International Journal of Testing, 8: 315–333, 2008
Copyright © Taylor & Francis Group, LLC
ISSN: 1530-5058 print / 1532-7574 online
DOI: 10.1080/15305050802435037

Routledge
Taylor & Francis Group

Global Norms: Towards Some Guidelines for Aggregating Personality Norms Across Countries

Dave Bartram
SHL Group Ltd.

The article discusses issues relating to the international use of personality inventories, especially those in which organizations make comparisons between people from differing cultures or countries or those with different languages. The focus is on the issue of norming and the use of national versus multinational norms. It is noted that questions related to norming are in turn dependent on a range of factors that relate to how successfully an inventory has been adapted. However, good adaptation is a necessary but not sufficient condition for determining how various different approaches to norming might operate. In practice, data from the OPQ32i indicates that between-country and between-language differences in average scale scores are generally small and less than those often associated with other factors such as gender. In the light of the data reported here and other work, some guidelines on aggregation are proposed.

Keywords: global norms, international norms, personality assessment.

Increasing numbers of organisations are using personality assessment in international contexts. This raises a need to compare the results of people who have completed an instrument in different countries or using different language versions. The question frequently arises whether the results from two candidates for the same position who have completed different-language versions of the same instrument should be compared using a common (i.e., multi-lingual) norm or their own local language norms. Similar questions can arise in relation to possible country differences and culture differences in which there is no language translation issue.

Correspondence should be addressed to Dave Bartram, SHL Group Ltd., The Pavilion, 1 Atwell Place, Thames Ditton, Surrey, KT7 0NE, UK. E-mail: dave.bartram@shlgroup.com

Norms are part of a measurement procedure: they provide the scale that is needed to assign a value and meaning to the outcomes obtained by some operational technique or instrument. Although test responses reflect individual characteristics, they also depend on a number of other variables. Roe (2006, personal communication) has characterized these as:

A. Endogenous factors (biological characteristics such as gender, age, race)
B. Exogenous factors (environmental characteristics such as educational level and type, job level and type, organization, industrial sector, labor market, language, culture)
C. Examination factors, including format factors (paper and pencil vs. computer), setting, and stakes (e.g., pre-screening, selection, development, research)
D. Temporal factors (e.g., generation)

Any norm group reflects a particular profile of these factors; for example, Caucasian males aged between 30 and 50 (endogenous factors) with graduate-level education working in the banking and finance sectors (exogenous factors) who were tested using a computer-based version of the test as part of a job screening procedure (examination factors) during the period 1990 to 2005 (temporal factor).

Norms serve two main purposes. First, they provide a way of standardizing scores obtained from different instruments in the same contexts (e.g., two different ability tests). Second, they provide a frame of reference with links to relevant endogenous, exogenous, examination, and/or temporal factors. Norms are used to serve the needs of policymakers, test users, and test takers. For policymakers, they provide a basis for making comparisons between the performances of groups and can reveal or conceal effects of intervention (e.g., consider the whole issue of race-norming in the United States). Test users need norms because they have to produce correct assessments, predictions, recommendations, decisions, training, or career interventions. Finally, test takers need norms because they need to be able to understand their scores through reference to the performance of others.

DIFFERENT TYPES OF NORMS

In practice, we can differentiate three main types of norm in terms of sampling.

The Standardization Norm

The standardization norm is produced by careful sampling from a well-defined population. This type of norm requires a large sample (2,000 or more) and careful control of the sampling. Weighting can be used to balance samples up to match proportions in the population (e.g., if only 45% of the sample are male but 49% of the population are male, the weight given to males in the sample can be increased to rebalance this). In any standardization norm, the sampling needs to be well defined in terms of the ABCD factors and proportional to the numbers in the population defined by the particular profile of factors.

The Occupational Personality Questionnaire (OPQ32: SHL, 1999; 2006), for example, was standardized on a UK general population sample. However, there are very few other examples of instruments within the occupational testing field that have been standardized in the above fashion. What is far more common are user norms.

The User Norm

As the name implies, a user norm is created from the data provided by users of the instrument. The important difference here is that assignment of the ABCD profile to the data is post hoc. We look at the various samples of data that have been collected and then characterize them in terms of the ABCD variables to describe the population from which they come. Unlike the standardization sample, however, they have not been sampled from a population that was defined a priori.

Two test providers might each claim, for example, to have a norm for graduate applicants to the pharmaceutical industries. However, as each provider has different clients, some sampling biases may occur such that neither is truly representative of applicants to companies in this industry sector. In practice, the main requirement here is to sample sufficient different industries within the sector to be sure that the final norm is reasonably likely to be representative of this group for this industry. These norms need to draw subsamples from a sufficiently diverse range of exemplars of the various sources of people that make up the population of interest (e.g., different pharmaceutical companies).

