Research Report

What Do Cross-National Comparisons of Personality Traits Tell Us?

The Case of Conscientiousness

Steven J. Heine, Emma E. Buchtel, and Ara Norenzayan

University of British Columbia

ABSTRACT—*Much research contrasts self-reported personality traits across cultures. We submit that this enterprise is weakened by significant methodological problems (in particular, the reference-group effect) that undermine the validity of national averages of personality scores. In this study, behavioral and demographic predictors of conscientiousness were correlated with different cross-national measures of conscientiousness based on self-reports, peer reports, and perceptions of national character. The predictors correlated strongly with perceptions of national character, but not with self-reports and peer reports. Country-level self- and peer-report measures of conscientiousness failed as markers of between-nation differences in personality.*

Canadians and Americans agree that Canadians are generally more agreeable than their southern neighbors. Is this judgment accurate? One way to investigate this question would be to compare the average agreeableness of Canadians with that of Americans to see how well the country scores align with people's perceptions. In an influential analysis (McCrae & Terracciano, 2006; Terracciano et al., 2005), there was no correlation between profiles of actual reported personality traits and people's perceptions of the character of their own country. On this basis, McCrae and Terracciano (2006) argued that such perceptions of national character (PNCs) are illusory.

We found these results surprising (also see McGrath & Goldberg, 2006). Past research has shown that agreement on

Address correspondence to Steven J. Heine, 2136 West Mall, University of British Columbia, Vancouver, British Columbia V6T 1Z4, Canada, e-mail: heine@psych.ubc.ca.

cultural differences becomes stronger with increased cultural contact (Triandis & Vassiliou, 1967), and that perceptions of group differences are often, although not always, quite accurate (Jussim, 2005; McCauley, 1995). Conceptions of culture as a shared meaning system suggest that one can make meaningful and relatively accurate judgments about one's own culture (Wan et al., 2007).

Relying on self- and peer reports to assess individual differences in personality has become a widely used approach to investigating cross-cultural differences (e.g., McCrae & Terracciano, 2006; Schmitt et al., 2007). Therefore, determining the validity of this method across cultures is of utmost importance (Norenzayan & Heine, 2005). In the study reported here, we investigated the validity of self- and peer reports and PNC comparisons across cultures. We argue that methodological problems compromise the validity of aggregate self- and peer-report data across cultures. Focusing on conscientiousness as a test case, we present new behavioral and demographic evidence suggesting that PNC data are more valid as a measure of between-culture differences in behavior than are aggregate self-report and peer-report data.

Comparing means on subjective Likert self-report scales is the most commonly used method for investigating cross-cultural differences, yet there are many methodological challenges associated with this approach (for discussion, see Cohen, 2007; van de Vijver & Leung, 2001). One such challenge is the reference-group effect (RGE; Heine, Lehman, Peng, & Greenholtz, 2002; Peng, Nisbett, & Wong, 1997), the tendency for people to respond to subjective self-report items by comparing themselves with implicit standards from their culture (Heine et al., 2002). Consider how one would respond to the items on the Revised NEO Personality Inventory (NEO-PI-R; Costa & McCrae, 1992). For example, one's response to "I am not a very methodical

person" would hinge greatly on one's understanding of the norms for being methodical. Because norms differ across cultures, the RGE systematically distorts cultural differences. Note that the RGE is a problem only for comparing aggregate data between cultures; it is not implicated in contrasts of individuals within a shared cultural context.

The RGE is an issue regarding self- and peer reports, as respondents need to compare their evaluations against standards largely based on observations of people within their cultures. However, the RGE is less likely to contaminate PNCs. This is because when people are asked to evaluate their "average compatriot," they likely bring to mind a standard that lies outside their own culture, for example, a perceived international norm. Some evidence supporting this idea was obtained in the influential study mentioned earlier: The variance of the PNC cultural means was much larger than the variance of the cultural means of self- or peer ratings (Terracciano et al., 2005).

