APPLIED PSYCHOLOGY: AN INTERNATIONAL REVIEW, 2002, *51* (1), 170–178

Commentary on "An International Study of the Psychometric Properties of the Hofstede Values Survey Module 1994: A Comparison of Individual and Country/Province Level Results"

The article, "An International Study of the Psychometric Properties of the Hofstede Values Survey Module 1994: A Comparison of Individual and Country/Province Level Results", was published in *Applied Psychology: An International Review* in April 2001. The article has elicited a response by Geert Hofstede. The authors have written a rejoinder to this response, and both are published in this section. One of our goals at *Applied Psychology: An International Review* is to facilitate debate and discussion. I hope this section will serve that purpose.

Miriam Erez, Editor

The Pitfalls of Cross-National Survey Research: A Reply to the Article by Spector et al. on the Psychometric Properties of the Hofstede Values Survey Module 1994

Geert Hofstede*

Tilburg University, The Netherlands

In the survey questionnaire of their Collaborative International Study of Managerial Stress in 23 countries, Spector, Cooper, and Sparks (2001) have included the questions of the IRIC Values Survey Module 1994 (VSM94).

* Address for correspondence: Institute for Research on Intercultural Cooperation (IRIC), Tilburg University, P.O. Box 90153, 5000 LE Tilburg, The Netherlands. Email: iric@Kub.nl

In their article they comment on the fact that they found no internal consistency of the VSM94 questions across individuals within their countries. They concluded "that the construct validity of the five VSM scales is suspect, and that they should be used with caution".

Spector et al.'s conclusion is based on a misunderstanding of the nature of cross-cultural survey research, which made them stumble into several methodological pitfalls. A more careful preparation on their side might have prevented this misadventure.

Their lack of preparation is already evident from the "References" list of their article. This list refers to the 1984 abridged edition of my 1980 book *Culture's consequences*. The abridged edition was issued for students; it starts with a warning that this version does not contain the source data, statistical proofs and sections on methodology of the original book. Scholars planning to engage in a worldwide survey and to base part of it on my work should be thorough enough to go to the source and read it.

Their References list also does not refer to my 1991 book *Cultures and organizations: Software of the mind*. This more recent book contains an Appendix entitled "Reading Mental Programs", with methodological suggestions for conducting cross-national research in general and replicating my research in particular. A pity they did not see it. Some of its warnings were repeated in the Manual of the VSM94 to which Spector et al. do refer, but which they evidently did not read either. The 1991 book has appeared in 16 languages and Spector et al. might have made its various language versions available to their collaborators in the various countries surveyed as support in the translation process.

The very essence of cross-national (or cross-aggregate) survey analysis, as exposed in Chapter 1 of both editions of *Culture's consequences*, is that the same two questionnaire items can show entirely different correlations at the individual and at the aggregate level: in the latter case we correlate the mean values of the items for the aggregates. Comparing forests is not the same as comparing trees. A national group is a symbiosis of different individuals and their relationships, just like a forest is a symbiosis of different trees, shrubs, animals, and organisms. This means that tests developed for the individual level are not necessarily relevant for the aggregate level, and vice versa. Applying aggregate-level reasoning at the individual level is known as the "ecological fallacy" and it has mainly been signaled among political scientists. Applying individual-level reasoning at the aggregate level is frequent among naïve psychologists going international, and in my book I have called it the "reverse ecological fallacy".

A consequence of the reverse ecological fallacy is the assumption that applying a reliability formula like *Cronbach alpha across individuals* provides information about reliability across countries. A test for assessing country-level values (like the VSM) is not a test for assessing individual

personality. A study by Bosland in 1985 (referred to in the VSM94 Manual) already showed that the VSM items produce low reliability scores at the individual level. Spector at al.'s calculations at this level (their Table 2) therefore contained no news; they were hammering on an open door but their efforts were wasted for a cross-country study.

The reliability of an instrument designed for comparing country means can only and should only be tested across countries; for the original IBM questions this was done across two surveys four years apart. Since then there have been several large-scale replications on various populations. Their scores are all listed, with their correlations with the original IBM scores, in the new edition of *Culture's consequences* (2001). Spector et al.'s study would have been a candidate for addition to this set if it had met the criteria of properly matching its samples. The VSM94 Manual warns that:

> . . . comparisons of countries or regions should inasmuch as possible be based on samples of respondents who are matched on all criteria other than nationality or region. So, respondents from one country to another should be chosen from the same gender, age, education level, occupation, manager/non-manager status, employer etcetera—they should be matched on any criterion other than nationality that can be expected to affect the answers.