The Benchmark

A benchmark describes the performance of some specific reference group, such as top performers in finance roles in a particular organization or set of organizations. These are typically defined in terms of a mean and standard deviation and may relate to quite small samples; i.e., less than those required to define a norm group. Benchmarks are useful for seeing whether people are at, above, or below some predefined point on a scale. In personality instruments, an important aspect of benchmarking will be the shapes of a profile and associated fit measures.

SOME BASIC ISSUES

When do effects of culture matter? This involves two related issues. The first regards how we define culture and how we aggregate data across cultures. The second is one of effect sizes: When does size matter? In the present article I focus on the issue of using various levels of aggregation of data for norming purposes in occupational assessment.

When measuring personality for occupational assessment (e.g., as part of a job selection procedure), organizations will often use various adaptations of an instrument to assess people from various countries and will then make direct comparisons between these people. In so doing, we face many questions: Are these measures of common constructs, or are we comparing chalk and cheese? If they are common constructs, are there impacts of culture on general score levels? Is so, how should these be taken into account?

When talking about culture, we need to consider and be able to disentangle effects of language, nationality, and local culture. We cannot begin to answer the "Does it matter?" question without addressing the question of "What is culture?" In the literature on culture people seem to confuse culture, nationality, and language. Definitions variously talk about:

- Shared values (i.e., a common understanding of what is more important in life and what is less so)
- Shared cognitions (i.e., a common way of looking at and making sense of the world and relationships with people)
- Shared knowledge (i.e., a shared understanding of what constitutes common sense, what we can assume other people know, and what we can take for granted)
- Shared standards or cultural norms (e.g., how one behaves in social setting, dress codes, how one expresses emotions in public). One could perhaps call this shared social skills.
- Shared language

In practical terms, culture matters when it is related to a group of people for whom within-group variability in terms of relevant constructs is relatively small compared with variability between them and other groups. How clear the boundary is between in- and out-groups is another issue. Common history and common geography may be important factors in defining cultural boundaries and are likely to be necessary but not sufficient for the development of shared values, knowledge, and understanding, and for the insulation of a group from the diluting influences of other cultures.

More problematic is the whole notion of national culture and the identification in testing of norms with nationally defined standardization samples. Increasingly, nations are culturally heterogeneous, politically defined geographical areas. The tendency to confuse culture and nation permeates much of the literature. From Hofstede (1980; 2003) through to the GLOBE project (House et al, 2004) there is a focus on countries as the unit of analysis. While it is very convenient to aggregate data by country and then to call differences between countries "national cultural" effects, it is less clear that these differences are necessarily cultural. It is still standard practice to norm tests using national samples, generally with some acknowledgement to ethnic mix demographics but often with no analysis of the size of effects associated with cultural demographics. This probably has more to do with publishers needing to provide norms that relate to their national market (often defined in terms of the language the test is in) than it does to any notion that nations are mono-cultural.

Countries may vary in the degree to which they are mono- or multi-cultural and the degree to which differing cultural groups have retained their distinctiveness over time or have been blended. Countries will also differ in terms of the mix of languages that they contain. From an assessment point of view, this can be a major issue in which direct comparisons are needed to draw between people within a country from different cultures. This may be formally recognized, as in Canada or Belgium, or informally so, as in many other countries.

Aggregation of country-level data into country clusters may exacerbate the problem by appearing to identify multi-national or even global cultures while concealing a high level of within-country cultural diversity. It may be more sensible to apply multi-level analysis approaches to data and to define levels of aggregation in terms of demographic variables that are related to substantive differences in scores.

The unit of analysis and the level of aggregation of data should not be defined in terms of some arbitrary political construct (like a nation) unless it can be shown that this corresponds with a single culture or homogeneous group. Definition of the unit of analysis should be tied to the operational definition of culture and the basic notion of relative homogeneity within and heterogeneity between groups. Culture only matters for assessment purposes when it is related to some effect or impact on scores that is a group-level effect and that is large enough to result in misinterpretation of individual-level scores.

Staying with the definition issue, should we distinguish more clearly between notions of culture as shared values, shared norms of behavior, shared perceptions, and shared knowledge, or is culture a complex of all of these? It is certainly important to make these distinctions when considering organizational cultures and looking at research on culture change, and it may well be that variations between groups in some aspects of culture result in far grater cultural distance than others.