ESTABLISHING VALIDITY: BEHAVIORAL AND DEMOGRAPHIC PREDICTORS OF PERSONALITY ACROSS CULTURES

What is the existing evidence that cross-national comparisons of personality traits are valid? One piece of evidence is that some of Hofstede's (2001) cultural dimensions correlate with some of the Big Five measures (McCrae, 2002). However, these correlations are not a priori theoretically relevant (e.g., neuroticism and masculinity are correlated, but for theoretically unknown reasons), and we note that no significant correlations between Big Five and Hofstede's dimensions were replicated across three independent measures of the Big Five (McCrae, 2002; McCrae et al., 2005; Schmitt et al., 2007). Another piece of evidence is that geographically and historically related cultures (such as Germany and Austria) have similar personality profiles (McCrae & Terracciano, 2006). However, cultures with similar reference groups (i.e., cultures that share similar norms) would yield this same pattern. Other criteria, particularly those that do not rely on self-report measures, are needed to validate cross-national comparisons of personality traits. We focus here on a criterion that is frequently relied on in evaluating trait measures (e.g., Gosling, Ko, Mannarelli, & Morris, 2002): whether the trait scores predict actual behavior.

METHOD AND RESULTS

A challenge in validating personality traits across cultures is assessing behaviors that unambiguously map on to the Big Five dimensions. Past within-culture studies have suggested that the personality trait of conscientiousness is relatively easily identified, through both behavior and markers of behavior (see Gosling et al., 2002; Roberts, Kuncel, Shiner, Caspi, & Goldberg, 2007). A second challenge is the sheer cost and difficulty of obtaining such measures in a large number of different cultural contexts. Fortuitously, a large-scale research project by

Levine and Norenzayan (1999) assessed behavioral measures of "pace of life" in major cities from 31 countries. Specifically, walking speed, postal workers' speed, and the accuracy of clocks in public banks were measured. For the measure of walking speed, 35 men and 35 women were timed unobtrusively as they walked alone for a distance of 60 ft in unobstructed public squares during business hours in major downtown locations. Postal workers' efficiency was measured by the amount of time it took to buy one stamp, using a bill that required change in both coins and paper. At least eight postal workers at randomly selected postal offices were timed in each city. Finally, the accuracy of 15 randomly selected bank clocks was assessed in each city. These variables, which were adequately intercorrelated, are arguably behavioral indicators of conscientiousness (see Allik & McCrae, 2004).

There are also demographic variables associated with personality. Roberts et al. (2007) found that occupational success can be reliably predicted by personality traits, and that among the Big Five, conscientiousness is the best predictor of such success (see Judge, Higgins, Thoresen, & Barrick, 1999). We did not have direct measures of the average occupational success of nations. However, one national indicator of economic success is per capita gross domestic product (GDP). No doubt, a nation's GDP is a complex variable reflecting a variety of historical, political, and sociodemographic factors. Nevertheless, a country's per capita GDP is related to the occupational success of its citizens, so per capita GDP should correlate with the average conscientiousness of its citizens. We included per capita GDP as reported in *The World Factbook* (Central Intelligence Agency, 2007) as a demographic variable in this study.

Personality is also reliably associated with longevity (Roberts et al., 2007), and conscientiousness is the best predictor of longevity among the Big Five. The average life expectancy of nations was also obtained from *The World Factbook* (Central Intelligence Agency, 2007).[1]

We correlated our validity criteria (walking speed, postal workers' speed, accuracy of clocks in public banks, per capita GDP, and longevity) with a variety of measures of conscientiousness.[2] Specifically, conscientiousness was measured by aggregated self-reports on the NEO-PI-R (McCrae, 2002) and the Big Five Inventory (Benet-Martínez & John, 1998; Schmitt et al., 2007), aggregated peer reports on the NEO-PI-R (McCrae et al., 2005), and average PNC scores from Terracciano et al. (2005). Table 1 reports the correlations in two ways. In each cell, the correlation on the left was calculated using pair-wise deletion of data. These correlations are based on between 17 and 55

---

[1]We have provided a table of all the values for each country on the Web at http://www.psych.ubc.ca/~heine/personalitydata.xls.

[2]The analyses reported by Terracciano et al. (2005) were conducted by correlating the individual facet scores from the different measures; in contrast, we report the correlations between factor scores for personality and behavioral and demographic variables. We were unable to analyze the facet scores, as these data were not available to us.

