitiveness, Hofstede, UN Human Development Indices

There has long been an interest in whether differences in personality scale score averages across countries or cultures are meaningful or due to sources of systematic bias (see, for example, Barrett & Eysenck, 1984; Lynn & Martin, 1995; McCrae, 2001, 2002; McCrae, Terracciano et al, 2005a, 2005b; Schmitt, Allik, McCrae, & Benet-Martinez, 2007). The present article extends this research to consider a large data set from an instrument that is widely used around the world in the field of occupational assessment (i.e., for selection or development in the workplace) but which has received little attention in the field of cross-cultural psychology. The OPQ32 is a 32-scale personality inventory. It was originally developed in 1999 from the earlier OPQ Concept Model and was produced in two forms: OPQ32n used single item statements each rated on 5-point Likert-type scales, and OPQ32i used forced-choice item quads—sets of four statements

[1]SHL Group Ltd, Thames Ditton, UK

Corresponding Author:
Dave Bartram, SHL Group Ltd, The Pavilion, 1 Atwell Place, Thames Ditton, Surrey, KT7 0NE, United Kingdom.
Email: dave.bartram@shl.com

from which the candidate chooses one as "most like me" and one as "least like me" (SHL, 1999, 2006). OPQ32i has been widely used around the world and has been translated and adapted into more than 30 languages. Most recently, the OPQ32i has evolved into OPQ32r, which uses forced-choice item triplets with a multidimensional Item Response Theory (IRT) scoring model to recover normative scale data from the forced-choice item formats (Brown & Maydeu-Olivares, 2011; SHL, 2009).

OPQ32 measures 32 work-related personality traits from which "Big Five" scale scores can be produced (by scale aggregation: Bartram & Brown, 2005). Bartram (2012) has reported on a series of studies that have demonstrated construct invariance for the OPQ32i scales across a wide range of countries and between ethnic groups within countries such as South Africa.

The focus of the present article is on the variation in average scale scores across countries and looks at the degree to which such variation, where it occurs, can be accounted for by independent country-level measures of culture and other "outcome" measures. This research was carried out primarily to see whether scalar equivalence across different country versions of OPQ32 could be established and, therefore, the aggregation of data across countries justified. This is a necessary step in building norm groups that are suitable for use in multinational assessment programs (Bartram, 2008). While the primary aim of the present research was to assess evidence in support of scalar equivalence of scale scores across countries, it also provides an opportunity to test the validity of the OPQ32i-based Big Five personality scores by investigating their associations with a range of country-level criterion measures and with country-level scores based on other Big Five measures.

The research is of general relevance to the question of how stable cross-cultural effects are, to what extent they reflect culture-related or instrument-related bias in the way instruments are responded to, and to what extent they reflect true score differences in the underlying traits that are being measured. For this reason, the present article focuses on analyses in terms of the Big Five personality model as this has been widely used in previous research, notably by McCrae (McCrae, 2001, 2002; McCrae & Terracciano, 2007; McCrae, Terracciano et al, 2005a, 2005b) and Schmitt (Schmitt et al., 2007; Schmitt, Realo, Voracek, & Allik, 2008). Earlier work by Barrett and Eysenck (1984), later expanded by Lynn and Martin (1995), is also of relevance and will be considered.

The Problem of Establishing Scalar Equivalence

Scalar equivalence of a trait measure across two groups (Van de Vijver & Leung, 1997) requires that a given raw score represents the same amount of the trait in both groups. If there is some systematic bias, such that for Group A raw scores are always inflated by a couple of points, then scalar equivalence could be achieved by an adjustment to the raw scores. In practice, the identification of such systematic biases is difficult to achieve. There is no single test that can prove scalar equivalence; rather, it is established through the accumulation of evidence. As a first step, it is necessary to show that the constructs being measured are equivalent. This can be achieved by demonstrating invariance in the pattern of relationships between measures across groups, either by showing invariance in the group-related correlation matrices or in the instrument's factor structure, where there is a well-defined factor model.

Scalar equivalence can be demonstrated in a bottom-up or a top-down fashion. The bottom-up approach uses individual-level data and is based on examination of item-level differential item functioning (DIF) and bilingual retest studies. Both of these approaches have problems with them (see McCrae & Terracciano, 2007, for a discussion). For example, a measure can have scalar equivalence even when all its items have DIF so long as the DIF is random. What is more, DIF will not detect systematic response biases such as the group-related differences in acquiescence effects one can get with item formats using Likert-type scales.

The use of bilingual studies is based on the fact that the people responding to the two versions of the instrument have a fixed trait level associated with the measure, and hence should produce the same raw scores in both language versions or, if scores differ, be usable as a basis for equating any systematic differences. A study of Greek-English bilinguals is reported in the OPQ32 Technical Manual (SHL, 2006, pp. 71-72). This showed high retest correlations between the two language versions (median $r = 0.86$) with generally small differences between scale means. Only two scales showed significant difference, with effects sizes of $d = 0.15$ and $d = 0.13$. These results are very similar to what one would expect from retesting in the same language.

However, there are problems in using bilingual studies to support scalar equivalence across countries. To do so requires assuming that bilinguals are typical of people in both countries (i.e., for Greek-English bilinguals, England and Greece). In practice, bilinguals tend to be from one country and to be fluent in a second language. As such, they can only demonstrate the quality of the translation and not of the cultural adaptation.

The top-down approach to scalar equivalence argues that if there is scalar equivalence at the individual level, then any differences in means or *SD*s at the group level must represent true score variance and not systematic measurement error. Suppose, for example, that trait X is related to criterion Y. If group A is low on this criterion measure and group B is high, we would expect to find a difference in the mean levels of trait X between the two groups. If the groups are the same on criterion Y, then any difference in average trait level would tend to indicate a lack of scale invariance. This assumes, of course, that group differences in the criterion measure and in the trait measure are not both affected by some common third source of systematic bias. However, any such source of bias would be instrument- and method-independent (i.e., it might relate instead to some separate cultural characteristic that affected both the trait expression and the criterion measure).

Thus, the most powerful approach to testing for scalar invariance is to see whether group-level differences in means and *SD*s of the measures in question are systematically related to a range of other independent measures of group-level effects. McCrae and Terracciano (2007) report on data from two projects relevant to this. The first (from McCrae, 2001, 2002) reported secondary analyses of self-report data on the NEO-PI-R from volunteers in 36 cultures around the world, while the second (McCrae, Terracciano et al, 2005a, 2005b) considered observer-rating data from college students from 51 cultures. In both cases, the data were considered in relation to the Five-Factor model of personality. NEO-PI-R (Costa & McCrae, 1992) is a 240-item instrument providing measures of 30 facet scales, six for each of the five personality factors: Neuroticism, Extraversion, Agreeableness, Openness, and Conscientiousness. Schmitt et al. (2007) report data from 56 nations using a different instrument: the BFI (Benet-Martinez & John, 1998; John & Srivastrava, 1999). This is a short 44-item self-report Big Five measure and was included in the International Sexuality Description Project. While NEO-PI-R attempts to control for acquiescence response bias by a balance of negative and positive items, the BFI items are generally all positive in direction.

Other than these, few studies have been reported of country-level effects in personality and their relation to other measures. This is due to the difficulty, in the past, of getting large samples of data across large numbers of countries using an instrument that has demonstrated construct invariance. Lynn and Martin (1995) related Eysenck Personality Questionnaire (EPQ) data from 37 countries to a range of social indicators, finding that country-level Extraversion was related to homicide rate (positive) and suicide rate (negative). Diener, Diener, and Diener (1995) found that national levels of a subjective well-being measure (SWB) were strongly positively related to income, individualism, human rights, and societal equality. Steel and Ones (2002) found that SWB was related to Neuroticism and Extraversion at the aggregate level in the same way as at the individual level for 39 cultures (with the EPQ) and 24 cultures (with the NEO-PI-R). McRae

and Terracciano (2007) report a large number of relationships between personality scores and aggregate country differences in cancer rates, life expectancy, substance abuse, and obesity. Many of these effects were partly or wholly accounted for by country differences in GDP. They conclude that their results are "not easy to interpret, nor particularly impressive" (p. 261). In particular, they cite variables like GDP as accounting for large amounts of between-group variance and being covariate with other measures. While this is a problem if one is trying to disentangle the individual and social factors that cause differences in cancer rates, it is not a problem if one is simply trying to establish scalar equivalence of a measure. For the latter task, the presence of strong correlations with GDP, for example, is sufficient to indicate that between-group differences on the trait measure are unlikely to be measurement bias. The fact that they may be difficult to interpret is not an issue in relation to evidencing the scalar equivalence of the instrument across groups, but is an issue for understanding culture-level phenomena.