When considering the impact of culture on assessment and the issue of instrument adaptation, various effects of culture have very different implications. For example, within limits:

- differences in norms of behavior are not likely to create problems of construct non-equivalence but may well result in shifts in manifest levels of trait expression, even if underlying latent trait levels are not different;
- differences in values are also unlikely to create problems, although they should be detectable through the use of relevant instruments, assuming that there is a shared common set of values (Schwartz, 1992); and
- differences in perception and ways of looking at and understanding the world are the issues that are more likely to have an impact on equivalence of assessment across cultures and may ultimately result in there simply being no construct equivalence.

Methodologically it might help to define cultural difference or cultural distance in ways that would relate to the issues involved in the adaptation of measurement instruments (Hambleton, 2005).

COMPARING SCORES OF PEOPLE FROM DIFFERENT COUNTRIES OR CULTURES

Differences between the personality scale scores for people from different countries (whether using the same or different languages) can arise for a number of reasons (apart from real differences in their personalities). The use of country-specific user norms when assessing multi-national samples can have three potential impacts.

1. It has the negative impact of causing bias if the country norms are not comparable in terms of the demographics of the samples they contain.
2. It removes (i.e., controls for) possible effects of cultural differences in the expression of traits in which there are real differences between countries. This may be deemed as desirable in some cases and not in others. It is analogous to using gender-specific norms to remove any effects of gender in raw score differences or separate black applicant and white applicant group norms to remove ethnic group differences. For some applications this may be reasonable, for others it may not.
3. It has the positive impact of equating for differences arising from bias because of translation effects; i.e. in effect, removing apparent scale differences that have arisen from lack of equivalence in translations.

Each of these is considered in more detail below.

Demographic Differences between National Norm Groups

The particular mix of demographics in the samples used to make up a country norm is likely to differ from country to country. Where those demographics have an impact on scores, we will find sample bias effect. Consider first, two people: one in Country A and one in Country B. Both are English-speaking and have similar cultural backgrounds. Both people get the same raw score on a scale. However, the norm group for Country A contains people drawn from different populations to the samples used to provide the norm group for Country B. For example, Country A's norm may be based on samples drawn predominantly from management in the finance and banking sectors, while Country B's norm may be graduates, with a majority being applicants for management trainee positions in the retail or manufacturing sectors. In this case, by using country-specific norms, these identical raw scores are reported as different standard scores.

Implication: People from different countries who have the same raw scores may appear to be different when these scores are standardized using their country norms with different mixes of demographics. In this case, the designation of such norms as national is misleading. Where full measurement equivalence can be demonstrated, it would be better to combine these into a single aggregate norm group and to use the same multinational group for all the people. Alternatively, one might use just one of the countries' norms for everyone (whichever was the more relevant in terms of industry sector and population).

If this was the only consideration, we could combine norms across countries at will, focusing on the user demographics rather than the incidental issues of which country they come from or which language was used. However, there are two other considerations that argue against doing this.

Cultural Bias

Personality inventories are designed to measure the enduring traits that underlie behavior. The way in which a trait is expressed in behavior is partly a function of the level of that trait and partly a function of the setting in which the behavior occurs. For example, in the Netherlands it is expected that people will be outspoken, while in the United Kingdom the norm is for diplomacy.

Some settings have universal impacts on expression; for example, even extraverts stop being talkative when they are listening to a performance in a theater. Others are culturally related; for example, cultural norms for extravert behaviors are very different in Latin and Scandinavian cultures. Thus, a person with a given latent or source trait level of extraversion will express that more in some cultures than in others. As the OPQ32 and related instruments use ratings of descriptions of behaviors to infer traits, people brought up in Latin cultures may appear to be more extraverted than those brought up in Scandinavia. By norming each person's raw scores using their own country norm, we can equate across cultures for these effects.

Implication: If we used a common (aggregated) norm group, two people with different raw scores would appear to have different levels of the trait. However, a raw score of 15 on the Outgoing scale may indicate the same latent trait level in Culture A as a raw score of 20 indicates in Culture B. Only by having culture-related norms can we remove that difference, giving each person the same standard score.

A further consideration here is that countries are not mono-cultural. Subcultural groups within countries would also need to be considered if we wanted to control properly for cultural bias (e.g., Hong Kong Chinese in Canada; Polish immigrants in the United Kingdom or United States; Pakistani cultures in the United kingdom; and so on).

