

## Reference

1. R. J. Charlson, F. P. J. Valero, J. H. Seinfeld, *Science* **308**, 806 (2005).

## How to Measure National Stereotypes?

BECAUSE IT IS PARTICULARLY DIFFICULT TO EVALUATE the accuracy of national stereotypes, the Report by A. Terracciano *et al.* ("National character does not reflect mean personality trait levels in 49 cultures," 7 Oct. 2005, p. 96) examining the relations between ratings of national character and ratings of individuals in 49 different cultures represents quite a technical achievement. Studies of stereotypes usually suggest that stereotypic beliefs contain a kernel of truth: The perceived differences between groups do in fact exist, but they are smaller than the stereotype would suggest (1, 2). Terracciano *et al.* instead found that, on average, there was no relation between national stereotypes and self and other descriptions. Some methodological weaknesses of their study must be considered, however.

One issue is their almost exclusive reliance on college student samples. Although there is some evidence that cross-cultural comparisons between college students may generalize to broader populations (3), there is also substantial evidence that findings with college stu-

dents frequently do not so generalize (4). These findings do not invalidate college student samples as representations of broader national populations, but neither do they justify assuming college students provide an acceptable proxy for the population as a whole.

A second issue is whether the authors have provided a sufficient evaluation of national character. The authors reduce national character to personality traits. This ignores other potential elements of stereotype, most particularly differences in values, beliefs, or perceptions that are not adequately included in the measures used in this study.

Finally, Terracciano *et al.*'s measures of perceived national character were the mean ratings of the culture by members of that culture. Stereotypes are usually defined in terms of perceptions of the target group by outside observers. Moreover, their measure of actual national character was the mean ratings of oneself or a significant other. In other words, the measurement of national character was based on the ratings of a culture, whereas the measurement of actual character was based on the ratings of a person. The contexts of the two kinds of assessments were quite different and potentially not comparable.

It is increasingly evident that context is an important contributor to outcomes on rating scales (5). There is even evidence that cultural

## Letters to the Editor

Letters (~300 words) discuss material published in *Science* in the previous 6 months or issues of general interest. They can be submitted through the Web ([www.submit2science.org](http://www.submit2science.org)) or by regular mail (1200 New York Ave., NW, Washington, DC 20005, USA). Letters are not acknowledged upon receipt, nor are authors generally consulted before publication. Whether published in full or in part, letters are subject to editing for clarity and space.

differences by themselves can produce differences in the context of the measurement (6). A person familiar to the respondent will likely be evaluated in relation to other individuals familiar to the respondent, while a person asked to rate the culture will rate it in relation to other cultures. It is not surprising then to find that these ratings were on average unrelated to ratings of the country's national character.

It is possible that there really is no relation between national stereotypes and actual behaviors. One must wonder, however, what is the source of the variability in the ratings of cultures. Why, for example, do the German Swiss believe they are so conscientious? Even more curious is why Indonesians and Chileans accept that they are not. It seems likely that when asked to rate themselves on conscien-

tiousness, German Swiss evaluate themselves in light of those around them. A more definitive test would be to have the German Swiss rated by members of other cultures, but then that is presumably the kernel from which cultural stereotypes germinate in the first place.

**ROBERT E. MCGRATH<sup>1</sup>\* AND LEWIS R. GOLDBERG<sup>2</sup>**

<sup>1</sup>School of Psychology, Fairleigh Dickinson University, 1000 River Road, Teaneck, NJ 07666, USA. <sup>2</sup>Oregon Research Institute, 1715 Franklin Boulevard, Eugene, OR 97403-1983, USA.

\*To whom correspondence should be addressed. E-mail: mcgrath@fdu.edu

#### References

1. J. S. Hyde, *Am. Psychol.* **60**, 581 (2005).
2. H. M. Coon, M. Kemmelmeier, *J. Cross-Cultural Psychol.* **32**, 348 (2001).
3. S. H. Schwartz, A. Bardi, *J. Cross-Cultural Psychol.* **32**, 268 (2001).
4. R. A. Peterson, *J. Consumer Res.* **28**, 450 (2001).
5. N. Schwarz, *Am. Psychol.* **54**, 93 (1999).
6. N. Schwarz, *J. Consumer Res.* **29**, 588 (2003).

CONSISTENT WITH A LONG-HELD VIEW IN SOCIAL psychology, A. Terracciano and colleagues claim that national stereotypes lack accuracy ("National character does not reflect mean personality trait levels in 49 cultures," Reports, 7 Oct. 2005, p. 96). Although it is possible that their findings demonstrate people's inability to discern the attributes of their own groups, three alter-

native explanations need to be considered.