What Is "Culture" and When Does It Matter?

There are many different definitions of *culture*. For example, Hofstede (1980) defined *culture* as "the collective programming of the mind which distinguishes the members of one human group from another" (p. 25). For our present purposes, we can say that culture is a set of exogenous variables relating to shared values, cognitions, knowledge, language, and standards or cultural norms. In practical assessment terms, culture only matters when it is related to a group of people for whom within-group variability on relevant constructs is relatively small compared to variability between them and other groups.

One of the dominant cultural values frameworks was developed by Hofstede (1980, 2001) originally using survey data from IBM employees from 40 countries (88,000 employees in 72 countries were surveyed, but small country samples were excluded). Though based on one organization, this work has had an enormous impact on thinking in cross-cultural research and has tended to overshadow other frameworks (e.g., Trompenaars, 1993). Hofstede defined four main dimensions:

1. Individualism (IDV) is "the degree to which people in a country prefer to act as individuals rather than as members of groups" (Hofstede, 1994, p. 6). This dimension is often referred to as "individualism-collectivism."
2. Power-distance (PDI) relates to the extent to which it is accepted that the power vested in institutions and organizations is distributed unequally. This gets reflected in the degree to which people feel free to disagree with their superiors and the degree to which their superiors feel the need to consult.
3. Masculinity (MAS), often referred to as "masculinity-femininity," contrasts valuing assertiveness, success, competition over personal relationships, caring for the weak, and cooperation.
4. Uncertainty Avoidance Index (UAI) reflects a preference for clear formal rules and guidance. High scores are associated with intolerance for deviant behaviors and a belief in absolute truths.

Until recently, research on Hofstede's dimensions has been mainly subject to qualitative review (e.g., Kirkman, Lowe, & Gibson, 2006). A relatively limited quantitative review was carried out by Oyserman, Coon, and Kemmelmeier (2002) of the Individualism-Collectivism dimension along with other cultural value frameworks (e.g., from the GLOBE project: House, Hanges, Javidan, Dorfman, & Gupta, 2004), which looked at 83 studies carried out between 1980 and 2000. Taras, Kirkman, and Steel (2010), however, have recently examined the

relationships between Hofstede's original four cultural dimensions and organizationally relevant outcomes through a meta-analysis of 598 studies. In relation to cultural tightness, Taras et al. (2010) found some evidence to support the role of this construct as a moderator of the effects of cultural value. Cultural values were found to have stronger effects on outcomes in culturally tighter ($\rho = 0.28$) rather than looser countries. In terms of limitations, this meta-analysis, like all meta-analyses, is limited by the quality and quantity of the studies it reviews.

Gelfand, Nishii, and Raver (2006) expanded on Hofstede's dimensions by adding the construct of cultural tightness-looseness, defining this as the degree to which institutions in society promote narrower limits on socialization and have higher levels of constraint and systems for monitoring and sanctioning behavior. While there may be some overlap between Hofstede's constructs and the construct of tightness, Gelfand et al. (2006) propose that variance in individual attributes would be lower in tight than in loose cultures. In other words, the tighter the cultural constraints the more the behavior of individuals will be predicted by the cultural norms rather than by individual differences. This is important as it implies that the variability of individual behavior may be subject to change as well as the "norm" for that behavior. By implication, the tighter the cultural control, the less validity measures of individual difference will have as predictors of workplace behavior (even if the variability in the predictors is not constrained). Gelfand et al. (2011) report the results of a 33-nation study that explores a range of differences between countries that relate to "tightness scores" (with tightness being measured by a six-item rating scale). Their results indicate that "tight" nations, among other things, are more autocratic, have less open media with more controls, fewer civil liberties, lower crime rates, and less participation in collective action.

Introduction to the Studies

The three studies presented in this article examine the relationships between country-level OPQ32i Big Five scale means and *SD*s, on the one hand, and other country-level measures, on the other hand. For Study 1 the other measures are Hofstede's culture dimensions and Gelfand's index of "tightness" (for a subset of the countries). Here it is expected from previous research that findings relating personality to individualism and power-distance will be replicated with Neuroticism (N) and Extraversion (E), in particular, strongly relating to these cultural dimensions: with low N and high E being associated with high individualism and low power-distance. Previous research (e.g., McCrae, 2002) has tended to find that *SD*s in general are reduced in high power-distance collectivist cultures relative to low power-distance individualistic ones. We would expect to find negative correlations between Big Five *SD*s and Gelfand's "tightness" measures as well as finding that *SD*s decrease as IDV decreases and PDI increases.

It could be argued that Hofstede's cultural dimension scales sit in an analogous construct space at the country level to personality measures at the individual level of measurement. In a sense they describe the "personality" of a culture. For Study 2, the relationships with more independent national "performance" or "outcome" measures are examined (global competitiveness, life expectancy, educational provision, etc.).

Study 3 compares the present results with those from previous research, notably the published data from Lynn and Martin (1995), McCrae (2002), McCrae, Terracciano et al. (2005b), and Schmitt et al. (2007). The comparisons with Hofstede and with country "performance" measures are reexamined in the light of similarities and differences with previous research.

While the OPQ32 produces 32 scale scores, the emphasis in this article is on measurement in terms of the aggregation of a subset of 25 of those scales into Big Five model measures. This both provides a common basis for comparisons with earlier research using different instruments and reduces the problems of capitalizing on chance findings that one would have in

considering 32 scales with data from only 31 countries. For that reason, the next section looks in more detail at the OPQ32 as a Big Five measure.

The OPQ32 and the Big Five

OPQ32i (SHL, 1999, 2006) is the forced-choice item format version of OPQ32 and is very widely used in occupational assessments around the world for both selection and development. It is available in more than 30 languages. It consists of 104 sets of item quads: Each quad is a set of four statements, where each statement relates to a different scale. For each quad, the candidate chooses one statement as "most like me" and one as "least like me."

One of the key advantages of forced-choice item formats is that they provide control over systematic rating biases that can affect cross-country and cross-cultural comparisons. They have also been shown to be more fake-resistant in high-stakes occupational assessment settings than conventional Likert-type rating format items (Christiansen, Burns, & Montgomery, 2005; Griffith, & McDaniel, 2006; Jackson, Wroblewski, & Ashton, 2000; Martin, Bowen, & Hunt, 2002).

For OPQ32i, traditional scoring methods produce 32 ipsative scales (i.e., a fixed total number of points, 416, are distributed between the 32 scales). More recently, multidimensional IRT scoring has been developed that recovers normative trait scores from the patterns of choices people have made. Details of the relationship between the ipsative and normative trait measures obtained from the same forced-choice format are presented in the OPQ32r Technical Manual (SHL, 2009), and further details of the measurement methodology can be found in Brown and Maydeau-Olivares (2011). With 32 scales, the differences between the normative and ipsative scoring methods are small (Baron, 1996; Bartram, 1996), especially where subsets of individual scales are aggregated. Use of a subset of the 32 scales, such as for their aggregation into Big Five measures, removes the strict "ipsative constraint," which is present when all 32 scales are considered.

For the Big Five, 25 of the 32 OPQ scales are used with positive or negative loadings as shown in Table 1. The process employed to map the OPQ32 to the Five Factor Model (FFM) and establish the scales' weights in the equations is described in detail in the OPQ32 Technical Manual (SHL, 2006) and the Big Five supplement to that (Bartram & Brown, 2005). Detailed interpretations of the individual OPQ32 scales can be found in the OPQ32 User Manual (2005). It should be noted that for OPQ32 Big Five, Neuroticism is reverse scored and labeled Emotional Stability.

Around 50% of the variance in OPQ32 primary scale scores is typically explained by the five personality factors. Some OPQ scales are not related strongly to the Big Five, indicating that the OPQ measures a broader domain. Factor analyses of OPQ suggest that the Five Factor Model is embedded in a larger factor solution (typically six or seven factors), including, for example, aspects of motivation (need for achievement and need for power and control) that are not generally included within the definitions of the Big Five.

SHL adopted definitions of the Big Five close to those defined by Costa and McCrae (1992) as used in the NEO PI-R. Scores on the five factors are derived from OPQ scales that have consistently shown strong relations to the underlying factors in research.