For the OPQ32, the most widely used and translated version is the ipsative version of the instrument. It is known that there are cultural differences in social desirability responding and that these differences can bias scores on personality scales that use Likert ratings. Corrections for these effects have been suggested (Hanges, 2004). While Hanges argued against the use of ipsatization or ipsative item formats as a method of controlling for response bias in cross-cultural comparisons, it is acknowledged that the constraints imposed by the ipsative measurement approach will reduce any systematic differences due to socially desirable responding associated with different countries or cultures. Forced-choice format instruments, like the OPQ32i, are therefore likely to show smaller raw score differences between cultures than instruments that use Likert scale format items, like OPQ32n.

The main argument raised against the use of ipsative measures for between-country comparisons is the same as that used for arguing against their use for between-person comparisons—the constraints imposed on the degrees of freedom for variance between scales. However, it is now well established (e.g., Baron, 1996; Bartram 1996) that these constraints have minimal impact when the number of scales is large (more than 20 or so) and the average scale intercorrelation is low (as for the OPQ32). Furthermore, Brown and Bartram (2008) have shown how latent normative scale scores can be recovered from ipsative data when the number of scale is sufficiently large, and that the recovered normative scores are not constrained in the same way as the ipsative data from which they derive.

Any process for adjusting or correcting scores (as suggested for reducing cultural bias effects on Likert-rating scale scores) raises a number of other issues. Ultimately, it may be impossible (whether using ipsative methodologies or normative scale corrections) to prove that differences between scores for two cultural groups represent real differences in terms of average amounts of a latent trait. The main advantage of ipsative methods is that they avoid the need to make post hoc changes to people's scores while, in the case of OPQ32i and other instruments with large numbers of scales, not constraining their freedom to display score differences.

Language and Translation Bias

The third issue relates to translation and language issues. A multi-lingual norm only has meaning when candidates have responded to equivalent items, whatever the language of presentation of the questionnaire. Without the degree of consistency between versions there could be no question of comparing raw scores from different versions against the same norm.

When translating a test the objective is to preserve trans-linguistic and trans-cultural meaning (e.g., Daouk, Rust, & McDowall, 2005), but this may not always be possible. Some items may sound stronger or weaker in the target language and hence result in a shift in the relationship between raw scores and latent trait level. Hambleton and Patsula (1999) pointed out that some constructs have very different meanings across cultures. This makes it very difficult or impossible to create a translation with an equivalent meaning.

One of the major motivations for the OPQ32 development was the need to create a questionnaire that is applicable in a wide range of countries and cultures. Special attention was being paid to the translatability of the scale constructs and the items during the development process (SHL, 2006). The localization process for each country was also clearly defined and was centrally controlled and supported.

AGGREGATION OF NORM GROUPS ACROSS COUNTRIES

Norm groups can be produced either with the restriction that samples are only aggregated within countries, or we can also allow aggregation across countries. Where groups are aggregated across countries, the norm group formed can be called a multi-national norm. Where groups are aggregated across languages, we can speak of a multi-lingual norm, and where the aggregation is across cultures it would be a multi-cultural norm. In practice, most national norm groups will be multi-cultural because they represent the local mix of cultures.

Multi-lingual aggregations within countries could be envisaged, for example, in multi-lingual countries like Belgium or Canada, or within countries where there is more than one main applicant language for job recruitment (e.g., Spanish and English in the United States). Single language multi-national norms would be useful for making comparisons between people from different countries, all of whom share a common language.

The use of aggregated norms for different countries sharing a common language has the potential benefits of not concealing effects of country-related cultural differences (which are hidden by using country-specific norm groups) and of reducing the impact of country-related sample biases. It also has the potential negative effect, in which there are real differences in how the common language is used, of treating language–related bias effects as real country differences in average trait levels.

In practice, for international assessment projects in which people from different countries or people who have used different language versions are concerned, it is recommended that if aggregated norms are used then they should be used in combination with national norms. Comparison of the local country norm with an aggregated norm will show where the average profiles diverge and, hence, where one will find differences in individual scores. Areas in which main differences occur can then be highlighted and considered in the light of what is known about possible sample, cultural, or translation effects.

Clearly, caution needs to be exercised where there are differences between average raw scores for country norms, where such differences relate to different language versions of OPQ32, and where such differences cannot be explained by differences in demographic mix (i.e., they are not due to sampling bias) or by cultural factors.

Country Differences

Earlier, the question was posed: "When do effects of culture matter?" This section summarizes information about OPQ32 score distributions for a range of different countries to see whether the use of different local national norms or aggregate norms would actually result in the reporting of different standard scores. More details of these data can be found in SHL (2006). The data look at country differences across 12 Western European countries, the United States, Australia, and South Africa for a total of 61,438 working adults who completed OPQ32i.

