University of Wollongong
Research Online

Faculty of Commerce - Papers (Archive)

Faculty of Business

2007

Assessing analytical robustness in cross-cultural comparisons

Sara Dolnicar
University of Wollongong, sarad@uow.edu.au

Bettina Grun
Vienna University of Technology, Austria, bettina@uow.edu.au

Publication Details

This article was originally published as: Dolnicar, S & Grun, B, Assessing analytical robustness in cross-cultural comparisons, International Journal of Culture, Tourism and Hospitality Research, 2007, 1(2), 140-160.

Research Online is the open access institutional repository for the University of Wollongong. For further information contact the UOW Library: research-pubs@uow.edu.au

Assessing analytical robustness in cross-cultural comparisons

Abstract
Response styles can distort survey findings. Culture-specific response styles (CSRS) are particularly problematic for researchers using multicultural samples because the resulting data contamination can lead to inaccurate conclusions about the research question under study. This article critically reviews past recommendations to correct for cultural biases in responses, and proposes a framework that enables the researcher to assess the robustness of empirical findings from CSRS. This approach also avoids the disadvantages of ignoring the problem and interpreting spurious results or choosing one single correction technique that potentially introduces new kinds of data contamination.

Keywords
cross-cultural research, response styles, robustness, standardization

Disciplines
Business | Social and Behavioral Sciences

Publication Details
This article was originally published as: Dolnicar, S & Grun, B, Assessing analytical robustness in cross-cultural comparisons, International Journal of Culture, Tourism and Hospitality Research, 2007, 1(2), 140-160.

This journal article is available at Research Online: http://ro.uow.edu.au/commpapers/326

Assessing Analytical Robustness in Cross-Cultural Comparisons

Sara Dolnicar, University of Wollongong, Australia

Bettina Grün, Vienna University of Technology, Austria

Submission: November 2006

Revision: January 2007

Acceptance: February 2007

Authors' names appear in alphabetical order. Send correspondence to Sara Dolnicar, School of Management & Marketing and Marketing Research Innovation Centre (mric), University of Wollongong, NSW 2522, Australia, Telephone: (61 2) 4221 3862, Fax: (61 2) 4221 4154 (email: sara_dolnicar@uow.edu.au). Bettina Grün, Department of Statistics and Probability Theory, Vienna University of Technology, Wiedner Hauptstrasse 8-10, A-1040 Vienna, Austria, Telephone: (43 1) 58801 10716, Fax: (43 1) 58801 10798, bettina.gruen@ci.tuwien.ac.at.

Cross-Cultural Comparisons in Assessing Analytical Robustness

Abstract

Response styles can distort survey findings. Culture-specific response styles (CSRS) are particularly problematic for researchers using multicultural samples because the resulting data contamination can lead to inaccurate conclusions about the research question under study. This article critically reviews past recommendations to correct for cultural biases in responses, and proposes a framework that enables the researcher to assess the robustness of empirical findings from CSRS. This approach also avoids the disadvantages of ignoring the problem and interpreting spurious results or choosing one single correction technique that potentially introduces new kinds of data contamination.

Keywords: cross-cultural research, response styles, robustness, standardization.

Cross-Cultural Comparisons in Assessing Analytical Robustness

Introduction

The existence of response styles is a well-known and much-studied phenomenon in various disciplines within the empirical social sciences (Baumgartner and Steenkamp 2001; Bhalla and Lin 1987; Hui and Triandis 1989; Paunonen and Ashton 1998; Sekaran 1983). Different respondents have different ways of using the answer formats that researchers offer them, independent of the content. Paulhus (1991: 17) claims that this behavior results in a response bias that has "a systematic tendency to respond to a range of questionnaire items on some basis other than the specific item content (i.e., what the items were designed to measure)." He also claims, "To the extent that an individual displays the bias consistently across time and situations, the bias is said to be a response style."

The tendency to make more use of extreme answer options is one possible response style. In the case of the much-used multi-category answer format with five answer options, respondents with such a tendency tick the first and the fifth option more frequently than the others. For instance, when asked to rate satisfaction with the food in a hotel on a scale from "highly satisfied" to "highly dissatisfied," a respondent displaying an extreme response style or ERS (Baumgartner and Steenkamp 2001) would favor the two end points independently of the content.

Other respondents feel more comfortable avoiding extreme answers and make more use of the middle answer categories, such as "mildly satisfied" in the above example (Roster, Rogers

and Albaum 2003). Even if a respondent with a mild response style and a respondent with an ERS feel the same level of satisfaction, their answers on the multi-category answer format are likely to differ, wrongly leading researchers to conclude that their satisfaction levels differ (Kozak, Bigne and Andreu 2003). In other words, interpretation of response scales is subjective.