Following identification of relevant scales on the basis of content and construct overlap, confirmatory factor analysis was used to examine the fit of each of the scale composites to the reference NEO scale and also to confirm scale loadings on each of the factors. After some minor modifications to the original models, they showed a reasonable fit (all CFIs are greater than 0.95 and RMSEAs are close to 0.05). As OPQ32 was not designed to be a Big Five instrument, it is not possible to get a good overall FFM fit to the scales. For this reason factor structure was not examined country by country. Instead, invariance of the whole correlation matrix was assessed (see Bartram, 2012, for details) to establish construct equivalence. The same set of primary

Table 1. OPQ32 Scales Used to Produce Each of the Big Five Measures

Big Five Scale	OPQ Scales With Positive Loadings	OPQ Scales With Negative Loadings
Emotional Stability (Neuroticism reversed)	Relaxed, Tough-Minded, Optimistic, Socially Confident	Worrying
Extraversion	Outgoing, Socially Confident, Affiliative, Persuasive, Controlling	Emotionally Controlled
Openness to Experience	Variety Seeking, Innovative, Conceptual, Behavioural	Conventional
Agreeableness	Caring, Democratic, Trusting	Competitive, Independent-Minded
Conscientiousness	Conscientious, Detail Conscious, Vigorous, Forward Thinking, Achieving	

scales and the same scale weights were used for scale aggregation for creating the Big Five scales for all language versions of OPQ32i.

Normative Nature of the Big Five Scales

While computed from ipsative primary scale scores, the omission of six scales and the aggregation procedures result in the ipsative constraints being removed for the Big Five scale scores. The expected average correlation between five ipsative scales would be –0.25. For the present data set, it is near zero: –.063.

Further evidence that the aggregate Big Five scores are not ipsative comes from examining the distribution of their totals for each individual. If they are ipsative, the totals will be a constant (i.e., there will be no between-subject variance in the total sum of the five scales). For the present data set of over one million cases, the individual candidate sten ($M = 5.5$, $SD = 2$) score averages across the five scales ranged from 3.09 to 7.21, with an SD of 0.50 and an average of 5.343. For ipsative scales, the mean would be fixed at 5.5 and the SD would be zero.

Relationship Between OPQ32 Big Five and Other Measures of Big Five

Correlations between the OPQ32 Big Five and two other measures were obtained in two independent studies. These data were obtained from previous research on the OPQ32 and are reported in detail in the OPQ32 Technical Manual (SHL, 2006). They do not form part of the data reported in this article. The first study was carried out with university students in the United Kingdom and involved the NEO-FFI and the OPQ32. The second was online data collected from a volunteer sample that completed the full International Personality Item Pool item set (IPIP; Goldberg, 1999) scales and the OPQ32. Again, the sample was all English-speaking and from the United Kingdom. The results are summarized in Table 2, which shows moderate to good coefficients of equivalence both for the NEO (Costa & McCrae, 1992) and the IPIP scales. In both cases, OPQ32 Emotional Stability, as expected, correlates negatively with NEO and IPIP Neuroticism. For the IPIP data set, divergent correlations (i.e., off-diagonal correlations in the full matrix) with OPQ range from –.197 to 0.472. For the NEO, they range from –.293 to 0.343. All convergent correlations are larger than discriminant ones. Table 2 also provides estimates of reliability based on data from the original OPQ32i UK standardization sample of $n = 807$ (see Bartram & Brown, 2005). It is clear from this that the OPQ32 Big Five measures are highly reliable.

Table 2. Correlations of OPQ32 Big Five Scales With the Goldberg IPIP Big Five and the NEO Big Five, Together With Internal Consistency Reliabilities for the OPQ32-Based Scales (See Text for Details)

	Emotional Stability	Extraversion	Openness	Agreeableness	Conscientiousness	*N*
Goldberg IPIP	–.725	0.820	0.610	0.621	0.657	188
NEO	–.669	0.522	0.444	0.575	0.670	261
Reliability of OPQ32i-based scales	0.920	0.920	0.870	0.870	0.890	807

Note: Correlations for Emotional Stability are with Neuroticism and are therefore expected to be negative.

Study 1: Relationships Between Hofstede Dimensions and Country-Level Personality Means and *SDs*

Method

This analysis is based on OPQ32i data from 31 different countries with a mixture of different languages. The data cover a time period from 2007 to 2010. Data were obtained through online administration from people who were being assessed either for job selection or succession planning purposes or for personal development within a job. Table 3 lists the countries and languages, with their sample sizes, and shows the breakdown by gender. It can be seen that there is considerable variation in sample sizes (from a minimum of $N = 336$ to a maximum of $N = 370{,}955$). The average gender ratio was 64% male, which is consistent with SHL data on working population gender ratios, which show this to be around 60% to 65% rather than 50% as in the broader general population. Complete demographic data on age were not available, but age was known for $N = 129{,}517$ of the sample. For this subset, the average age was 37.68 ($SD = 9.74$). Ages ranged from under 18 to over 65, with 99.8% of the sample being between 18 and 64 and 66% between 25 and 44. All the data were from *in vivo* use of the instrument for occupational assessment rather than from research use.

It was possible to match up the 31 the countries with data for Hofstede's four dimensions: PDI, IDV, MAS, and AUI (taken from Hofstede, 2001, Appendix A5, pp. 500-502). It was also possible to match 21 of the countries to data on cultural tightness taken from Gelfand et al. (2011).

Results

In order to summarize the main effects of score variation in relation to culture, Big Five scores were calculated for all the people in the data set (i.e., 31 countries, $N = 1{,}002{,}907$; see Table 4). Big Five scale scores are created through weighted aggregation of 25 of the 32 scales (Bartram & Brown, 2005, p. 6). As noted above, as they use only a subset of the 32 scales, the Big Five scale scores are no longer ipsative.

The 31 countries' average Big Five scores and the *SDs* of these averages were correlated with the Hofstede country dimensions scores (Table 5). The results show strong relationships between cultural dimensions and both means and *SDs* of Big Five scores at the country level of aggregation. As expected, there are strong negative correlations with PDI and positive ones with IDV for both Extraversion and Emotional Stability, indicating that these scale means tend to be higher in countries where individual ism is high and power-distance is low. MAS and UAI also show

Table 3. Sample Sizes and Gender Balance (*N* = 1,002,907)

Country	Language	*N*	Male	Female	Percent Male
Argentina	Spanish	17,300	10,435	5,892	60.32%
Australia	English	179,769	112,631	62,433	62.65%
Belgium	French	6,010	3,341	2,427	55.59%
Belgium	Flemish	8,358	4,313	3,876	51.60%
Brazil	Portuguese	12,549	8,705	3,791	69.37%
Canada	French	1,274	643	584	50.47%
China (PRC)	Simplified Chinese	14,835	8,126	6,348	54.78%
Denmark	Danish	7,773	4,875	2,754	62.72%
Finland	Finnish	11,076	6,210	4,603	56.07%
French	France	9,075	5,149	2,761	56.74%
Germany	German	8,076	5,735	1,890	71.01%
Greece	Greek	380	225	146	59.21%
Hong Kong	Traditional Chinese	7,735	3,045	3,063	39.37%
Hungary	Hungarian	1,658	882	708	53.20%
India	English	166,823	142,363	24,440	85.34%
Indonesia	Indonesian	530	424	103	80.00%
Italy	Italian	8,605	5,574	2,787	64.78%
Japan	Japanese	343	250	88	72.89%
Malaysia	English	11,236	6,896	4,339	61.37%
Middle East	Arabic	2,037	334	26	16.40%
Netherlands	Dutch	50,104	29,099	17,548	58.08%
New Zealand	English	2,119	1,117	1,002	52.71%
Norway	Norwegian	24,367	15,651	8,348	64.23%
Poland	Polish	811	575	211	70.90%
Portugal	Portuguese	1,026	229	127	22.32%
Russia	Russian	336	184	140	54.76%
South Africa	English	9,338	5,365	3,290	57.45%
Spain	Spanish	4,770	2,855	1,798	59.85%
Sweden	Swedish	30,863	15,759	11,757	51.06%
United Kingdom	English	370,955	221,803	135,783	59.79%
United States	English	32,776	18,350	10,657	55.99%
	Minimum	336	184	26	16.40%
	Maximum	370,955	221,803	135,783	85.34%
	Total	1,002,907	641,143	323,720	63.93%

negative correlations with Emotional Stability, and MAS also correlates negatively with Extraversion and Agreeableness. A negative correlation is also seen between Agreeableness and PDI. Openness correlates positively with uncertainty avoidance and Conscientiousness positively with power-distance.