English Versions of the OPQ32i

Some of the countries examined used UK and US English versions of the OPQ32. These include the United Kingdom, Australia, South Africa, and the United States. The British sample contained 7,784 applicant and employee subsamples and the American data set 1,414. The South African data set consisted of 6,057 applicants and employees from various industry sectors, including insurance, finance, and safety. A total of 64.8% were male and 35.2% were female. The ages ranged from 16 to 67, with a mean age of 34.07 years (SD = 10.0). The ethnic composition included 52.2% black and 47.8% white respondents. The Australian data set included 5,292 applicants and employees from various industry sectors, predominantly banking and finance (55%), manufacturing, mining and production (8%), public sector (5%), retail (3%), professional services (2%), and 27% other. A total of 69.1% were male, 30.9% were female. Around 50% of the respondents were graduates and 50% were managerial and professional.

Translated Versions of the OPQ32i

We also examined 11 European countries that used translated versions of the OPQ32i. Country sample sizes ranged from 861 to 8,222, with an average of 3,768. Respondents completed local ipsative language versions of the OPQ32i online for selection or assessment purposes in real high-stakes situations.

ANALYSES

Further details of the various samples and demographics can be found in SHL (2006), as can tables of the various country specific means and standard deviations for the 32 scales.

The main finding of structural equation modeling analyses comparing the OPQ32 scale structure across different English-speaking samples and different European language versions was that the pattern of correlations between the 32 scales was invariant across different language versions. However, similarity of structure does not imply equivalence of means or of standard deviations between countries. The structural invariance simply indicates that the constructs being measured have the same relationships with each other from country to country.

Consistent and clearly interpretable differences between males and females were found on some OPQ32 scales. Therefore, for country comparisons it is essential to control for gender effects. Other sampling variables, such as job level or industry, could potentially cause bias in the data. In the present research, however, these variables were not considered due to the lack of information on the demographic mix in the various samples.

Univariate analyses of variance showed highly significant results for country, gender, and their interaction for most of the OPQ32 scales. Only one scale (OPQ32 Trusting) yielded no significant differences between males and females; the country effect, however, was significant. This result is not surprising, considering the very large sample sizes which entail that almost any score difference will be statistically significant even if it is of no practical consequence.

The magnitude of differences were analyzed for each scale in relation to the scale's standard deviation, adopting Cohen's classification for the magnitude of the standardized differences (Cohen, 1988). This considers effects of 0.8 to be large, effects of 0.5 to be moderate, and those of 0.2 to be small.

Comparison of English-speaking Countries

This comparison included the United Kingdom, the United States, Australia, and South Africa (N = 20,122 people). The United Kingdom, Australia, and South Africa use the UK English version of the OPQ32i. The US English version is used in the United States. There was very limited variation in mean country scores for both males and females. Mean scale score deviations from the overall average were typically within 0.2 of a standard deviation, which corresponds with a small effect size in Cohen's classification. No scale means deviated from the overall mean by more than 0.5 SD.

Comparison of European Countries

This comparison included 12 European countries: Belgium, Denmark, Finland, France, Germany, Italy, the Netherlands, Norway, Portugal, Spain, Sweden, and the United Kingdom (N = 49,091 people). Each country used their local version of the OPQ32i: French and Flemish for Belgium, Danish, Finnish, French,German, Italian, Dutch, Norwegian, Portuguese, Spanish, Swedish, and UK English versions, respectively. There was a greater variation in mean country scores than was observed for English-speaking countries. Mean scale score deviations from the overall average were typically within 0.3 of a standard deviation. No scale means deviated from the overall mean by more than 0.8 SD.

Country Differences in Terms of the Big Five Personality Factors

Table 1 shows average sten scores (the OPQ32 using the sten scale for reporting results, where one sten is equal to half an SD) by country for the Big Five personality factors derived from the OPQ32i scores.

The Scandinavian male respondents were most Emotionally Stable, with the Danish males scoring the highest at 6.4 stens. The Portuguese and the French females scored low (4.2). On Extraversion, country means ranged within the medium effect size band, and no obvious outliers existed. The same is true for Openness to Experience.