Studies repeatedly show that the cultural background of respondents has a systematic effect on their response style. Respondents from different cultural backgrounds tend to use survey answer formats in different ways (see, for instance, Hui and Triandis 1989; the section on culture-specific response styles below includes a review). This effect does not influence the results of empirical studies within one discrete cultural area. However, if the sample consists of individuals from varied cultural backgrounds with significantly different styles of using survey formats, the results derived from this data set could be distorted. Such data sets will henceforth be referred to as “multicultural data sets”.

This study provides a replication-based approach to assessing the danger of misinterpretation due to CSRS for each of the items used to compare respondents from different cultural backgrounds. This article achieves this by (a) reviewing prior work studying CSRS, (b) critically reviewing proposed techniques to correct for CSRS, (c) proposing a robustness-based approach to assessing cross-cultural findings, and (d) illustrating both possible misinterpretations if data are not corrected or are inappropriately corrected.

The approach here differs distinctly from prior propositions by building on robustness analysis of cross-cultural research findings across various conditions in correcting for bias. The main advantage of this approach (as opposed to prior recommendations) is to minimize the risk of either misinterpreting raw data contaminated by response styles, or choosing an incorrect transformation of original answers and drawing the wrong conclusions from the corrected data.

Hence the solution proposed in this study takes the perspective of robustness, and assesses the degree to which findings actually depend on the potential sources of contamination.

If substantial differences between cultures derive from analyses based on corrected and uncorrected data, the researcher can have more confidence when reporting such findings, because the result cannot be an artifact resulting from the expected response styles or data transformations undertaken to correct the data. However, if the analysis of uncorrected data leads to findings entirely different from the analyses of corrected data sets, the researcher should interpret such results with care.

Prior Work

The central problem of cross-cultural studies is that if variant response styles are present in data sets, the researcher can no longer interpret differences in group mean values (Chun, Campell, and Yoo 1974). This represents a major problem where, for instance, the central research aim is to test cultural differences, and when using mean-based tests (such as t-tests and F-tests).

The literature review draws on sources from the fields of psychology (for instance, Arce-Ferrer and Ketterer 2003), sociology (for instance, Watson 1992), and marketing research (for instance, Greenleaf 1992a). Prior work falls into four classifications: (1) articles that discuss the kinds of methodological problems that may distort cross-cultural research findings in general, and response styles specifically; (2) empirical studies that investigate the existence and nature of response styles in cross-cultural studies; (3) publications in which techniques propose to detect whether or not data sets are contaminated by response styles; and (4) work that proposes correction techniques.

Methodological problems in cross-cultural research

The literature on methodological and theoretical problems of cross-cultural research (research stream 1) shows a vast number of potential pitfalls awaiting cross-cultural research. A comprehensive review by Sekaran (1983) lists nine different areas capable of affecting the validity of findings, and argues that researchers must ensure equivalence at different levels, ranging from conceptual/functional over construct operationalization, to item and scalar equivalence. Sekaran's review illustrates the many potential pitfalls in empirical cross-cultural research, and thus provides a useful reference point for the present study, which is interested in one particular aspect among those discussed by Sekaran: *measurement bias*. Drasgow (1987: 19) claims that measurement bias occurs if individuals with equal standing on the trait measured by the test (but who are sampled from different subpopulations) fail to have equal expected observed test scores.

Bhalla and Lin's (1987) article on methodological requirements of cross-cultural studies discusses measurement bias in a section called "Scalar Equivalence," where they describe how "Cultures differ in their response set characteristics, such as social desirability, acquiescence, and evasiveness, which influence response scores" (Bhalla and Lin, 1987: 278). Smith and Reynolds' (2002) review discusses the aspect of measurement bias in more detail, and differentiates between two sources of bias: response sets and response styles. *Response sets* imply that respondents wish to paint a certain picture of themselves, such as the Japanese not wanting to boast about their achievements. *Response styles* are systematic differences in responses that result from the format of the questions presented to respondents. Smith and Reynolds conclude that "Failure… to detect differences in cross-national response bias will… affect data comparability,

may invalidate the research results and could therefore lead to incorrect inferences about attitudes and behaviors across national groups" (2002: 450). The present study focuses on *response styles*.