As we would expect, countries that are high on collectivism (i.e., low IDV) and high on power-distance (PDI) show reduced variation between people in Big Five scores. This is consistent with the view that cultural "tightness" will tend to reduce the amount of individual variation that is expressed. The correlations between *SD*s for the Big Five are all high, indicating the

Table 4. Average OPQ32 Big Five Sten Scores for the 31 Countries: Means and *SDs* (for Sample Sizes, See Table 3), Together With the Average *SD* Across the Five Scales (Final Column)

		Emotional Stability		Extraversion		Openness		Agreeableness		Conscientiousness		
Country	Language	*M*	*SD*	*M*	*SD*	*M*	*SD*	*M*	*SD*	*M*	*SD*	*M SD*
Argentina	Spanish	5.47	1.59	5.98	1.66	5.21	1.74	4.81	1.77	5.51	1.74	1.70
Australia	English	5.58	2.03	5.39	2.06	5.40	2.01	5.68	1.95	5.68	2.00	2.01
Belgium	French	4.87	2.00	5.53	1.89	5.59	1.91	5.67	1.88	5.73	2.06	1.95
Belgium	Flemish	5.42	2.18	5.86	2.14	5.48	2.05	6.21	1.94	5.38	2.08	2.08
Brazil	Portuguese	4.91	1.69	5.63	1.86	5.12	1.88	5.58	1.73	4.75	1.80	1.79
Canada	French	5.83	1.86	6.39	1.89	5.24	1.92	5.68	1.87	5.64	1.92	1.89
China (PRC)	Simplified Chinese	5.46	1.91	5.13	1.91	5.03	1.79	5.91	1.61	5.36	1.75	1.79
Denmark	Danish	6.38	1.82	6.37	1.98	5.60	2.01	5.76	1.86	5.51	2.06	1.95
Finland	Finnish	6.22	2.06	5.94	2.24	5.01	2.11	6.08	1.98	5.44	2.14	2.11
France	French	4.85	2.00	5.47	1.84	5.67	1.92	5.62	1.76	5.55	2.00	1.90
Germany	German	6.08	2.00	5.96	1.88	6.08	1.86	5.87	1.80	5.32	1.84	1.88
Greece	Greek	4.32	1.90	5.76	1.67	5.54	1.97	4.97	1.80	5.87	2.10	1.89
Hong Kong	Traditional Chinese	5.13	2.16	5.43	2.09	5.58	1.95	5.67	1.87	5.36	1.84	1.98
Hungary	Hungarian	5.73	2.08	5.21	1.96	4.97	2.07	5.52	1.81	5.43	2.12	2.01
India	English	5.41	1.72	5.18	1.74	5.16	1.79	4.19	1.69	5.52	1.74	1.74
Indonesia	Indonesian	4.98	1.67	5.02	1.81	5.08	1.75	5.18	1.67	4.93	1.86	1.75
Italy	Italian	5.09	2.00	5.70	1.85	5.86	1.93	5.00	1.82	5.02	1.89	1.90
Japan	Japanese	4.46	2.27	5.16	1.91	6.45	1.84	5.62	1.78	4.15	1.86	1.93
Malaysia	English	4.90	1.81	4.58	1.92	5.11	1.74	5.29	1.67	5.17	1.93	1.82
Middle East	Arabic	4.98	1.51	4.86	1.44	5.33	1.68	4.96	1.62	5.18	1.69	1.59
Netherlands	Dutch	6.01	2.05	5.84	1.98	6.08	1.93	5.73	1.84	4.49	1.89	1.94
New Zealand	English	5.61	1.98	5.20	1.95	5.27	1.98	5.91	1.88	6.18	2.06	1.97
Norway	Norwegian	6.22	1.83	6.29	1.91	5.30	1.88	6.14	1.91	5.26	1.89	1.88
Poland	Polish	4.99	2.01	4.64	2.03	5.84	1.81	4.75	1.69	5.35	2.07	1.92
Portugal	Portuguese	4.77	1.76	5.88	1.79	5.79	2.01	5.81	1.68	5.12	1.96	1.84
Russia	Russian	4.76	1.96	5.27	1.76	6.19	1.98	5.10	1.78	5.02	1.86	1.87
South Africa	English	5.28	1.89	4.93	1.94	5.45	1.89	5.06	1.82	5.60	1.97	1.90
Spain	Spanish	4.79	1.78	5.75	1.90	5.61	1.88	5.40	1.80	5.13	1.94	1.86
Sweden	Swedish	6.24	1.90	6.37	1.91	5.61	1.87	6.28	1.89	5.83	2.00	1.91
United Kingdom	English	5.38	2.04	5.52	2.00	5.66	2.03	5.79	1.94	5.61	2.05	2.01
United States	English	5.70	2.05	5.84	2.09	5.29	2.05	5.34	1.97	5.72	2.03	2.04

presence of a general factor of within-country variability. It is this general factor that varies most clearly with PDI (negative) and IDV (positive). From a practical viewpoint, while these correlations are substantial, the magnitude of the variation in *SD*s is relatively small.

This pattern of results, where effects are positively related to IDV and negatively to PDI, implies a strong negative relationship between PDI and IDV. This is supported by the fact that they correlate $r = –0.73$ in the sample of 31 countries considered here and $r = –.60$ for the set of 69 countries listed in Hofstede (2001).

In order to explore further the relationship between score variability and cultural "tightness," "tightness" scores for 21 of the current 31 countries were obtained from Gelfand et al. (2011).

Table 5. Correlations Across 31 Countries of Mean Country OPQ32 Big Five Scores and *SD*s of Country Big Five Scores With Hofstede Dimensions Country Scores

Means	PDI	IDV	MAS	UAI
Emotional Stability	–.708*	0.545*	–.443*	–.601*
Extraversion	–.647*	0.391*	–.542*	–.058
Openness	–.064	0.112	0.025	0.353*
Agreeableness	–.482*	0.305	–.366*	–.208
Conscientiousness	–.321	0.368*	–.036	–.217
*SD*s	PDI	IDV	MAS	UAI
Emotional Stability	–.269	0.445*	0.294	0.072
Extraversion	–.403*	0.452*	–.053	–.287
Openness	–.504*	0.552*	–.090	0.085
Agreeableness	–.715*	0.754*	–.179	–.157
Conscientiousness	–.424*	0.537*	–.058	0.154

*$p < .05$.

The correlation between the overall average Big Five *SD* (see final column in Table 4) was –.507 ($p < .05$), supporting the view that variability is reduced (low *SD*s) in tight cultures (high tightness score). This is further supported by examining the individual Big Five correlations with tightness. The strongest effects are with Openness *SD* ($r = –.631$) and Conscientiousness *SD* ($r = –.481$), both significant. Negative but non-significant correlations are found with Emotional Stability *SD* ($r = –.270$), Extraversion *SD* ($r = –.239$), and Agreeableness *SD* ($r = –.406$).

Tightness, while showing the expected relationship with overall score variability, does not show significant correlations with differences in mean levels of Big Five scores between countries. Correlations range from $r = –.006$ for Openness to $r = –.246$ for Extraversion. Thus, the effect of tightness as measured by Gelfand et al.'s scale does seem to focus on how variable scores are within countries rather than the mean level of those scores.

Overall, these results indicate that a lot of the variance in country mean scores can be accounted for by country-level cultural value dimensions. It also shows that the variability in individual scores within countries is related to cultural values.

Table 6 summarizes the results of regression analyses that show how much variance in the country means for each of the Big Five can be accounted for by the Hofstede dimension values. For both Emotional Stability and Extraversion, the effects are large and significant. Effects are more modest and, with the current sample size of 31, not significant for two of the other three scales.