While females scored consistently higher than males on Agreeableness, particularly high scores were observed for the Norwegian and the Swedish females (6.5). The American and South African men showed low scores on this factor (4.3 and 4.2 stens, respectively). Females also scored consistently higher than males on Big Five Conscientiousness. The Dutch men were least Conscientious (4.1),

TABLE 1
OPQ32i Based Big Five Mean Sten Scores by Country (From SHL, 2006)

	Male					Female				
	Emotional Stability	Extraversion	Openness	Agreeable ness	Conscientious ness	Emotional Stability	Extraversion	Openness	Agreeable ness	Conscientious ness
Belgium (Dutch)	5.59	5.33	5.26	5.32	5.13	5.28	6.17	5.28	6.25	5.94
Belgium (French)	5.21	5.39	5.48	5.27	5.39	4.84	5.52	5.57	5.94	6.41
Denmark	6.57	5.96	5.49	5.23	5.29	6.17	6.03	5.57	6.09	5.85
Finland	6.44	5.66	4.61	5.53	4.71	6.23	5.83	4.82	6.47	5.38
France	4.90	5.31	5.42	5.05	4.76	4.22	5.06	5.69	5.36	5.60
Germany	6.21	5.41	5.97	5.24	5.03	5.74	5.83	6.09	5.65	5.64
Italy	5.08	5.13	5.71	4.82	4.66	4.55	5.15	5.92	5.15	5.11
Netherlands	6.15	5.46	5.96	5.08	4.12	5.45	5.50	6.10	5.98	4.68
Norway	6.44	5.78	5.07	5.63	5.02	5.90	5.78	5.09	6.52	5.51
Portugal	4.72	5.35	5.39	4.74	5.32	4.19	5.18	5.45	5.38	6.02
Spain	5.55	5.31	5.81	5.68	5.43	5.02	5.56	5.86	5.96	6.58
Sweden	6.40	5.96	5.41	5.79	5.37	5.96	6.09	5.66	6.53	5.95
United Kingdom	5.56	5.39	5.63	5.00	5.33	5.10	5.64	5.65	5.97	6.15
United States	5.87	5.21	5.16	4.32	5.70	5.41	5.71	5.38	5.88	6.68
South Africa	5.30	4.62	5.38	4.25	5.19	5.13	4.89	5.33	4.86	5.79
Australia	5.66	5.22	5.33	5.17	5.72	5.17	5.57	5.46	5.90	6.51

while the Spanish, the American, and the Australian women scored high at 6.6, 6.7, and 6.5, respectively.

Country Clusters

It is of interest to consider whether there is any pattern in the differences and similarities between countries. In order to do this, cluster analysis was carried out on the standardized mean country OPQ32i scores for both males and females. Squared Euclidean distance was used as the measure of distance between the average profiles and Ward's method was used for recalculating distance to new clusters. The analysis revealed very similar groupings for males and females.

Three clusters were consistent for males and females:

1. The four Scandinavian countries (Finland, Denmark, Sweden, and Norway) plus Germany and the Netherlands;
2. Belgium (French-speaking), France, Italy, and Portugal;
3. The United Kingdom, the United States, Australia, and Spain.

The South African females joined the second cluster and the males joined the third. The Flemish-speaking males from Belgium were in the third cluster while the females were in the first, close to the Dutch women. Performing the cluster analysis with the genders together, males and females from the same country were grouped closely together for the following countries: France, Belgium (French-speaking), South Africa, Italy, Portugal, Germany, and Finland. For other countries, rather closer clusters were formed for the same gender on geographical or cultural basis. For example, the British, American, and Australian men were grouped together as were the females from these countries. Similar effect was observed with men and women from Sweden, Norway, the Netherlands, and Denmark. They seemed to be more similar to the same gender from culturally similar countries than they were to the other gender from their own country.

These results provide further support to the conclusion that most of the differences described in these analyses are in fact culture-related rather than attributable to any language or translation bias. For example, the Finnish language is very different to other Scandinavian languages; however, the average profile is very similar to the other Scandinavian countries.

RELATING GENDER DIFFERENCES TO CULTURAL FACTORS

Since the OPQ32i analyses reported in SHL (2006), the pool of OPQ32i data has been increased to over 74,000 people from 19 different countries using 14 different

TABLE 2
Size of Gender Difference (d-value: Male minus Female) Correlated with Each of Four Hofstede Dimensions for N = 19 Countries

	Power Distance	Individualism	Masculinity	Uncertainty Avoidance
Emotional Stability	–0.47*	0.55*	–0.39	0.31
Extraversion	0.06	–0.22	–0.48*	0.18
Openness	–0.01	–0.27	0.16	–0.08
Agreeableness	0.59*	–0.51*	0.19	0.22
Conscientiousness	0.11	–0.54*	–0.12	–0.56*

*$p < 0.05$.

languages (now including Chinese). These data have been examined for functional (or construct), measurement unit and scalar equivalence between countries. The research provides strong support for construct equivalence across countries. As in the earlier work, scale means show effects of country, gender, and interactions between country and gender. Gender differences were largely consistent with the findings of McCrae & Terracciano (2005) and Costa, Terracciano, and McCrae (2001). They showed systematic relationships with Hofstede's (1980) cultural dimensions, with larger gender differences (male minus female) on the Big Five being found in countries with low power distance, high individualism, and high uncertainty avoidance (see Table 2).