Empirical Evidence for Response Styles in Cross-Cultural Research

Smith and Reynolds (2002: 463) also point out that "There is already evidence to suggest that, at a minimum, extreme and neutral response styles differ cross-culturally." Research stream 2 offers such evidence, which consists of studies empirically investigating the existence and nature of response bias between cultures. Early work in this area investigated whether answers from black and white respondents differed systematically, concluding that black respondents tended to use the extreme points of the scale (for instance, the "strongly agree" or "strongly disagree" options on a Likert scale) more frequently than white respondents (Bachman and O'Malley 1984). Several other empirical studies detect systematic differences in response styles between respondents from Asian and Western cultures. Chen et al. (1995) conclude that individuals from collectivist cultures tend to avoid extremes; while Shiomi and Loo (1999) and Si and Cullen (1998) find that respondents from Asian countries use middle categories more than do Western respondents; Das and Dutta (1969) identify a moderate response style among Indian respondents; and Chun et al. (1974) detect a stronger ERS (the tendency to use the extreme values as exemplified above) among American students compared to Korean students.

The present study contradicts this general tendency. Roster, Rogers, and Albaum (2003) find that US and Filipino respondents are more likely to respond to attitudinal scales with extreme answers compared to Chinese or Irish respondents. Another set of response style studies focuses on Hispanic respondents, and consistently finds that these respondents tend to an ERS (Hui and Triandis 1989; Marin, Gamba, and Marin 1992). Triandis and Triandis (1962)

demonstrate the same effect for Greek students in comparison to American students. Very few reports show that no differences exist between cultures: Cheung and Rensvold (2000) and Yates et al. (1997) find no difference in response styles between American and Taiwanese students.

Such systematic differences are not random occurrences; cultural influences lead to differences in how people respond to a question. For instance, Hui and Triandis (1989: 298) explain the differences between Hispanic and non-Hispanic respondents thus: "In cultures around the Mediterranean, by contrast an extreme response style is used because people consider such a response sincere. To use the middle of the scale would be considered trying to hide one's feelings, which is normatively disapproved." Hui and Triandis also suggest that modesty and caution drive Asian cultures to make less use of extreme points on an answer scale.

Detecting Response Styles

Research stream 3 consists of studies that propose detection methods of response styles. Most researchers derive measures for specific kinds of response styles to quantify the contamination of the data. For instance, Chun et al. (1974) use individual standard deviations to detect ERS. Greenleaf (1992b) and van Herk et al. (2004) propose further methods to detect and measure ERS, and use the proportion of extreme responses. Johnson et al. (2005) use the number of items with extreme responses. Hui and Triandis (1985) use the individual means to detect *acquiescence response style* (ARS, the tendency to agree independently of the content of the question), standard deviation for *response range* (RR, a measure to assess how wide is the range of answer options used), and the number of times the respondent ticked the end points, to assess the level of ERS. Baumgartner and Steenkamp (2001) give a detailed discussion of seven different response-style types, and the determination of corresponding measures that can be used

to assess the different response style types. Cheung and Rensvold (2000) take a different approach, and use multi-group confirmatory factor analysis to check for contamination of the data with ARS and ERS. They propose a stepwise procedure, in which they first check for form invariance between two cultures, where they test if the same items are associated with each construct. They test for ERS by checking factorial invariance, that is, for equality of the factor loadings. They then test for ARS, by checking intercept invariance, that is, for the equality of the item values where the latent variable is equal to zero. If invariance is confirmed in each test, the differences in latent means indicate substantive differences between cultures. If variance is not confirmed, ERS and ARS could be present in the data. Furthermore, checking the factorial invariance fails to detect uniform ERS differences (that is, the same bias exists with respect to all items associated with a construct).

Correcting for Response Styles

Work classified as research stream 4 contains techniques proposed to correct for response styles, once identified. Standardization is the most commonly used correction technique. If the researcher believes that CSRS are present in the data and that they might distort research findings, they will standardize the data. Standardization is based on the assumption that response styles are constant over time, and that the differences in aggregated answer patterns actually reflect differences in response styles and not content-related differences. This assumption might be justified by checking if the differences are consistent over different, unrelated constructs, which reduces the possibility that the differences are content related, because for a single construct, overall differences may exist in attitude towards the construct under study between the respondents. While standardization leads to the removal of response bias, the danger associated with standardization is that it also eliminates content-related differences. Fischer (2004) reviews

standardization methods commonly used to adjust for response styles in cross-cultural research and provides a classification of the different methods. He distinguishes between different forms and different units of adjustment, and an overview of these is given in Table 1.