First, the criterion scores, which were obtained from responses on a personality inventory [the Revised NEO Personality Inventory (NEO-PI-R)], were less variable than the stereotype scores, which were obtained with a new instrument [the National Character Survey (NCS)]. Arguably, the greater length of the NEO-PI-R facet scales (eight items) relative to the NCS scales (one item) contributed to this difference. Furthermore, the nonrepresentative sampling of respondents could have reduced the variability of the criterion scores, as college students tend to share similarities in different cultures.

Second, the similarity of the sample profiles was assessed with intraclass correlation coefficients (ICCs). ICCs are used for dyadic data that cannot be sorted. When judgments are correlated with criteria, Pearson correlations are more appropriate. These indices are only sensitive to profile similarity, not to differences in variability.

Third, national characteristics and stereotypes can be specific. The Japanese may be uniquely characterized by their deference, whereas people from the United States may be known for their materialism. If so, measures of profile similarity gravitate toward zero as a function of profile length.

Failures to reject a null hypothesis are usually not newsworthy. A typical response is to design a

study to minimize contaminating effects. Here, however, the embrace of the null hypothesis is also a conceptual surprise. Historically, research on the five-factor model of personality has been predicated on observer agreement, where agreement was thought to imply accuracy. Now, the role of observer agreement is to signal inaccuracy. It is certainly possible that perceptions of nations are qualitatively different from perceptions of individuals, but to find out we need a process model that specifies how people judge national character and how they might agree without being accurate.

**JOACHIM I. KRUEGER AND JACK C. WRIGHT**

Department of Psychology, Brown University, Hunter Laboratory 295, 89 Waterman Street, Providence, RI 02912, USA.

#### Response

WE AGREE WITH MCGRATH AND GOLDBERG THAT national stereotypes include more than national character, and beliefs about national differences in appearance, attitudes, or athletic abilities may or may not be accurate. Our study focused on personality traits, which seem to define the core of national character. To the extent that the five-factor model (FFM) is comprehensive, our National Character Survey (NCS) measured key features of national character, and we found no evidence for a kernel of truth in these stereotypes.

Student samples may or may not generalize to

the general population, but there is reason to think they did in this instance. Previous work comparing personality ratings from students and adults showed generalizability (1, 2). In our study, the self-report criterion included data from adults, and in cultures where personality profiles were based solely on adult self-reports, we found no support for the accuracy of national character stereotypes. Finally, national character ratings by adults in Ethiopia, Italy, and the Philippines (3) agreed with students' NCS rating. All the available data suggest that the age of the raters did not affect the outcome.

McGrath and Goldberg suggest that our "national character was based on the ratings of a culture" and thus we compared cultures with people. In fact, the NCS asked respondents to describe the typical member of their culture. Italian raters, for example, were asked to respond to the stem "Italians are likely to be ...". Factoring NCS responses led to the familiar FFM, as one would expect ratings of persons to do (4). Personality and national character assessments were fully comparable; they were simply different.

McGrath and Goldberg also raise the issue of changing frames of reference. There is no evidence that these compromise personality ratings. The reference group effect would tend to eliminate any cross-cultural differences in personality traits and render them meaningless [see note (27)

of our Report], but our aggregate scores varied systematically across cultures, formed clear geographical clusters, and showed meaningful correlations with culture-level variables (1). Thus, the reference effect cannot explain the failure to find correlations with NCS scales.

Finally, McGrath and Goldberg also suggest that a more definitive test of stereotype accuracy would employ out-group judgments. However, the literature (3, 5, 6) and our own data (7) indicate that out-group ratings of national character are very similar to in-group ratings, at least between neighboring cultures. Given such similarity and the simple fact that people presumably know members of their own culture better than foreigners do, it is not clear how out-group ratings would be accurate.

Krueger and Wright propose alternative explanations for our finding that national stereotypes are inaccurate, but none seems justified.

The first concerns reliability. The eight-item NEO-PI-R (8) scales are presumably more reliable than the single-item NCS scales. Greater reliability means less random error, and that should increase, not decrease, the variability of the NEO-PI-R scales across cultures. The greater variability of the NCS scores that we in fact observed is more likely due to the exaggeration that is characteristic of stereotyping.

Although the observer rating criteria were

obtained from college students, the inaccuracy of stereotypes was confirmed by self-report data [our Report; (2)] from samples of adults as well as students.

Krueger and Wright note that ICCs are used for interchangeable dyadic data. They are also used to assess absolute agreement that takes into account the means and variances of two sets of measures. Pearson correlations are sensitive only to the shape of a profile; they ignore differences in elevation. Imagine two sets of 30 scores with perfectly parallel profiles but with a constant difference of 10 *T*-score points. The Pearson correlation would be 1.0, suggesting perfect agreement despite the large mean differences. The ICC method we used would give a much lower or negative coefficient because of its sensitivity to the 10 *T*-score points difference, correctly showing that the profiles are substantively very different. However, we reanalyzed the data using Pearson correlations and found very similar results, with median correlations of 0.08 and -0.01 for the observer rating and self-report data, respectively.