Study 2: Relationships With Country-Level "Outcome" Measures

It is of interest to find that country-level variations in personality scores, while small in magnitude, are systematically related to independently assessed country metrics like those of Hofstede. It suggests that these variations are "real" and not just effects of instrument- or measurement-related response biases. However, the significance of such differences for country-level "performance" is another matter. In Study 2, the same data set was examined in relation to two sets of national outcome measures: The United Nations Human Development Index (UNDP, 2009) and the Global Competitiveness Index (GCI: WEF, 2010). UNDP's Human Development Index data were obtained from the 2009 Human Development Report. The overall index is based on three

Table 6. Prediction of Country OPQ32 Big Five Means Using Hofstede Dimension Scores (31 Countries)

					Beta Values			
Scale	*R*	*R* Square	*F*	*p*	PDI	IDV	MAS	UAI
Emotional Stability	0.872	0.760	20.588	< .001	–.400	0.201	–.266	–.415
Extraversion	0.799	0.638	11.464	< .001	–.676	–.088	–.457	0.211
Openness	0.393	0.154	1.188	*ns*	–.097	0.084	–.043	0.397
Agreeableness	0.557	0.310	2.917	< .05	–.432	–.020	–.272	–.037
Conscientiousness	0.411	0.169	1.319	*ns*	–.053	0.311	0.020	–.174

subindices: quality of education, life expectancy, and gross domestic product. The Education Index (EI) is based on the combined gross enrolment ratio in education from UNESCO Institute for Statistics data 2007; life expectancy is based on 2007 UN data and the GDP index is based on GDP per capita data from the World Bank. The GCI is produced by the World Economic Forum, and the 31st GCI covers the years 2010-2011. To assess stability over time, the GCI for 2009 was also examined. The GCI covers a range of variables under the heading of the "12 Pillars of Competitiveness." These include property rights, institutional corruption, judicial independence, government efficiency, quality of infrastructure, health and education, higher education and training, market efficiency, financial market sophistications, market size business sophistication, and innovation. As well as statistical data, it includes data from the WEF Executive Opinion Survey of 13,000 business leaders in 133 countries. The GCI has three main subindices covering the "basic requirements" of economies, "efficiency enhancers," and "innovation and sophistication factors."

It can be expected that the cultural attributes that relate more to the developed world than the developing world (individualism, low power-distance) may result in an association between personality factors like extraversion, emotional stability, agreeableness, openness and conscientiousness with "success" in terms of economic performance, GDP, educational standard, and life expectancy. Clearly, these are not independent metrics and the causal relationships between them are complex, however we would expect to find increased *SD*s and higher means (with Neuroticism positively scored as Emotional Stability) for the OPQ32 Big Five in more successful countries using the UN and WEF measures as criteria. However, we should note that previous work of McCrae (2002), McCrae and Terracciano (2005), and Schmitt et al. (2007) have tended to find negative correlations between country "performance" measures and country-level Conscientiousness. This seems to conflict with the results of individual-level meta-analyses using job performance as criteria, where the scale that shows most consistent validity generalizability is Conscientiousness (e.g., see Barrick & Mount, 1991; Barrick, Mount, & Judge, 2001). Despite the previous findings of negative correlations, given the work-based nature of the OPQ32 item content and the literature on the positive impact of Conscientiousness on a range of validity criteria, we would expect to find positive relationships at the country level.

Results

Correlations between each of the Big Five means and *SD*s and the various country performance indicators are shown in Table 7. GCI 2009 and 2010 are not broken down into the three subindices, as the pattern of correlations for these was the same as for the overall index. As far as means are concerned, all correlations are positive and Extraversion and Agreeableness again

Table 7. Correlations Between GCI and UN Development Report Indices and OPQ32 Big Five Country Means and *SDs* (n = 31)

Means	GCI 2009	GCI 2010	UNDI 2009: EI	UNDI 2009: LEI	UNDI 2009: GDPI
Emotional Stability	0.551*	0.547*	0.242	0.130	0.226
Extraversion	0.430*	0.369*	0.527*	0.456*	0.489*
Openness	0.170	0.191	0.292	0.228	0.394*
Agreeableness	0.693*	0.698*	0.577*	0.518*	0.528*
Conscientiousness	0.108	0.065	0.135	−.003	0.116
SDs					
Emotional Stability	0.496*	0.520*	0.409*	0.329	0.438*
Extraversion	0.578*	0.581*	0.335	0.210	0.270
Openness	0.339	0.293	0.536*	0.249	0.470*
Agreeableness	0.603*	0.559*	0.608*	0.367*	0.636*
Conscientiousness	0.283	0.251	0.580*	0.225	0.424*

Note: GCI = Global Competitiveness Index; UNDI = United Nations Development Index; EI = Education Index; LEI = Life Expectancy Index; GDPI = Gross Domestic Product Index.
*$p < 0.05$.

show strong correlations with all indices. (Note that for the OPQ32 data, correlations are with Emotional Stability and not Neuroticism. Correlations with Neuroticism would, of course, be expected to be negative.) In addition, Emotional Stability correlates with the GCI and Openness with the UN Human Development GDP index. The variability of personality scores within countries (i.e., country-level *SD*s) are positively correlated across all indices with much the same pattern as for the country-level means. This indicates that higher levels of economic performance, higher levels of educational enrollment, life expectancy, and GDP are related to there being more variability in personality within countries as well as their being higher average levels of extraversion, emotional stability, and agreeableness.

It could be argued that the relationship between personality and outcomes is actually a reflection of the effect of cultural values. This can be tested by seeing whether removing the effects of Hofstede's dimensions from prediction of GCI and UN Human Development indices leaves any variation that can be accounted for by personality data. The results of a hierarchical regression analysis are presented in Table 8 for the case of the 2010 GCI. It can be seen that one can account for 43.7% of the between-country variation in GCI with the Hofstede dimensions. However, a further 26% on top of this is accounted for by introducing the three main Big Five measures: Emotional Stability, Extraversion, and Agreeableness. A further 7% is accounted for by adding the three personality scales' *SD*s, but this addition is not significant. Similar results are obtained when other indices are examined in this way.

Openness and Conscientiousness show the lowest correlations with GCI and UN criteria. However, apart from a small negative correlation between UNDI Life Expectancy Index and Conscientiousness, all correlations are positive as expected, but in contrast with some previous research findings (see below).

Study 3: Comparisons With Other Research

It is of interest to see what convergence or divergence there might be in terms of other data sets where Big Five scales have been examined at the country level. In particular, it is of interest to know whether the findings reported for these other measures can be replicated with the particular subset of countries examined here.

Table 8. Prediction of 2010 Global Competitiveness Index From Hofstede and Big Five Means and *SD*s (31 Countries)

Model	*R*	*R* Square	Adjusted *R* Square	*R* Square Change	*F* Change	Significance of *F* Change
Hofstede scales	0.661	0.437	0.350	0.437	5.05	$p = .004$
Hofstede scales with means for Extraversion, Emotional Stability, and Agreeableness	0.835	0.697	0.605	0.260	6.59	$p = .002$

It was expected that we should find the highest level of convergence between OPQ32 and the NEO self-report data from McCrae (2002), as the OPQ32 Big Five scales were modeled on the construct definitions used for the NEO and in both cases the data are self-report—though the populations sampled are rather different and the instrument construction is different. It was expected that both the peer-report data on the NEO (McCrae & Terracciano, 2005) and the BFI data reported by Schmitt et al. (2007) would show more divergence due to both method (in the first case) and construct (in the second case) differences.

At the country level, the data from the EPQ (Lynn & Martin, 1995), Schmitt et al. (2007), McCrae (2002), and McCrae, Terracciano et al. (2005b) were compared with the present data where there was an overlap in the countries covered. McCrae (2002) summarized a large number of data sets (36 cultures) of self-report assessment using NEO-PI-R (McCrae, 2002, p. 112, Table 3). The 2005 data, on the other hand, are peer-report data and covered 51 cultures (McCrae, Terracciano et al., 2005b, p. 415, Table 2). From the present data sets, there was an overlap of 21 countries for the 2002 NEO data and 23 countries for the 2005 NEO data. Of the overlapping countries, there were 15 in common between all three data sets. In what follows these are referred to NEO2002-SR for the self-report data and NEO2005-PR for the peer report data.

Schmitt et al. (2007), as noted earlier, used a shorter instrument (the BFI) and described self-report data from 56 countries, of which 23 overlap with the present data set. The data for the present comparisons are taken from Schmitt et al. (2007, pp. 188-189, Table 5). In what follows, this data set is referred to as BFI2007-SR.

At the level of country means, Table 9 shows there are some strong correlations between the OPQ-based Big Five and the other three sets of data, with at least one of the three other data sets being significantly correlated with OPQ32 Big Five for Extraversion, Openness, and Conscientiousness. For Neuroticism, all three of the other data sets correlate significantly with the OPQ32 based Big Five scale. None of the other three data sets show correlations with OPQ32 Agreeableness at the country level. Low and even negative correlations are found for Agreeableness and Conscientiousness, for the NEO and OPQ comparisons despite the reasonable correlation between NEO and OPQ for these scales at the level of individual data. The average convergent correlation for the NEO self-report data is 0.422, while the average absolute discriminant correlation is 0.264. For the NEO peer report data the average convergent correlation is 0.356, while the average absolute discriminant correlation is 0.260.

For the BFI data, while there are significant positive correlations with OPQ for Neuroticism and Conscientiousness, correlations for the other three Big Five scales are nonsignificant and negative. The consequence of this is that the average convergent correlation for OPQ with BFI is only 0.064, while the absolute discriminant average is $r = 0.218$ (similar to that found for the NEO data sets).