Table 1 here

The different units in Table 1 are *within-subject*, *within-group*, *within-culture*, and *double*. If the unit of adjustment is the subject, standardization occurs using all variables for each individual. This compares to within-group standardization, which uses all individuals for each variable. Within-culture standardization uses all variables and individuals in each culture. If a study includes double standardization, the research combines within-subject standardization with standardization within group for each culture.

The forms of adjustments in Table 1 differ with respect to the measure they use for correction: means, dispersion indices, both means and dispersion indices, or covariates. Means help correct for a possible ARS in the data, while dispersion indices (commonly standard deviations) can account for ERS. However, subtracting the mean leads to ipsatized scores known to reflect only intra-unit (relative) differences (Chan 2003), and furthermore, forces the row and column sums in the correlation matrix to zero. This affects all correlation-based analysis techniques, such as factor analysis. Hence the generally recommended use of ipsatization is only for scales with low inter-item correlations (Bartram 1996). Therefore, within-subject standardization using means is based on the assumption that no content-related difference exists between respondents. If any content-related differences exist, these would be lost in the standardization process.

Within-group standardization means that the transformation is made across variables, thus ensuring that each variable has the same overall mean and/or the same variance. This step is

important for multivariate analyses, and if aggregated scores are determined. Within-group standardization assumes that the overall average score and/or variance are comparable over variables. Researchers undertake within-culture standardization if they have assumed that the response styles differ between cultures, but are equal within culture. Under this assumption, within-culture standardization might be preferable to within-subject standardization. The estimates for the different forms of adjustments are more reliable because they take into account all respondents in a cultural group. However, the assumption of homogeneity in response styles within culture might be questionable, given that other socio-demographic characteristics are associated with response styles.

Leung and Bond (1989) propose double standardization for individual analysis. Within-subject standardization removes individual response style, and within-culture standardization for each item removes differences between the average responses of the individuals of each culture. These differences in average responses are also called differences in positioning.

If a researcher assumes that CSRS might distort the results of the analysis and wants to correct the data to account for this contamination, they cannot randomly choose a correction technique. The choice of correction technique should consider what assumptions each of the different standardization methods makes, in order to identify the one most appropriate for the research problem and the data conditions faced. If such identification were possible, the researcher could transform the data using the appropriate method, and the corrected data set becomes the bases for all analyses.

Unfortunately, and in most cases, deciding which correction technique is most appropriate is not simple. Often no criteria are available to assess whether the assumptions each of the possible transformations makes are appropriate for the data and the problem. The researcher

essentially has two choices. They can either ignore the possibility of CSRS contamination, arguing that any transformation would lead to a different form of contamination of the data anyway (for instance, loss of content-related differences); or they can choose the correction technique that appears to be most suitable, and accept that the assumptions made by this correction technique may be inappropriate, and thus cause unwanted new distortion effects on the data.

A Robustness-Based Approach

Neither of the presently available solutions for dealing with CSRS is entirely satisfactory for a researcher. This problem is due to the high levels of insecurity about how best to trade off original data contamination with data contaminations potentially introduced by inappropriate correction. Essentially, the problem faced is one where the true values of the answers given by respondents cannot be retrieved any more, and are therefore unknown. Consequently, the challenge is to assess whether differences postulated between respondents from different cultures are true, or merely artifacts of response styles or response style corrections.

Motivation

The replication approach is successful in dealing with precisely this situation, where true values are unknown and an assessment of the reliability of conclusions is needed. For instance, in market segmentation studies, the true segment membership is unknown *a priori* and not directly observable. The researcher can segment a market in myriad ways, none of whose memberships may be the best or most appropriate. One solution proposed for this problem is replication. Extensive repetition and comparison of results can extract the most reliable results —or the most reliable findings. Furthermore, replication allows distinguishing between stable segments that

represent “natural” clusters and unstable segments that represent “artificial” clusters (Kruskal 1977).

The most common problem of this nature is that true empirical values for any problem can in general only be estimated. This is the basis of inferential statistics: the significance level informs the researcher how likely any particular result is true and not a random result. While finding what the true values of respondents are is not a trivial problem, powerful ways exist to assist the empirical researcher to assess the dangers of misinterpretation and interpret findings only that have a fairly low probability of being wrong. For example, using a typical significance level of 95 percent, an empirical researcher takes a five percent risk of claiming a finding that is not true. Using a similar approach for the problem of CSRS offers a promising avenue for dealing with the dangers of potential misinterpretation based on cultural differences of respondents, and the fact that true views of respondents are typically not known.