TABLE 1

Correlations Between Personality Traits and Behavioral and Demographic Measures of Conscientiousness

Personality measure	Validity criterion					
	Postal workers' speed	Clock accuracy	Walking speed	GDP	Longevity	Average
Self-report NEO-PI-R	−.57*/−.58	−.34/−.39	−.27/−.43	−.66**/−.16	−.38*/−.38	−.43
Self-report BFI	−.28/−.42	−.07/−.18	−.14/.15	−.32*/.08	−.51**/−.16	−.19
Peer-report NEO-PI-R	.16/.42	.07/−.10	−.11/.01	.03/−.12	.33*/−.11	.06
PNC	.68**/.72*	.69**/.71*	.50*/.68*	.40**/.69*	.01/.74*	.61

Note. In each cell, the correlation on the left was calculated using pair-wise deletion of missing data ($n = 17–55$), and the correlation on the right was calculated using case-wise deletion of missing data ($n = 10$). BFI = Big Five Inventory (Benet-Martínez & John, 1998; Schmitt et al., 2007); GDP = gross domestic product (Central Intelligence Agency, 2007); NEO-PI-R = Revised NEO Personality Inventory (McCrae, 2002; McCrae et al., 2005); PNC = perceptions of national character (Terracciano et al., 2005).
$*p < .05$. $**p < .01$.

countries. The right-hand correlation in each cell was calculated using case-wise deletion of data, which reduced the sample to the 10 countries for which we had data for all measures.[3] Although the pair-wise-deletion analyses had a larger sample that was thus more representative of a worldwide sample, the case-wise-deletion analyses allowed for more accurate comparisons across measures, as in these analyses the measures were compared across the same countries.

The analyses revealed that the aggregate self-report data from the NEO-PI-R correlated negatively with all the validity criteria for both sets of countries (average $r = −.43$).[4] Likewise, the aggregate self-report data from the Big Five Inventory correlated negatively with the validity criteria (average $r = −.19$). The aggregate peer-report data from the NEO-PI-R showed mixed correlations with the validity criteria. Overall, the pattern yielded null correlations ($r = .06$). In sum, none of the self-report or peer-report data were reliably associated with the validity criteria in the predicted direction.

In contrast, the PNC data correlated positively and significantly with all five of the validity criteria, with the sole exception of longevity with pair-wise deletions. The average correlation between the PNC data and the validity criteria was .61. This is evidence that the PNC data have greater validity for measuring conscientiousness than the aggregate self- and peer-report data do.

The five validity criteria we used are not independent; indeed, the correlations among them ranged from .25 to .71, and the average correlation was .47—a value higher than is typically obtained for correlations among items in trait measures. One possible explanation for these high interitem correlations is that the behavioral criteria simply reflect the consequences of wealth, and that participating in a wealthy society leads people to work more effectively, walk faster, and live longer. We there-

fore recalculated the correlations after partialing out per capita GDP (see Table 2). The average partial correlation between the NEO-PI-R self-report data and the criteria was weakly negative, as was the average correlation between the Big Five Inventory self-report data and the criteria. The average correlation between the NEO-PI-R peer-report data and the criteria was weakly positive ($r = .15$). In contrast, the average correlation between the PNC data and the validity criteria was moderately positive ($r = .41$), and seven of the eight partial correlations for the PNC data were positive. In sum, after per capita GDP was partialed out, the aggregate self- and peer-report measures did not reliably predict the remaining validity criteria, whereas the PNC data still showed good validity.

DISCUSSION

The lack of correlation between aggregate self- and peer-report personality measures and people's views of their compatriots, as reported in Terracciano et al. (2005), is a problem. However, we submit that it is caused by various methodological challenges that plague cross-cultural comparisons of mean scores on Likert scales. We suspect that the primary confound undermining the aggregate personality measures is the RGE; however, a number of other response biases may also be implicated (see Cohen, 2007; van de Vijver & Leung, 2001). In contrast, people's impressions of their compatriots' personalities appear to reasonably reflect measures of their compatriots' behaviors.

Cross-cultural comparisons of aggregate self- and peer-report measures of personality fail a number of validity tests. First, as McCrae (2001) noted, country rankings do not correlate with how cultural experts view the ranked countries, and thus are not validated by expert opinions. Second, we have found that aggregate self- and peer-report measures of personality (at least in the case of conscientiousness) do not correlate with either behavioral or demographic indicators. Third, these measures do not correlate with people's impressions of their compatriots (viz., the PNC data). The fact that aggregate self- and peer-report data do not correlate with validity criteria is consistent with the fact

[3]The 10 countries that had data for all dependent variables were Canada, the Czech Republic, France, Germany, Hong Kong, Indonesia, Italy, Japan, South Korea, and the United States.