Spector et al. have not taken this warning to heart; as elephants in the cross-cultural china shop they compared country scores including 47 per cent women in the UK, against 1 per cent women in Germany and mean ages of 34.6 years in Hong Kong against 54.2 years in India. Their respondents were convenience samples "*mostly* in administrative/managerial positions" and working for a variety of mostly local companies. They might have tried to correct their scores by controlling for demographic, educational, and occupational differences, but their article shows no evidence that they took this precaution. This means that their samples were insufficiently matched for drawing any cross-national conclusions, and I wonder how valid the outcomes of their main research concern, managerial stress, will be.

Given this poor matching it is not surprising that they did not find good reliabilities at the country level either. It is almost a miracle that in spite of this they still found significant correlations with my original country index scores across the 16 overlapping countries, pointing to cross-national reliability, for Individualism ($r = 0.71$, $P < 0.001$), for Uncertainty Avoidance ($r = 0.59$, $P < 0.01$), and for Power Distance ($r = 0.57$, $P < 0.05$). For Masculinity they found $r = 0.24$ but for measuring this dimension matching genders is essential and this, as we saw, was nonexistent in their case.

In the introduction to their article the authors demand ". . . that, at the very least, the construct validity of these dimensions is based on the bedrock of their internal consistencies". This may be true but as they unwittingly

demonstrated, testing the internal consistency of a cross-cultural instrument is a job for which they were not prepared. They do not try to explain why scales for which as they state "the construct validity is suspect" are so widely used and have produced so many meaningful correlations.

The relationship between reliability and validity goes two ways. The reliability of an instrument is implicitly tested through its proven validity. An unreliable test (one in which the measures contain too much random noise and too little information) cannot produce valid results, so if validity is proven, reliability has to be assumed. Validity is shown by correlating test results with outside criteria expected to correlate according to some kind of theory or logic. The new 2001 edition of *Culture's consequences* describes and analyses all published cross-national studies (up till the late 1990s) in a variety of disciplines for which the results were significantly and meaningfully correlated with scores on the dimensions. These continue validating the dimensions and the reliability of the instrument that led to their identification.

REFERENCES

Bosland, N. (1985). *The (ab)use of the Values Survey Module as a test of personality*. Working Paper 85–1. Maastricht, NL: IRIC.

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

Hofstede, G. (1984). *Culture's consequences: International differences in work-related values* (Abridged edition). Beverly Hills, CA: Sage.

Hofstede, G. (1991). *Cultures and organizations: Software of the mind*. London and New York: McGraw-Hill.

Hofstede, G. (1994). *Values Survey Module 1994: Manual*. Tilburg, NL: IRIC.

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

Spector, P.E., Cooper, C.L., and Sparks, K. (2001). An international study of the psychometric properties of the Hofstede Values Survey Module 1994: A comparison of individual and country/province level results. *Applied Psychology: An International Review*, *50*(2), 269–281.

The Pitfalls of Poor Psychometric Properties: A Rejoinder to Hofstede's Reply to Us

Paul E. Spector*
University of South Florida, USA

Cary L. Cooper
University of Manchester Institute of Science and Technology, UK

Juan I. Sanchez, *Florida International University, USA*; Kate Sparks, *University of Manchester Institute of Science and Technology, UK*; Andre Büssing, *Technical University of München, Germany*; Philip Dewe, *Birkbeck College, England*; Luo Lu, *Kaohsiung Medical University, Taiwan*; Karen Miller, *University of Witwatersrand, South Africa*; Lucio Renault de Moraes, *Federal University of Minas Gerais, Brazil*; Michael O'Driscoll, *University of Waikato, New Zealand*; Milan Pagon, *College of Police and Security Studies, Slovenia*; Horia Pitariu, *Babes-Bolyai University, Romania*; Steven Poelmans, *IESE Business School, University of Navarra, Spain*; Phani Radhakrishnan, *University of Toronto at Scarborough, Canada*; Jesus Salgado, *University of Santiago, Spain*; Oi Ling Siu, *Lingnan University, Hong Kong*; Jean Benjamin Stora, *Hautes Etudes Commerciales Groupe, France*; Peter Vlerick, *University of Ghent, Belgium*; Mina Westman, *Tel Aviv University, Israel*; Maria Widerszal-Bazyl, *Central Institute for Labour Protection, Poland*; Paul Wong, *Trinity Western University, Canada*

Hofstede (this issue) has taken exception to our conclusions (Spector, Cooper, Sparks, et al., 2001) that the poor internal consistency reliabilities of his Values Survey Module 1994 (VSM94) should be cause for concern. He suggests that (1) individual level psychometric properties for a scale are irrelevant when one uses it to make inferences to a group (or country) level, (2) that our samples were poorly matched on demographics (and were not very representative) and this prevented our finding internal consistency at the country (ecological) level, and (3) the VSM94 must be reliable because there is so much evidence for validity. We will address each of these points, explaining our case for concern.