The prior discussion clarifies that each data set requires customized assessment, and that a general deterministic solution cannot offer the optimal way of dealing with potential CSRS-related misinterpretation of results. Even if determining correction factors for certain nationalities were possible, these would have to be different for different constructs under study. For instance, questions about satisfaction are more likely to trigger different response style effects compared to questions about vacation activities undertaken, with respondents likely to perceive the latter as personal or even confidential in nature. Even if determining a set of bias values for a range of commonly studied constructs were achievable, CSRS are dynamic phenomena, and likely to change over time. Optimally, researchers should develop a technique that allows them to assess for their data sets the extent to which CSRS biases results.

The problem this study targets does not encompass all possible mistakes a researcher might make in the context of a cross-cultural study. Specifically, problems arising from badly operationalized constructs, or the lack of structural equivalence of the construct across cultural backgrounds, cannot be solved with the proposed approach. Both these problems essentially make the responses unusable, because each respondent (or each culture) may have entirely different perceptions about what the question is about. Social science research typically uses many constructs that are ill-defined, as Kampen and Swyngedouw (2000) discuss in detail. Many other studies discuss the problem of structural equivalence of constructs across respondents from different cultural backgrounds, emphasizing the need for extensive exploratory work before questionnaire development (Kozak, Bigne and Andreu 2003; Sekaran 1983; Bhalla and Lin 1987). Issues of operationalization and structural equivalence should be addressed at an earlier stage of the research project. The problem dealt with in the present study, CSRS, occurs during the quantitative phase, and occurs even if the construct under study has perfect structural equivalence.

Classification of Variables with Respect to Robustness of Findings

The robustness of results is useful for assessing the reliability of findings because whether raw or corrected answers are closer to the truth of the matter investigated is not known, as discussed above. Robustness in this context means independence of CSRS and corrections for CSRS. Researchers can consider robust a result from an empirical study that includes respondents from different cultural backgrounds if the original answers, as well as various suitable corrections for response styles, lead to the same conclusion. In the worst case, the original values, and each of the alternative ways of correcting for CSRS, lead to different results, indicating a very low level of robustness of findings.

Figure 1 depicts the alternative outcomes of this scenario. The term “corrected” indicates that a wide variety of different corrections is possible. For simplicity of illustration, Figure 1 only compares the results of one correction technique to the results for the raw data. However, in general, several correction techniques might be assumed suitable in addition to the raw data.

Figure 1 also distinguishes between the analysis of corrected (vertical dimension) answers and the analysis of raw (horizontal dimension) answers. In each case the research question (whether differences exist between respondents from different cultural backgrounds) can be tested for each of the variables in the data set. The test result can be significant (indicating that a difference exists) or not significant (indicating that no difference exists). The combinations of test results based on corrected and raw data can be used to assess the danger of misinterpretation of multicultural data sets in general, as well as the classification of each variable into a high- or low-risk category.

Figure 1 depicts a case where only two results are compared. However, the proposed approach would typically include a set of corrected data sets derived by applying all theoretically suitable transformations. Different results can occur for different variables because each respondent’s answer consists of a *true value* component and a *response style* component. While the response style component is assumed constant across all answers by an individual respondent (they may always tend to use extreme answer options), the true values are assumed to be different (they may be more interested in relaxing vacations than in action-packed or culture-oriented vacations). The relation to true values and response styles across a culture will determine whether the true and corrected data will lead to the same or different conclusions.

The top right-hand corner in Figure 1 represents high-risk items (HRIs) because the two tests do not lead to the same conclusions. The test based on raw data leads to the conclusion that

a difference exists (for instance, that French tourists are significantly more interested in city packages than German tourists, and therefore may be the better target groups for such offers). The test based on corrected data indicates that no difference is evident (that French and German tourists are equally interested in city packages). The bottom left-hand quadrant illustrates the opposite situation: raw data leads to insignificant differences, whereas corrected data leads to significant differences.

Both these situations can lead to misinterpretations. If only the corrected or uncorrected data is interpreted, differences are or are not claimed, which may well be artifacts of response styles or response style corrections only. Items of this nature are therefore referred to as high risk. Two possible ways of dealing with such high-risk items exist: one is to omit reporting on the findings on these items, possibly a data-dumping exercise that may not be possible, given that clear answers are needed. However, this is quite a usual procedure, particularly in psychological studies. Items for which freedom from structural or scalar inequivalence cannot be established are often omitted to "purify" the scale (see, for instance, Cheung and Rensvold 2000; Huang, Church and Katigbak 1997).