[4]All average correlations were calculated from r-to-Z-to-r transformations (see McNemar, 1962).

TABLE 2

Partial Correlations Between Personality Traits and Behavioral and Demographic Measures of
Conscientiousness, Controlling for Per Capita Gross Domestic Product

| Personality measure | Validity criterion | | | | |
	Postal workers' speed	Clock accuracy	Walking speed	Longevity	Average
Self-report NEO-PI-R	−.26/−.57	−.02/−.41	.33/−.56	.03/−.47	−.26
Self-report BFI	−.11/−.49	.10/−.35	.10/.16	−.42**/−.45	−.20
Peer-report NEO-PI-R	.18/.51	.06/−.01	−.18/.22	.39**/−.02	.15
PNC	.60*/.69*	.61**/.42	.35/.25	−.31*/.42	.41

Note. In each cell, the correlation on the left was calculated using pair-wise deletion of missing data ($n = 17$–55), and the correlation on the right was calculated using case-wise deletion of missing data ($n = 10$). BFI = Big Five Inventory (Benet-Martínez & John, 1998; Schmitt et al., 2007); NEO-PI-R = Revised NEO Personality Inventory (McCrae, 2002; McCrae et al., 2005); PNC = perceptions of national character (Terracciano et al., 2005).
*$p < .05$. **$p < .01$.

that cross-cultural comparisons of mean scores on Likert scales often yield findings sharply at odds with those obtained using experimental measures (Heine et al., 2002).

A key tenet of cultural psychology is that culture and self are mutually constituted (Shweder, 1990). We think the present investigation demonstrates this point well. Participating in a world where clocks are on time and people work effectively and earn a good income leads people to adopt conscientious habits, and these habits continue to shape and reinforce the surrounding culture. Culture and self thus come to make each other up. We suggest that people participating in such a culture develop personalities that are, on average, more conscientious than the personalities of people who participate in "event-time cultures," where activities transpire according to a spontaneous schedule (Levine & Norenzayan, 1999). Participation in a cultural world comes to shape the cultural norms that people perceive as reflecting general conscientiousness. These norms are apparent when people evaluate their culture (Wan et al., 2007), although they might not be detected in self-report assessments as people implicitly compare themselves with these perceived norms.

Our investigation was limited to conscientiousness. It is conceivable that for the remaining four Big Five personality traits, the PNC measure is not accurate, but the self- and peer-report measures are. However, it is not obvious why a different pattern would emerge for the other traits. One clear direction for future research would be to obtain measures of behavioral markers of other personality traits across a large array of cultures. A major challenge lies in first identifying cross-culturally valid behavioral markers of the remaining Big Five traits.

A second limitation of our study is that we did not assess our various measures with the same group of participants: Different individuals completed each of the personality and behavioral measures, and, moreover, our GDP variable was not an aggregate of individual responses, as the other measures were. The ideal study would involve a multilevel framework in which personality

traits, behavioral measures, and economic variables are collected from the same individuals.

Another limitation of our investigation concerns the quality of the validity criteria. It can be debated how well the specific criteria that we chose map on to conscientiousness. More generally, behavioral measures have a number of weaknesses for assessing personality (see Funder, 2007). Furthermore, a given behavior might have different meanings in different cultures; for example, frequency of bowing would be a poor indicator of politeness in the United States, although it would be a good indicator in Japan. However, in some key respects, behaviors are especially strong measures for cross-cultural research as they are immune to most of the methodological artifacts that plague cross-cultural comparisons (Peng et al., 1997). Behavioral and demographic measures are valuable but not faultless criteria by which to validate cultural differences. Research on cross-cultural differences would be improved by greater use of such validity criteria to corroborate findings that emerge from questionnaire and laboratory studies.

CONCLUSION

Mindlessly applying inaccurate stereotypes opens the door for misunderstanding and mistreating people from other cultures (McCrae, 2001). Yet equally problematic to cross-cultural relations is insensitivity to real cultural differences. Such insensitivity encourages ethnocentric projections, perpetuating the unfounded belief that people in other cultures are "just like us." The evidence discussed here suggests that PNCs, not aggregate self- or peer reports of personality, reflect the kinds of cultural differences in personality that are important for increasing intercultural understanding.