Table 9. Comparisons Between NEO, BFI, and OPQ Data for Self-Report and Peer-Report

	Neuroticism	Extraversion	Openness	Agreeableness	Conscientiousness	Countries
NEO2002-SR with OPQ	0.772*	0.735*	0.268	0.138	0.197	21
NEO2005-PR with OPQ	0.564*	0.387	0.481*	–.016	–.331	23
BFI2007-SR with OPQ	0.550*	–.127	–.349	–.239	0.484*	23

Note: All comparisons are based on the current data set (i.e., the subset of countries where there are data on OPQ and on the other measure). For OPQ32, Emotional Stability scores have been reversed.
*$p < .05$.

EPQ data were taken from Lynn and Martin (1995, p. 404, Table 1). They report data for 37 countries, of which 20 overlap with the present data set. OPQ32 Emotional Stability correlated –0.80 with EPQ Neuroticism, though OPQ32 Extraversion only correlated 0.22 with EPQ Extraversion.

It is possible that the differences found between country-level and individual-level data are related to country- or culture-specific biases that differentially affect the two sets of data. Alternatively, they reflect bias in the selection of countries included, such that they do not represent a true random sample of all countries. The extent to which scales correlate with other external criteria can provide an indication of whether bias effects are present, assuming the same bias does not affect the external criteria as well as the scale measures.

Table 10 compares the correlations with Hofstede found for OPQ Big Five with those found for the two McCrae data sets and the Schmitt data. Table 11 looks at relationships with the UNDP Human Development Index and the Global Competitiveness Index (GCI) that were explored in Study 2. In the tables, the top panel provides the results for the full OPQ32 data set of 31 counties, and then each comparison is made with the relevant subset of countries shared between this data set and the other (NEO2002, NEO2005, or BFI2007).

The data in Table 10 can be summarized by correlating the two sets of 20 correlations for each pair of instruments, where the 20 correlations include PDI correlated with each of the Big Five, IDV correlated with each of the Big Five, and so on for MAS and UAI. These correlations between instruments will indicate the similarity or otherwise of the pattern of correlations across these 20 relationships. For NEO self-report data with OPQ32, the correlation is 0.76, indicating a high degree of similarity. It is lower for the BFI data ($r = 0.49$) and lowest for the NEO peer report data ($r = 0.36$). While there are some similarities in the patterns of results, especially for the OPQ and NEO self-report data, it is worth noting that results differ for Conscientiousness. For the Hofstede dimensions, OPQ Conscientiousness has the same pattern of correlations as the BFI and opposite to that for NEO. Most other relationships are similar and effects are generally greater for the self-report data.

Again for the HDI and GCI indices (Table 11), we see that NEO Conscientiousness is negatively correlated, while OPQ Conscientiousness, for the same country subset, is positive, as was hypothesized in Study 2 based on the findings of individual-level meta-analyses of occupational criterion-related validity studies. What is interesting is that correlations with Agreeableness are far stronger for the OPQ than for the other measures.

Conclusions and Practical Implications

The forced-choice format version of OPQ32 has been shown to be very robust in terms of construct equivalence across countries and languages; between countries in relation to different

Table 10. Correlations With Hofstede Culture Dimensions for the Subsets of Countries in the Current Data Set Where Information From Other Studies Was Available

	Neuroticism	Extraversion	Openness	Agreeableness	Conscientiousness	Countries
OPQ32 countries						31
PDI	0.708*	–.647*	–.064	–.482*	–.321	
IDV	–.545*	0.391*	0.112	0.305	0.368*	
MAS	0.443*	–.542*	0.025	–.366*	–.036	
UAI	0.601*	–.058	0.353	–.208	–.217	
NEO2002 subset						*19*
NEO2002-SR with						
PDI	0.536*	–.656*	–.261	0.058	0.430	
IDV	–.287	0.656*	0.440	–.097	–.282	
MAS	0.609*	–.440*	0.305	–.469*	0.114	
UAI	0.666*	–.072	0.394	–.547*	–.279	
OPQ32 with						
PDI	0.758*	–.839*	–.135	–.453	–.224	
IDV	–.590*	0.635*	0.087	0.217	0.313	
MAS	0.433	–.545*	–.012	–.406	–.226	
UAI	0.564*	–.132	0.444	–.174	–.368	
NEO2005 subset						*23*
NEO2005-PR with						
PDI	0.386	–.596*	–.562*	–.118	0.448*	
IDV	–.246	0.578*	0.488*	0.115	–.499*	
MAS	–.230	–.055	–.039	–.547*	–.061	
UAI	0.433*	0.016	–.093	0.000	0.049	
OPQ32 with						
PDI	0.673*	–.645*	–.141	–.400	–.452*	
IDV	–.533*	0.446*	0.155	0.231	0.540*	
MAS	0.163	–.328	0.220	–.085	–.222	
UAI	0.557*	0.032	0.514*	–.084	–.426*	
BFI2007 subset						*23*
BFI2007-SR with						
PDI	0.081	–.320	–.145	0.014	–.082	
IDV	–.152	0.125	0.288	0.102	0.168	
MAS	0.486*	–.023	–.482*	–.225	–.326	
UAI	0.588*	–.385*	0.374	–.085	–.093	
OPQ32 with						
PDI	0.622*	–.523*	–.189	–.431*	–.296	
IDV	–.588*	0.309	0.155	0.297	0.340	
MAS	0.303	–.321	0.302	–.281	0.029	
UAI	0.5038	0.225	0.375	–.016	–.191	

Note: The top panel of the table gives the correlations for the full OPQ32 data set, while the remaining panels provide the results for the common NEO2002, NEO2005, and BFI2007 subsets of countries. For OPQ32, Emotional Stability scores have been reversed.
*$p < .05$.

versions of English and Chinese, and for South Africa, within-country between first-language and ethnic groups (Bartram, 2012). However, as we have seen, differences do occur between and within countries in terms of average scale scores and *SD*s. The variations in means are compa-

Table 11. Correlations With GCI and HDI Indices for the Subsets of Countries in the Current Data Set Where Information From Other Studies Was Available

	Neuroticism	Extraversion	Openness	Agreeableness	Conscientiousness	Countries
OPQ32 all countries:						
OPQ32 & GCI 2010	–.547*	0.369*	0.191	0.698*	0.065	31
OPQ32 & HDI2007	–.215	0.534*	0.339	0.589*	0.086	31
NEO2002-SR subset:						
NEO2002-SR & GCI 2010	–.500*	0.196	0.089	0.389	–.400	19
OPQ32 & GCI 2010	–.533*	0.459*	0.145	0.576*	0.207	19
NEO2002-SR & HDI2007	–.071	0.407	0.292	0.090	–.464*	19
OPQ32 & HDI2007	–.272	0.755*	0.299	0.572*	–.080	19
NEO2005-PR subset:						
NEO2005-PR & GCI 2010	–.457*	0.229	0.487*	–.037	–.306	23
OPQ32 & GCI 2010	–.438*	0.233*	0.186	0.621*	0.236	23
NEO2005-PR & HDI2007	–.058	0.453*	0.364	–.065	–.381	23
OPQ32 & HDI2007	–.175	0.433*	0.419*	0.575*	0.163	23
BFI2007-SR subset:						
BFI2007-SR & GCI 2010	–.177	–.004	–.350	–.192	–.236	23
OPQ32 & GCI 2010	–.512*	0.185	0.258	0.674*	–.064	23
BFI2007-SR & HDI2007	0.242	–.050	0.012	–.263	–.189	23
OPQ32 & HDI2007	–.141	0.511*	0.413*	0.657*	0.023	23

Note: The top panel of the table gives the correlations for the full OPQ32 data set, while the remaining panels provide the results for the common NEO2002, NEO2005, and BFI2007 subsets of countries. For OPQ32, Emotional Stability scores have been reversed.
*$p < .05$.

rable in magnitude to those associated with demographics like gender, but probably smaller than the effects associations with variables like job level or managerial position (Bartram, 2009). Effect sizes are generally not of substantive significance in terms of their impact on individual profile interpretation but are of interest in that they can be used to test for effects of scalar equivalence. In addition, they may help us understand the implications of country differences in culture.

Between-country effects are comparable in size and extent to those we find for differences within countries between males and females (SHL, 2006). From the point of view of scale interpretation, most differences are small (i.e., less than 0.20 of one *SD*), but a substantial number are moderate (i.e., likely to have the effect of shifting a scale score half an *SD* or one sten scale point). Very few are large (i.e., likely to have the effects of shifting a scale score more than one sten scale point).