Figure 1 here

Two other kinds of items are not endangered by potential misinterpretations based on response styles. If significance tests based on the raw and corrected data lead to the same result, then either a difference between the two cultures or countries of origin exists, or does not. These results can be reported with a reasonable amount of confidence, given that they are based both on uncorrected and corrected variable values.

The bottom right-hand corner of the figure depicts the case where both significance tests indicate no difference between the cultures or countries of origin; variables of this nature can be

referred to as very low-risk items (vLRIs). Alternatively, if both significance tests indicate differences in responses (the top left-hand quadrant of the framework), the only possible misinterpretation would be that one test states that respondents from one country have higher values, and the other test indicates that respondents from the other country have higher values — a rather unlikely outcome, but one that should be checked. These variables are therefore referred to as low-risk items (LRIs). An asymmetry exists in the evaluation of the items introduced where all significance tests agree. In order to be able to draw unambiguous conclusions, the researcher must additionally check that the differences for the LRIs have the same sign.

Outline of the Procedure

A wide variety of different recommendations exists as to how to assess and correct for CSRS. These include many viable ways of dealing with the problem of CSRS contamination in the empirical analysis of multicultural data which are based on different assumptions. However, choosing any of the available approaches has at least one major drawback: the researcher assumes — without knowing the true nature of contamination by CSRS — that the chosen transformation leads to values closer to the true views of respondents than the raw data. This may or may not be the case. A chosen transformation may well lead to values that are further away from the true views of consumers. Typically, no way exists to determine which of the possible scenarios is the case.

Any attempt to find the transformation that recovers the true values is, by its very nature, a process that cannot be firmly validated. Hence the proposed procedure aims not at prescribing a way to transform the data, but rather, determines the robustness of findings to CSRS. Assessing robustness uses the original values and a set of possible transformations, and the researcher

undertakes the analyses required to answer the research questions for all data sets. They will classify as robust results that lead to the same conclusions under all data conditions. Therefore, CSRS robustness is an indicator that the rejection of a hypothesis was likely correct, despite the contamination with CSRS. This test functions as a guide for the researcher as to which results they can reliably report, and which they should interpret with care because the effects of CSRS could lead to misinterpretations.

The procedural proposal consists of four steps. However, these steps do not include other necessary steps of fieldwork design for cross-cultural studies, such as the assessment of structural equivalence of the construct under study. The starting point is any multicultural data set where the researcher must assume that CSRS is presentable.

Step 1: Selection of a set of correction techniques appropriate for the problem and data at hand.

Step 2: Correction of raw values according to all chosen correction techniques.

Step 3: Testing of cross-cultural differences based on the raw data and all transformed data sets.

Step 4: Computing of CSRS robustness indices — the proportion of identical analyses results to deviating results across all pairs of data sets. For each analysis, one value is derived.

Empirical Illustration

For the illustration, this section uses a data set from the tourism area where multicultural samples and comparisons naturally arise and are thus frequently encountered. The same data set

applies for the purpose of this illustration to: (1) demonstrate the potential misinterpretations that can result either from ignoring the existence of response styles or choosing inappropriate correction techniques, as discussed above; and (2) demonstrate how the proposed robustness-based approach can help researchers reduce the level of uncertainty in interpreting results from cross-cultural analyses.

Imagine the National Tourism Organization (NTO) of Austria would like to allocate their advertising budget to one country of origin, because the budget is insufficient for campaigns in several countries. The NTO therefore undertakes a study to compare countries of origin with respect to their travel motivations, in order to determine which country's travellers are most interested in certain aspects, for instance, culture, health, and beauty, or an unspoiled environment.

Description of the Data Set

The data set used resulted from the national guest survey conducted in the summer season of 1994 by the NTO, the so-called *Österreich Werbung*. While the original data set includes respondents from 14 different countries, the selection was only of a sub-sample of 1,351 respondents from four areas of origin, in order to keep the illustration simple (France: n = 312, Italy: n = 340, USA: n = 246, Vienna: n = 453). This illustration also assumes that all equivalence criteria that need to be assured during the survey development and data collection phase have been evaluated and found satisfactory. However, as previously noted, these equivalences are not the focus of this paper. Also, assessment of them *ex-post*, using statistical techniques, might not be possible. For example, Cheung and Rensvold (2000) state: "one type of form noninvariance, known as construct bias cannot be detected statistically."