While the overall magnitude of scale differences between countries is small, for any particular language or group there may be one or two areas of the profile which differ to a greater degree; these are the scales in which the choice of norms will make a difference. It is important that the interpreter is aware of the scales for which differences can occur and considers carefully how important or relevant they are for the particular assessment that is being undertaken.

CONCLUSIONS

The results are perhaps surprising in showing relatively small differences in average scores across languages and countries. While the present data focus on countries as the unit of analysis, they provide a means of comparing differences between countries where language is the same (or similar) and where it differs.

Very few scales show differences that would result in moving a person's score more than one sten point (i.e., half an SD) if different norms were used. For most cases, the reported sten scores would be unchanged. The differences found between countries and languages are far smaller than those found between males and females. The gender differences are, however, moderated by country and language differences.

Aggregation of Norm Groups across Countries and Languages

Returning to the practical question of what to do about multi-national or multicultural norming in applied settings, the key question to answer in choosing a norm reference groups was clearly stated by Cronbach (1990): "Does the norm group consist of the sort of persons with whom [the candidate] should be compared?" (p. 127). Cronbach makes clear that this does not even entail comparing people with others from their own demographic group. Some tests, for example, provide scoring against same-sex and opposite-sex norms to show how a person would appear in comparison with people of the opposite sex as well as of the same sex.

In answering Cronbach's question, the test user also needs to consider whether the focus should be on a broad or a narrow comparison. Cronbach cites the example of a person who complains to his doctor of only getting four hours sleep a night. The doctor at first considers this abnormal in comparison with the patient's age group. However, when he learns that this person is a hard-drinking male who is threatened with divorce and unemployment, he finds that the sleep pattern is no longer abnormal. Compared with a narrow norm group the person is average, compared with a broader one he is not.

Narrow groups can be considered as selected samples from broad groups (earlier such narrow groups were described as benchmarks rather than norms). Conversely, one might consider broad norm groups as aggregations of narrower ones. However, there is an important caveat here: a broad norm can only be constructed by aggregating narrow selected samples that cover the demographic spread of the target broader group. If one were to draw, for example, one sample of young white males and another of older black females from a data set of British employees, one could not then "reconstruct" a norm covering gender, age, and ethnicity by combining these samples; some judgment on the adequacy and meaning of the mix is needed. Just as we can aggregate across ages, gender, and other demographics within countries, we can also do this across countries, languages, and cultures, subject to the caveats discussed earlier.

In aggregating norms, the following points should be noted.

- The aggregation of norm groups across countries, language, or cultures should not be automated but based on judgment.
- The correlation matrices for the countries or languages concerned must not be different.
- The norm groups should be checked to ensure that they are comparable in terms of sample demographics. If they are not comparable, adjustment by re-weighting may be considered.
- The mixtures of countries should be reasonable. Increasing caution should be exercised when mixing data from countries that are more divergent in terms of language, geography, or culture. As a guide, countries in the same clusters

(as evidenced by cluster analysis of mean raw scores) may be combined with less concern than those from different clusters or those with very different labour markets.

- Where demographic factors are associated with score differences, the demographics mix within each sample should be weighted to ensure comparability of mix across samples. This may be carried in relation to: population statistics, known statistics on specific sub-sectors or industries, or equal weights.

  While weighting is a matter of judgment and it is not possible to prescribe what the correct method should be, whatever weighting has been used should be made clear.
- Prior to combining them, country, language, or culture samples need to be weighted to make them comparable in terms of overall weight in the final mix. For example, if the norm for Country A has 500 people and that for Country B has 5,000, combining the two sets of data allows for ten times more weight to Country B than A. Assuming that country means and standard deviations are used as the basis for norm aggregation, these may be equally weighted or weighted in proportion to country population sizes or in proportion to relevant industry sector sizes in the countries concerned.
- As adherence to these guidelines involves the exercise of considerable expert judgment, professional support should be sought in deciding how to proceed on a case by case basis.
- Professional judgment should also be referred to when aggregation of norms is not possible.

In practice, these guidelines can be followed to produce virtual norm groups. We can define a virtual norm group as one that has been created by the aggregation of suitably weighted population or user norms for a specific purpose (such as the sort of multinational comparisons being discussed here). Clearly, caution needs to be exercised where there are differences between average raw scores for country norms, where such differences relate to different language versions of the instrument, and where such differences cannot be explained by differences in demographic mix (i.e., they are not due to sampling bias) or by cultural factors.