* Address for correspondence: Department of Psychology, University of South Florida, Tampa, FL 33620, USA. Email: spector@chuma.cas.usf.edu

INDIVIDUAL LEVEL PSYCHOMETRIC PROPERTIES DON'T MATTER

Hofstede argues that the VSM94 was designed to assess values at the aggregate (country) and not the individual person level. Therefore, the psychometric properties at the individual level are irrelevant.

We first wish to point out that we reported the internal consistencies at the aggregate level, and results were not appreciably different from those at the individual level. Furthermore, Hofstede is absolutely correct that one must be careful about the ecological fallacy in mixing levels when drawing inferences. For example, if power distance correlates with mean personal income at the country level, one cannot assume that among individuals within a country, the same two variables will be related. However, we did not argue that if the scale lacks internal consistency at the individual level it will lack it at the country level. Statistically it is possible that the items don't relate at one level but they do at another. However, our results found internal consistency problems at both levels, which we believe is cause for concern.

A second issue concerns the real nature of what the VSM94 assesses—individual values or collective values. Morgeson and Hofmann (1999) noted that when individuals are used as informants to provide information about collectives, items should be framed at the target level. In other words, to assess country level values, questions should ask individual participants, not about their own values, but about the values of people within their country. Klein, Dansereau, and Hall (1994) provided a similar perspective and noted that asking respondents to report on their own unique experience is appropriate if one assumes they are independent of the larger collective. Klein, Conn, Smith, and Sorra (2001) provided empirical support for this with a study in which referent (self versus group) affected the extent to which participants agreed with one another in their ratings about work. The researchers argued that the individual referent focused attention on the person's own feelings and provides a less accurate measure of the group level construct.

A close inspection of the VSM94 makes it clear that the focus is on the individual's values, as the instructions state, "This section of the questionnaire is concerned with your values in life and what is important to you." The first section asks the respondent to indicate how important each item is to him or her. Another section asks the individual to indicate agreement with several statements. Given the individual focus of the items, we are unconvinced by the arguments that individual level internal consistency can be ignored if one uses the VSM94 (or many other scales for that matter) at the country level.

As long as one is using a scale that reflects an individual level construct (as opposed to a group level construct), the items should intercorrelate with one another at the individual level. If they don't the scale cannot be said

to assess a single construct, and the aggregation cannot be said to reflect a single culture or country level difference. Either the items assess different constructs, or these values comprise two or more unrelated components, a situation found with global Type A measures that have been abandoned by most researchers in favor of component measures (Edwards, Baglioni, & Cooper, 1990), and with measures of cynicism in police officers that were initially thought to be unidimensional but later proved to be multidimensional (e.g. Regoli, Crank, & Rivera, 1990).

POORLY MATCHED DEMOGRAPHICS CAUSED POOR ECOLOGICAL INTERNAL CONSISTENCY

Hofstede correctly notes that our sampling was not ideal. Our original plan to collect representative samples from each country was not totally achieved, and whereas some of us were able to use methods designed to yield fairly representative samples, a few were only able to get data from a small number of organisations. He suggests that this less than perfect approach had two important consequences—it attenuated the aggregate level internal consistencies and it invalidated results based on these datasets.

Hofstede raised an interesting issue that perhaps demographic differences among our samples attenuated internal consistency results. This would be particularly likely for the masculinity scale, because it has been shown that men and women from within the same country differ (Hofstede, 1984). Since some of our samples had a greater proportion of males than others, might this have produced distortions? We tested this idea by statistically controlling for gender, as well as organisational level, which Hofstede also mentioned. We did this by limiting the analysis to only managers, by limiting the analysis to only managers at the middle level or to the senior level, and by analysing for males only and for females only. If Hofstede is correct, this last control should have improved the internal consistency of the masculinity scale in particular, because of gender differences in means (Hofstede, 1984). Unfortunately, these controls had little effect on internal consistencies, and in many cases the coefficient alphas were even worse. For example, the 0.29 alpha for masculinity in our entire sample was reduced slightly to 0.23 for males only, and increased slightly to 0.35 for females only. Uncertainty avoidance which was 0.49 for the entire scale reduced to −0.09 for both males and females separately. Results were not much different when we controlled for gender and level together. Consequently, demographic differences among our samples do not explain the lack of internal consistency at the aggregate level.