The survey includes a set of 21 questions on vacation motivations, including asking respondents to state to what extent each of the listed aspects was a motivating factor for them on the vacation. The four-point ordinal answer format used the labels “not at all,” “a little bit,” “to some extent,” and “exactly.” The questions covered all relevant, but different, aspects of vacations which might influence the destination choice. One example is: “On holiday I want to exert myself physically and play sports.” (The full set of motivation statements is available from the authors.) Raw scores are determined by assigning equidistant values from 0 to 1 to the four levels of agreement.

Illustration of Possible Misinterpretations

The first step is to illustrate which misinterpretations could be made based on this data set if the possible existence of response styles is ignored. For this purpose, two independent sets of computations are undertaken, one based on the raw, uncorrected data (thus ignoring the possible contamination of the data by response styles) and one with corrected data. The data were corrected using within-subject standardization. This is achieved by subtracting the individual mean and dividing through the individual standard deviation. This particular correction technique was chosen because of its high popularity, and has been recommended in the past for removing both ARS and ERS. Using the corrected data as the basis of analysis assumes that either ARS or ERS contaminate the data, which eliminates response styles of these kinds; however, doing so also enables elimination of some of the actual content of respondents’ answers.

Table 2 shows a comparison of the mean answers of the raw and standardized scores for each region, together with the mean answers of the total population. For the raw data, the mean values are between 0 and 1, where 0 indicates no motivation at all, and 1 absolute agreement with

the statement. The corrected data can take negative and positive values, indicating if the motivation for an item is above or below the average motivation. Child care, for example, is of very low importance for all areas of origin, as indicated by having the smallest mean value.

The raw data attracts the conclusion that on average, the respondents have a motivation between not at all and a little, because the observed values are comparable to the original scale. In contrast, for the corrected data, the only conclusion is that this item is of least importance for each area of origin, and deviates most from the average motivation.

Table 2 here

The cross-cultural comparison of the results of the national guest survey data shown in Table 2 reveals that for the French respondents, the level of agreement with the motivation statements is higher for the raw than for the corrected data, for those motivations that are more important for the French. For those motivations not so important for the French respondents, the levels of agreement are lower when raw, uncorrected data is used. The opposite is true for the Italian respondents. No obvious differences between the raw and corrected data are observable for the American respondents. For the Viennese respondents, the appreciation for “atmosphere” might be underestimated when comparisons are computed on the basis of the raw scores. The importance of “sports” emerges as much higher for the Viennese respondents than for the average population in the corrected data; but this is not the case in the raw data.

A superficial assessment only is the basis for the above interpretation. When respondents’ answers to all 21 travel motives are tested for differences between each of the countries and the overall mean value using a t-test, 60 percent of the differences (50 out of 84) are significant for the raw scores and 57 percent for the standardized scores (48 out of 84), at a significance level of 0.05. A cross-tabulation of the test results shows that in 24 of the possible differences between

countries (27 percent), neither raw nor standardized data renders significant results. Thirty-eight differences (45 percent) are significant when tested on the basis of both raw and standardized data. In 10 tests (12 percent) the standardized data renders significant results, while the raw data does not; and the opposite holds for 12 comparisons (14 percent). This leads to an overall proportion of 74:26 with respect to the number of tests that returned the same results (low risk) as opposed to those that returned different results (high risk).

While these differences may seem academic, based on the above discussion, the practical relevance becomes very clear, considering that national tourism organizations typically use these kinds of nation profiles to develop communication strategies to attract tourists from certain countries. If market structure analysis used the raw data as a basis, the test would indicate that French tourists are significantly more interested in unspoiled nature than other tourists. The national tourist organization might use this information to develop a large, expensive "unspoiled nature" advertising campaign. However, the standardized data shows no significance, indicating that the French are no more or less interested in unspoiled nature than are other tourists. This comparison illustrates that the contamination of empirical data by CSRS is a serious problem that can lead to misinterpretation of results, because the conclusions drawn could depend on the chosen correction technique — or possibly the incorrect decision to ignore possible contamination by response styles and analyze the raw data only.

Application of the Proposed Robustness-Based Approach

Using the same data set, the next illustration demonstrates how the proposed robustness-based approach can help researchers to reduce the uncertainty revealed above, which is inherent in any data-analytic problem where the true values are unknown. The method follows uses the

four steps outlined above: first (Step 1), the researcher chooses a set of correction techniques appropriate for the problem. Because the aim of this analysis is a cross-cultural analysis, where the mean values on the attributes are compared for the different cultures, double standardization removes the very differences that are of interest, making the process inappropriate. Because only univariate comparisons are made, within-group standardization does not influence the results. The set of appropriate correction techniques therefore contains within-subject and within-culture standardization, where correction uses either means and/or standard deviations.