Krueger and Wright suggest that stereotypes may be specific for different cultures and their effects may be diluted by analyses of all 30 traits in each culture. Although some stereotypical traits may be more salient than others, the aggregate NCS ratings were highly reliable for all five factors and 30 facets, indicating that the raters

shared perceptions of the typical member of their culture on most traits.

Null findings are not newsworthy if they are based on a weak study. But our project used data from over 40,000 respondents in 49 cultures, employed a comprehensive selection of personality traits, examined agreement both across and within cultures, and replicated the null findings using two methods of assessing personality. Agreement between observers was taken as evidence of accuracy [e.g., (I)] and the failure to find agreement was considered evidence of inaccuracy.

#### ANTONIO TERRACCIANO AND ROBERT R. MCCRAE

National Institute on Aging, NIH, DHHS, Gerontology Research Center, 5600 Nathan Shock Drive, Baltimore, MD 21224, USA. E-mails: terraccianoa@grc.nia.nih.gov; mccraej@grc.nia.nih.gov

#### References and Notes

1. R. R. McCrae *et al.*, *J. Pers. Soc. Psychol.* **89**, 407 (2005).
2. R. R. McCrae, in *The Five-Factor Model of Personality Across Cultures*, R. R. McCrae, J. Allik, Eds. (Kluwer Academic/Plenum, New York, 2002), pp. 105–125.
3. A. T. Church, M. S. Katigbak, in *The Five-Factor Model of Personality Across Cultures*, R. R. McCrae, J. Allik, Eds. (Kluwer Academic/Plenum, New York, 2002), pp. 129–154.
4. W. T. Norman, L. R. Goldberg, *J. Pers. Soc. Psychol.* **4**, 681 (1966).
5. D. Peabody, *National Characteristics* (Cambridge University Press, New York, 1985).
6. J. Krueger, *Eur. J. Soc. Psychol.* **26**, 961 (1996).
7. A. Terracciano, R. R. McCrae, paper presented at the 113th Annual Convention of the American Psychological Association, Washington, DC, 2005.
8. P. T. Costa Jr., R. R. McCrae, *Revised NEO Personality Inventory (NEO-PI-R) and NEO Five-Factor Inventory (NEO-FFI) Professional Manual* (Psychological Assessment Resources, Odessa, FL, 1992).
9. This research was supported in part by the Intramural Research Program of the NIH, National Institute on Aging.

### TECHNICAL COMMENT ABSTRACTS

#### Comment on “Zircon Thermometer Reveals Minimum Melting Conditions on Earliest Earth” I

Andrew Glikson

Watson and Harrison (Reports, 6 May 2005, p. 841) proposed a model for early Earth magmatism based on crystallization temperatures of Hadean zircons. However, detrital zircon populations are skewed relative to the composition of their source terrains, Archaean isotopic and geochemical mantle signatures preclude reincorporation of Hadean continental crust into the early mantle, and the effects of early impacts should be considered.

Full text at [www.sciencemag.org/cgi/content/full/311/5762/779a](http://www.sciencemag.org/cgi/content/full/311/5762/779a)

#### Comment on “Zircon Thermometer Reveals Minimum Melting Conditions on Earliest Earth” II

Allen P. Nutman

Watson and Harrison (Reports, 6 May 2005, p. 841) interpreted low temperatures (~700°C) for Hadean zircons as evidence of the existence of wet, minimum-melting conditions within 200 million years of solar system formation. However, high-temperature melts (~900°C) are zircon undersaturated and crystallize zircon only after substantial temperature drop during fractional crystallization. Zircon thermometry cannot distinguish between low- and high-temperature Hadean igneous sources.

Full text at [www.sciencemag.org/cgi/content/full/311/5762/779b](http://www.sciencemag.org/cgi/content/full/311/5762/779b)

#### Response to Comments on “Zircon Thermometer Reveals Minimum Melting Conditions on Earliest Earth”

E. B. Watson and T. M. Harrison

The mean crystallization temperature of Hadean zircons based on titanium content is ~680°C. This value corresponds to the temperature of wet minimum melting in present-day crust. The low variance of the temperature distribution ( $\pm 25^\circ\text{C}$ ) also points to Hadean zircon growth under conditions that were highly reproducible and thermally regulated. Eutectic-like melting is particularly capable of providing such regulation and is consistent with Hadean zircon growth during wet crustal fusion.

Full text at [www.sciencemag.org/cgi/content/full/311/5762/779c](http://www.sciencemag.org/cgi/content/full/311/5762/779c)