While the present analysis has focused on aggregation of individual data to the country level, it is recognized that countries are not homogeneous entities. Rentfrow, Gosling, and Potter

(2008) have shown how there is systematic variation in mean Big Five scores between the states within the United States. Their results indicate relationships between state-level personality and geographic indicators of crime, social capital, religiosity, political values, employment, and health.

The present research has indicated surprisingly strong associations at the country level between some Big Five scale means and *SD*s, on the one hand, and some of Hofstede's country culture dimensions scores, on the other. Variation in *SD*s was also seen to relate to Gelfand et al.'s (2011) ratings of cultural tightness of countries.

Study 2 showed that not only are these variations in personality means and *SD*s related to cultural metrics like Hofstede's dimensions, but they also have strong relationships with "hard" outcome measures like quality of educational provision, life expectancy, GDP, and general economic competitiveness. Interestingly, scales that have been identified as showing validity generalizability in individual-level data meta-analyses (e.g., Conscientiousness; Barrick & Mount, 1991; Barrick et al., 2001) show relatively little variation across countries. Indeed, the evidence from previous research (summarized in Study 3) show negative correlations between outcome measures and Conscientiousness. It is also surprising that Agreeableness, which is rarely shown to have high validity for work performance at the individual level, has such strong correlations with country-level economic performance indicators.

Study 3 has shown a degree of convergence between, in particular, the NEO-PI-R self-report data from McCrae (2002) and the OPQ32 data, with a correlation of 0.76 in the pattern of results across the various Big Five–Hofstede correlations. Lesser degrees of convergence were noted for the other data sets. Correlations with OPQ-based measures tended to be somewhat higher for the other self-report data from the BFI than from the peer report data from the NEO. Considering the Big Five scales, we see the most convergence for Neuroticism and the least for Extraversion and Openness. The results for Agreeableness tend to be strong for the OPQ32 data but weaker for the other data sets, while there are contradictory results for Conscientiousness. One difference between the data sets relates to the populations they were drawn from. Both the McCrae and Terracciano (2005) and Schmitt et al. (2007) data were from college students. The various data sets examined by McCrae (2002) were a mixture of college student samples and adults. The present data were all working adults. It is possible that these differences in sampling may account for some of the differences in scale relationships between studies (such as the differing patterns of correlations of Conscientiousness with other measures), however further research will be needed to clarify the reasons for these differences.

What is clear is that the analysis of country-level effects is made difficult by the relatively small sample sizes we are dealing with. While we may have large samples within countries, that simply reduces the standard errors of the means and of the *SD*s. It does not make it any easier to detect small effects with small samples of countries. The standard error (SE) of correlations with an $N = 31$ is around 0.20 regardless of whether the N relates to individuals or countries. The aim for future research is to increase the number of countries studied and where possible break down countries into more homogeneous cultural subgroups, while ensuring sufficient within-group sample sizes for low SEs.

Overall, the results do show some very strong relationships between country-level means and *SD*s and independently assessed country-level metrics. As such, this provides some support for the scalar equivalence of the OPQ32 scores across countries. However, such evidence is only a starting point. For example, the presence of a strong correlation with an independent measure might be found when there is scalar equivalence across most language versions of the instrument but not all. Furthermore, it is possible that there are intervening variables that are reflected in personality data that also covary with metrics like GDP or other country-level statistics. While

such complications may make it difficult to interpret country effects, there is nevertheless a difference between effects that arise from method or instrument bias that we would not expect to find reflected in other instruments, methods or measures, and effects that covary with independently assessed country-level attributes.

Implications for Practice in Using National Versus Global Norms

The practical implications of these between-country effects can be considered in relation to a client organization with staff across the world that wishes to evaluate talent using a personality assessment tool, like the OPQ32. Clients typically want to use this in selection assessment, to help in drawing applicants from one set of countries (e.g., France, Germany, United Kingdom, Sweden, and Australia) for expatriate assignments in some other set of countries (e.g., Brazil and China). Clients may also want to carry out an audit of current managers to assess developmental needs globally or put in place a succession management process for which they need to audit top talent across the world and establish a portfolio of the capabilities that such an audit would yield.

All these activities involve making comparisons between people from different countries and cultures that may be using different languages. Such cross-group comparisons are challenging because differences between groups can arise from *real* cultural or other group-related differences as the present data indicate, rather than from instrument- or method-related systematic biases. While the need to ensure construct equivalence implies a need to restrict global assessment to constructs that are globally meaningful, this does not mean that every country and culture has to give the same value to such constructs, merely that they have the same meaning. Thus, a construct like "respect for one's elders" may have a much higher value associated with it in a country with a traditional family culture than in one with a more individualistic culture. However, people in both cultures will understand what the construct is and what it means. Within the world of multinational occupational assessment, the global organization, and the global economy, it is doubtful whether there are any particularly important local constructs that have no global counterpart. The GLOBE project (House et al., 2004) distinguished between characteristics of leadership that were regarded as universally important and those that varied from country to country. However, those that varied were universally understood. They simply did not have the same universal values.

The key practice question is: Where there is construct equivalence, can we aggregate across samples for norming purposes in order to preserve the underlying raw score difference effects? Or should we use country-based norms to remove these effects? The impact is directly analogous to that associated with gender: Should we remove gender differences by using gender norming or retain these effects by using combined male and female norm groups?

The traditional approach of aggregating data within but not between countries is difficult to justify and relates more to the history of test publishers operating within countries than it does to any good psychometric principles. As the research by Rentfrow et al. (2008) has shown, we can find large geographical differences in Five Factor scale scores, correlated with outcome variables, between-regions within-countries, as well as between-countries. Indeed, country-level aggregation is likely to represent the accumulation of heterogeneous within-country regional and cultural groups.

The present results support the guidelines developed for procedures to be followed in data aggregation for the production of multi-cultural, multi-national, and multi-lingual norms (Bartram, 2008) and are consistent with a view that attributes much of the variance between countries to "real" scale effects rather than sources of instrument or method bias.

Given that we will never "prove" scalar equivalence between groups, it is recommended that both local "national" and relevant multi-national aggregations of national norms are used for international or cross-cultural assessments and that areas where these give rise to differences in

scores should be highlighted and considered by users in the light of what is known about possible sample, translation, or cultural effects.

Limitations

There are limitations to the present research that may act to reduce the overall magnitude of the relationships with the Hofstede dimensions and the other indices. Clearly, there are very large variations in country sample sizes. A lower limit of $n = 300$ was chosen to ensure that means and *SD*s were estimated with reasonably small standard errors. However, for smaller country samples, there may also be some biases in sampling that reduce the representativeness of the sample as a "country" sample. While the present sample sizes tend to be larger than those previously reported on, and the people were drawn from a diverse working adult population, future research needs to be based on better demographic data so that differences due to gender, culture, language, country, and other factors can be disentangled from each other.

In general, there is a degree of concordance in the country-level findings between different measures, especially as regards Emotional Stability (Neuroticism). However, there are effects that are difficult to explain, such as the way Conscientiousness seems to be correlated between measures at the individual level but show different patterns of results at the country level. Study 3 showed interesting similarities and differences between forced-choice (OPQ) and rating formats (NEO and BFI), self- and peer report, short (BFI) and long (OPQ and NEO) instruments, but the differences in sampling and other factors make it difficult to attribute differences in outcomes to particular measurement method variables. Future research needs to consider the use of multi-method instrumentation and multi-level analysis with a common core of countries and cultures representing the expected diversity of score ranges that we are now able to predict from this and earlier research. Future work with the OPQ32 will be able to make use of the new IRT-based normative scoring combined with forced-choice item format (though preliminary research suggests this makes very little difference in terms of between-country effects).

The present data also do not allow us to discover what is "cause" and what is "effect." Do people in economically successful countries change their personalities or the expression of them, or is economic success dependent upon personality? Answers to such questions should be forthcoming in the future as we collect more and better personality data over time. It is only recently that the Internet has made possible the routine collection in one place of such large quantities of test data together with demographic information. This makes it practical to start examining score data in terms of time periods and to relate those time periods to the currency of indicators like those published by UNDP and WEF. In the next decade or two we should be able to look at such data across time to see which variables lead and which lag in the process of national economic change and development.

Declaration of Conflicting Interests

The author(s) declared no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.

Funding

The author(s) received no financial support for the research, authorship, and/or publication of this article.

Acknowledgments

The author is Chief Psychologist with SHL Group Ltd, which provided access to all the OPQ32 data used in the studies. The author is grateful to two anonymous reviewers for their helpful comments and suggestions.