The other issue is whether our sampling procedure is inferior to that used by Hofstede and some other researchers who gathered all data from a

single multinational organisation, and whether this essentially invalidated our results. Clearly the use of a single organisation provides the benefits of controlling for many organisational factors. However, it carries the liability of limited generalisability. Individuals who worked for a single company like IBM where Hofstede did his pioneering work are not necessarily good representatives of their native countries. Our plan was to achieve samples that were more representative of the native populations in each country by sampling from a variety of companies, and preferably from native organisations rather than American multinationals, and in almost all cases this latter goal was achieved. The correspondence between our results and Hofstede's for some of these scales is quite remarkable, and argues against our having samples that cannot generalise.

THE VSM94 MUST BE RELIABLE BECAUSE IT IS VALID

Hofstede is correct that validity assumes a certain level of reliability. The fact that the VSM94 significantly relates to other variables presupposes it can consistently measure something, but this does not mean it has internal consistency. If we repeatedly assess a person's weight (in grams) and telephone area code and then sum the result, we will likely get a reliable total score, but this doesn't presuppose both components measure the same thing and that the combination is meaningful. It is possible that the combination of weight and area code would correlate with height since one of the components (weight) does, but this does not provide construct validity to the combination, which is rather meaningless in our example. Our concern with the VSM94 is not that it cannot predict other variables, and not that it does not provide a relatively stable (test-retest reliability) measure, but that it lacks internal consistency. This leads to the conclusion that the scales (except for individualism in some samples and long-term orientation) do not assess a single homogeneous construct. As we stated in our original paper, we wonder what the combination might be. It is possible that these values are multidimensional and the individual items assess different subcomponents. It is possible that some items tap the hypothesised values and others additional constructs. More scale development work would seem in order to further develop these scales, and this might contribute to a richer understanding of these values. If we are concerned with the values of people and how they vary between countries, this should begin at the individual level, and then move to the aggregate. If our concern is with value constructs that are meaningful only at the country level, then the assessment procedure should focus on that level, perhaps by asking individuals to report on the values of people in general or on other methods that more directly reflect country level phenomena rather than individual.

REFERENCES

Edwards, J.R., Baglioni, A.J., Jr., & Cooper, C.L. (1990). Examining the relationships among self-report measures of the Type A behavior pattern: The effects of dimensionality, measurement error, and differences in underlying constructs. *Journal of Applied Psychology*, *75*, 440–454.

Hofstede, G. (1984). *Culture's consequences* (Abridged edition). Thousand Oaks, CA: Sage.

Hofstede, G. (this issue). The pitfalls of cross-national survey research: A reply to the article by Spector et al. on the Psychometric Properties of the Hofstede Values Survey Module 1994. *Applied Psychology: An International Review*.

Klein, K.J., Conn, A.B., Smith, D.B., & Sorra, J.S. (2001). Is everyone in agreement? An exploration of within-group agreement in employee perceptions of the work environment. *Journal of Applied Psychology*, *86*, 3–16.

Klein, K.J., Dansereau, F., & Hall, R.J. (1994). Levels issues in theory development, data collection, and analysis. *Academy of Management Review*, *19*, 195–229.

Morgeson, F.P., & Hofmann, D.A. (1999). The structure and function of collective constructs: Implications for multilevel research and theory development. *Academy of Management Review*, *24*, 249–265.

Regoli, B., Crank, J.P., & Rivera, G.F. (1990). The construction and implementation of an alternative measure of police cynicism. *Criminal Justice and Behavior*, *17*, 395–409.

Spector, P.E., Cooper, C.L., Sparks, K., Bernin, P., Büssing, A., Dewe, P., Lu, L., Miller, K., Renault de Moraes, L., O'Driscoll, M., Pagon, M., Pitariu, H., Poelmans, S., Radhakrishnan, P., Russinova, V., Salamatov, V., Salgado, J., Sanchez, J.I., Shima, S., Siu, O.L., Stora, J.B., Teichmann, M., Theorell, T., Vlerick, P., Westman, M., Widerszal-Bazyl, M., Wong, P., & Yu, S. (2001). An international study of the psychometric properties of the Hofstede Values Survey Module 1994: A comparison of individual and country/province level results. *Applied Psychology: An International Review*, *50*, 269–281.