RECOMMENDED PRACTICE

Where more than one language has been used in questionnaire administration or where a single language of administration has been used but the candidates are from different linguistic and cultural backgrounds, this should be taken into account during the interpretation process. This can be achieved through working with local experts who are familiar with the culture. Without this it is very easy to misinterpret behavior or to fail to appreciate the underlying potential because the person is conforming to unfamiliar business or social practices.

Comparing all candidates' scores to the same (multi-national, multi-cultural, or multi-lingual) norm will accentuate the differences due to the cultural behavior patterns of their background—even though these may be moderated by experience of the different environment. Because the scores are influenced by these arbitrary cultural differences in behavior, the measurement of the level of the underlying trait will suffer. On the other hand, using only the individual language norms for each candidate will reflect the underlying trait levels without relating to any differences between cultural norms, which may be relevant.

In international contexts, inferences from scores should take into account both factors. In interpreting the score it is important to know whether it shows, say, a moderate or extreme tendency to behave in a particular manner in general (relative to the home country norm) and how this would seem in a different context (multinational norm).

Where a questionnaire has been designed for international use and there has been careful adaptation into different languages, the differences between language norms are likely to be small. Comparison of each of the local language norms with an aggregated multi-national norm will show where the average profiles diverge. This information should be available to the interpreter, either through the norm information for the different groups, or through qualitative data on where a particular pair of countries or norms differs. For example, if a firm is recruiting candidates from Spain, Portugal, and Belgium, they might use an aggregate norm from those countries to create a base for comparison. If the firm wants to employ those people in Belgium and the Netherlands, an additional comparison to an aggregate norm from those two countries might be informative. Comparison of the local country norms with virtual aggregate norms will show where the average profiles diverge. Scores from scales in which there are differences should be interpreted with care, taking into account how candidates may moderate their behavior in different cultures.

In summary, for international assessments it is recommended that both local national and relevant multi-national aggregations of national norms are used and that areas where these give rise to differences in standardized scores should be highlighted and considered by users in the light of what is known about possible sample, translation, or cultural effects.

REFERENCES

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

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

Brown, A., & Bartram D. (2008). IRT model for recovering latent traits from forced-choice personality tests. Paper presented at the *Society for Industrial and Organizational Psychology (SIOP) Annual Conference*, San Francisco.

Cohen, J. (1988). *Statistical power analysis for the behavioral sciences* (Second ed.). New York: Lawrence Erlbaum.

Costa, P. T., Terracciano, A., & McCrae, R. R. (2001). Gender differences in personality traits across cultures: Robust and surprising findings. Journal of *Personality and Social Psychology*,*81*(2), 322–331.

Cronbach, L. J. (1990). *Essentials of Psychological Testing* (5th ed.). New York, Harper & Row.

Daouk, L., Rust, J., & McDowall, A. (2005). Testing across languages and cultures: Challenges for the development and administration of tests in the internet era. *Selection & Development Review*, *21*(4), 11–13.

Hambleton, R. K. (2005). Issues, designs and technical guidelines for adapting tests into multiple languages and cultures. In R. K. Hambleton, P. F. Merenda, & C. D. Spielberger (Eds.), *Adapting Educational and Psychological Tests for Cross-Cultural Assessment.* Mahwah, NJ: Lawrence Erlbaum Associates.

Hambleton, R. K., & Patsula, L. (1999). Increasing the Validity of Adapted Tests: Myths to be avoided and Guidelines for Improving Test Adaptation Practices. *Applied Testing Technology*, *1*(1), 1–16.

Hanges, P. J. (2004). Response bias correction procedure used in GLOBE. In P. J. H. R.J. House, M. Javidan, P. W. Dorfman, & V. Gupta (Eds.), *Culture, Leadership and Organizations: The GLOBE study of 62 societies*. Thousand Oaks, CA: Sage.

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

Hofstede, G. (2003). *Culture's consequences: Comparing values, behaviors, institutions and organizations across nations.* Beverly Hills, CA: Sage.

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

McCrae, R. R., & Terracciano, A. (2005). Universal features of personality traits. *Journal of Personality and Social Psychology*,*88*(3), 547–561.

Schwartz, S. H. (1992). Universals in the content and structure of values: Theoretical advances and empirical tests in 20 countries. In M. P. Zanna (Ed.), *Advances in experimental social psychology,* Vol *25* (pp. 1–65). New York: Academic Press.

SHL. (1999). *OPQ32 Manual.* Thames Ditton: SHL Group plc.

SHL. (2006). *OPQ32 Technical Manual.* Thames Ditton: SHL Group Ltd, and from http://www.shl.com/opqtechnicalmanual