References

Baron, H. (1996). Strengths and limitations of ipsative instruments. *Journal of Occupational and Organizational Psychology*, *69*, 49-56.

Barrett, P., & Eysenck, S. B. G. (1984). The assessment of personality factors across 25 countries. *Personality and Individual Differences*, *5*, 615-632.

Barrick, M. R., & Mount, M. K. (1991). The Big Five personality dimensions and job performance: A meta-analysis. *Personnel Psychology*, *44*, 1-26.

Barrick, M. R., Mount, M. K., & Judge, T. A. (2001). Personality and performance at the beginning of the new millennium: What do we know and where do we go next? *International Journal of Selection and Assessment*, *9*, 9-30.

Bartram, D. (1996). The relationship between ipsatized and normative measures of personality. *British Journal of Occupational and Organizational Psychology*, *69*, 25-39.

Bartram, D. (2008). Global norms: Towards some guidelines for aggregating personality norms across countries. *International Journal of Testing*, *8*(4), 315-333.

Bartram, D. (2009). Leadership competencies: Differences in patterns of potential across eleven European countries as a function of gender and managerial experience. In W. H. Mobley, Y. Wang, & M. Li (Eds.), *Advances in global leadership, Volume V* (pp. 35-64). Bingley, UK: Emerald Group Publishing.

Bartram, D. (2012). Stability of OPQ32 personality constructs across languages, cultures and countries. In A. M. Ryan, F. T. L. Leong, & F. Oswald (Eds.), *Conducting multinational research projects in organizational psychology: Challenges and opportunities* (chap. 3, pp. 59-89). Washington, DC: American Psychological Association.

Bartram, D., & Brown, A. (2005). *Five Factor Model (Big Five) OPQ32 report: OPQ32 technical manual supplement.* Thames Ditton, UK: SHL Group Ltd.

Benet-Martinez, V., & John, O. P. (1998). Los cincos grandes across cultures and ethnic groups: Multitrait multimethod analysis of the Big Five in Spanish and English. *Journal of Personality and Social Psychology*, *75*, 729-750.

Brown, A., & Maydeu-Olivares, A. (2011). Item response modeling of forced-choice questionnaires. *Educational and Psychological Measurement*, *71*, 460-502.

Christiansen, N. D., Burns, G. N., & Montgomery, G. E. (2005). Reconsidering the use of forced-choice formats for applicant personality assessment. *Human Performance*, *18*, 267-307.

Costa, P. T., Jr., & McCrae, R. R. (1992). *Revised NEO Personality Inventory (NEO-PI-R) and NEO Five-Factor Inventory (NEO-FFI) professional manual*. Odessa, FL: Psychological Assessment Resources.

Diener, E., Diener, M., & Diener, C. (1995). Factors predicting the subjective well-being of nations. *Journal of Personality and Social Psychology*, *69*, 851-864.

Gelfand, M. J., Nishii, L. H., & Raver, J. L. (2006). On the nature and importance of cultural tightness-looseness. *Journal of Applied Psychology*, *91*, 1225-1244.

Gelfand, M . J. , Raver, J. L., Nishii, L., Leslie, L. M., Lun, J., Lim, B. C., . . . Yamaguchi, S. (2011). Differences between tight and loose cultures: A 33-nation study. *Science, 332*, 1100-1104.

Goldberg, L. R. (1999). A broad-bandwidth, public-domain, personality inventory measuring the lower-level facets of several five-factor models**.** In I. Mervielde, I. Deary, F. De Fruyt, & F. Ostendorf (Eds.), *Personality psychology in Europe* (vol. 7, pp. 7-28). Tilburg, the Netherlands: Tilburg University Press.

Griffith, R. L., & McDaniel, M. (2006). The nature of deception and applicant faking behavior. In R. L. Griffith & M. H. Peterson (Eds.), *A closer examination of applicant faking behavior* (pp.1-19). Greenwich, CT: Information Age Publishing.

Hofstede, G. (1980). *Culture's consequences: International differences in work-related values*. Beverley Hills, CA: SAGE.

Hofstede, G. (1994). Management scientists are human. *Management Science*, *40*, 4-14.

Hofstede, G. (2001). *Culture's consequences: Comparing values, behaviors, institutions, and organizations across nations* (2nd ed.). London: SAGE.

House, R. J., Hanges, P. J., Javidan, M., Dorfman, P. W., & Gupta, V. (Eds.). (2004). *Culture, leadership, and organizations: The GLOBE study of 62 societies.* Thousand Oaks, CA: Sage Publications.

Jackson, D. N., Wroblewski, V. R., & Ashton, M. C. (2000). The impact of faking on employment tests: Does forced-choice offer a solution? *Human Performance*, *13*(4), 371-388.

John, O. P., & Srivastrava, S. (1999). The Big Five trait taxonomy: History, measurement, and theoretical perspectives. In. L. A. Pervin & O. P. John (Eds.), *Handbook of personality: Theory and research* (2nd ed., pp. 102-138). New York: Guilford.

Kirkman, B. L., Lowe, K. B., & Gibson, C. B. (2006). A quarter century of culture's consequences: A review of empirical research incorporating Hofstede's cultural values framework. *Journal of International Business Studies*, *37*, 285-320.

Lynn, R., & Martin, T. (1995). National differences for thirty-seven nations in extraversion, neuroticism, psychoticism and economic, demographic and other correlates. *Personality and Individual Differences*, *19*, 403-406.

Martin, B. A., Bowen, C. C., & Hunt, S. T. (2002). How effective are people at faking on personality questionnaires? *Personality and Individual Differences*, *32*, 247-256.

McRae, R. R. (2001). Trait psychology and culture: Exploring intercultural comparisons. *Journal of Personality*, *69*, 819-846.

McRae, R. R. (2002). NEO-PI-R data from 36 cultures: Further intercultural comparisons. In R. R. McCrae & J. Allik (Eds.), *The five-factor model of personality across cultures* (pp. 105-125). New York: Kluwer Academic/Plenum.

McCrae, R. R., & Terracciano, A. (2007). The five-factor model and its correlates in individuals and cultures. In F. J. R. van de Vijver, D. A. van Hemert, & Y. Poortinga (Eds.), *Individuals and cultures in multi-level analysis* (pp. 247-281). Mahwah, NJ: Erlbaum.

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

McRae, R. R., Terracciano, A., & 79 Members of the Personality Profiles of Cultures Project. (2005b). Personality profiles of cultures: Aggregate personality traits. *Journal of Personality and Social Psychology*, *89*, 407-425.

Oyserman, D., Coon, H. M., & Kemmelmeier, M. (2002). Rethinking individualism and collectivism: Evaluation of theoretical assumptions and meta-analyses. *Psychological Bulletin*, *128*, 3-72.

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*(5), 339-369.

Schmitt, D. P., Allik, J., McCrae, R. R., & Benet-Martinez, V. (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.

Schmitt, D. P., Realo, A., Voracek, M., & Allik, J. (2008). Why can't a man be more like a woman? Sex differences in Big Five personality traits across 55 cultures. *Journal of Personality and Social Psychology*, *94*, 168-182.

SHL. (1999). *OPQ32 manual.* Thames Ditton, UK: SHL Group plc.

SHL. (2006). *OPQ32 technical manual.* Thames Ditton, UK: SHL Group Ltd.

SHL. (2009). *OPQ32 development and psychometric properties of OPQ32r (supplement to the OPQ32 technical manual).* Thames Ditton, UK: SHL Group Ltd.

Steel, P., & Ones, D. S. (2002). Personality and happiness: A national level analysis. *Journal of Personality and Social Psychology*, *83*, 767-781.

Taras, V., Kirkman, B. L., & Steel, P. (2010). Examining the impact of culture's consequences: A three-decade, multilevel, meta-analytic review of Hofstede's cultural value dimensions. *Journal of Applied Psychology*, *95*, 405-439.

Trompenaars, F. (1993). *Riding the waves of culture: Understanding diversity in global business.* Chicago: Irwin Professional.

United Nations Development Programme. (2009). *Human development report 2009. Overcoming barriers: Human mobility and development. Statistical Annex*. New York: Author.

Van de Vijver, F. J. R., & Leung, K. (1997). Methods and data analysis of comparative research. In J. W. Berry, Y. H. Poortinga, & J. Pandey (Eds.), *Handbook of cross-cultural psychology: Vol. 1: Theory and method* (pp. 257-300). Boston: Allyn & Bacon.

World Economic Forum. (2010). *The global competitiveness report 2010-2011*. Geneva, Switzerland: Author.