REFERENCES

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

Benet-Martínez, V., & John, O.P. (1998). Los Cinco 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.

Central Intelligence Agency. (2007). *The world factbook*. Washington, DC: Potomac Books.

Cohen, D. (2007). Methods in cultural psychology. In S. Kitayama & D. Cohen (Eds.), *Handbook of cultural psychology* (pp. 196–236). New York: Guilford.

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.

Funder, D.C. (2007). *The personality puzzle* (4th ed.). New York: W.W. Norton.

Gosling, S.D., Ko, S.J., Mannarelli, T., & Morris, M.E. (2002). A room with a cue: Personality judgments based on offices and bedrooms. *Journal of Personality and Social Psychology*, *82*, 379–398.

Heine, S.J., Lehman, D.R., Peng, K., & Greenholtz, J. (2002). What's wrong with cross-cultural comparisons of subjective Likert scales?: The reference-group effect. *Journal of Personality and Social Psychology*, *82*, 903–918.

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

Judge, T.A., Higgins, C.A., Thoresen, C.J., & Barrick, M.R. (1999). The big five personality traits, general mental ability, and career success across the life span. *Personnel Psychology*, *52*, 621–652.

Jussim, L. (2005). Accuracy in social perception: Criticisms, controversies, criteria, components, and cognitive processes. In M.P. Zanna (Ed.), *Advances in experimental social psychology* (Vol. 37, pp. 2–97). San Diego, CA: Elsevier Academic Press.

Levine, R.V., & Norenzayan, A. (1999). The pace of life in 31 countries. *Journal of Cross-Cultural Psychology*, *30*, 178–205.

McCauley, C. (1995). Are stereotypes exaggerated? A sampling of racial, gender, occupational, academic, and political stereotypes. In Y.T. Lee, L. Jussim, & C.R. McCauley (Eds.), *Stereotype accuracy: Toward appreciating group differences* (pp. 293–312). Washington, DC: American Psychological Association.

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

McCrae, 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. 105–126). New York: Kluwer.

McCrae, R.R., & Terracciano, A. (2006). National character and personality. *Current Directions in Psychological Science*, *15*, 156–161.

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

McGrath, R.E., & Goldberg, L.R. (2006). How to measure national stereotypes? *Science*, *311*, 776–777.

McNemar, Q. (1962). *Psychological statistics* (3rd ed.). New York: Wiley.

Norenzayan, A., & Heine, S.J. (2005). Psychological universals: What are they and how can we know? *Psychological Bulletin*, *131*, 763–784.

Peng, K., Nisbett, R.E., & Wong, N.Y.C. (1997). Validity problems comparing values across cultures and possible solutions. *Psychological Methods*, *2*, 329–344.

Roberts, B.W., Kuncel, N.R., Shiner, R., Caspi, A., & Goldberg, L.R. (2007). The power of personality: The comparative validity of personality traits, socioeconomic status, and cognitive ability for predicting important life outcomes. *Perspectives on Psychological Science*, *2*, 313–345.

Schmitt, D.P., Allik, J., McCrae, R.R., Benet-Martínez, V., Alcalay, L., Ault, L., et al. (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.

Shweder, R.A. (1990). Cultural psychology: What is it? In J.W. Stigler, R.A. Shweder, & G. Herdt (Eds.), *Cultural psychology: Essays on comparative human development* (pp. 1–43). Cambridge, England: Cambridge University Press.

Terracciano, A., Abdel-Khalek, A.M., Ádám, N., Adamovová, L., Ahn, C.-k., Ahn, H.-n., et al. (2005). National character does not reflect mean personality trait levels in 49 cultures. *Science*, *310*, 96–100.

Triandis, H.C., & Vassiliou, V. (1967). Frequency of contact and stereotyping. *Journal of Personality and Social Psychology*, *7*, 316–328.

van de Vijver, F.J.R., & Leung, K. (2001). Personality in cultural context: Methodological issues. *Journal of Personality*, *69*, 1007–1032.

Wan, C., Chiu, C., Tam, K., Lee, S., Lau, I.Y., & Peng, S. (2007). Perceived cultural importance and actual self-importance of values in cultural identification. *Journal of Personality and Social Psychology*, *92*, 337–354.