Next (Step 2), the researcher corrects the raw values with respect to these six techniques. Then (Step 3), the researcher tests cross-cultural differences using t-tests for the raw data and all transformed data sets.

Finally (Step 4), the researcher computes the CSRS robustness indices and classifies each variable (motivation item contained in the questionnaire) as LRI, vLRI or HRI. This classification provides a quick insight into how problematic CSRS are for the given dataset, and allows decision making for the subset of items that are classified as LRI or vLRI. A further investigation of the HRIs will be necessary if they either include items central for the strategy to be chosen, or if most items are classified as HRI. In this case, the researcher might subjectively decide also to include HRIs with a very high robustness index, that is, when most datasets indicate a significant difference.

Classification as an LRI indicates that the cross-cultural comparison of this particular item has rendered significant differences across the four regions for all data sets, the raw data and all six data sets containing different appropriate corrections for response styles. The researcher must additionally check that the differences for these items have the same sign for all datasets. If so, then these cross-cultural differences can be interpreted safely as findings from the study.

Classification as a vLRI indicates that none of the analyses based on different underlying data sets led to the conclusion that the four regions differ with respect to this particular travel motivation. Again, the researcher can safely conclude the non-existence of difference in this case.

Classification as an HRI indicates that the tests based on the raw data and/or different corrections did not lead to the same results with respect to whether differences exist between the four regions. Thus cross-cultural differences with respect to these items should be interpreted with care, because response styles or chosen corrections thereof are likely to influence the results. Classification as an HRI allows the researcher to make statements about how many cases the tests for difference between cultures led to significant results, and in how many cases they did not. This information is contained in Table 3 in parentheses. For instance, in the case of the motivational item "rest and relax", the French respondents differed from the other three regions three times, and did not differ four times. This is the worst possible quota, because half the analyses claim differences and half do not. In the case of "creativity," Viennese respondents differed from the remaining regions once, and did not differ six times, which indicates only one disagreeing test result.

Regardless, HRIs present a challenge, because the researcher cannot safely draw any clear conclusions to report in an academic publication or to a client such as the NTO. In such a situation, the researcher has the option to report the discrepancy openly, or to undertake further qualitative research work for the regions under study with respect to the relevant HRIs. The qualitative study would have to try to determine the nature and extent of response styles with respect to the particular items in order to be able to validate the conclusion externally. Such an approach is expensive, and requires much time for further investigation. Therefore, such qualitative work is not reasonable for all items and all regions *a priori*. However, for selected

items that emerge as endangered by misinterpretation, such a follow-up study may well be a viable option.

Table 3 here

Table 3 provides all results for the Austrian NTO illustration. To enable a quick overview of the findings, cells in this table are shaded according to these classifications: LRIs are black, vLRIs are grey and HRIs are white. Optimally, this table would contain no white cells, indicating that conclusions with respect to cross-cultural differences can be safely drawn for all items. The higher the proportion of black cells in the table, the higher the extent of cross-cultural differences across all items and all regions.

In the empirical case depicted in Table 3, 36 motivation items are classified as LRIs (43 percent), and 19 as vLRIs (23 percent). The sign of the co-efficients are checked for all LRIs, and they are consistent over all data sets, which means that all the tests that found differences found the same regions to perceive a particular motivational item as more important. Therefore, the LRIs can be safely interpreted and used as a basis for marketing activities. Twenty-nine motivational items (35 percent) are classified as HRI, and should not be interpreted without explicitly reporting the potential impact of response styles and/or response style corrections on the conclusions, or without conducting a qualitative follow-up study to assess in detail the nature of the response style at work.

Reviewing the motivational items discussed in the context of the illustration of possible misinterpretation, Table 3 provides helpful insights to the Austrian NTO managers: “health and beauty” for the French tourists and “surroundings” for the American tourists cannot be considered safe advertising messages, because both are classified as HRIs. For the American respondents, “child care” emerges as a motivation item that can safely be interpreted as being

significantly different from those of other regions. Unfortunately, American tourists are significantly less interested in child care, thus making this item useless from a marketing point of view. This is not the case for “free and easygoing” for the Viennese tourists. This aspect is significantly more important to the Viennese than to other tourists, and would therefore be suitable for an advertising campaign targeting the Viennese.

Conclusions

Response style effects are a serious concern in empirical research throughout all disciplines of the social sciences, and are systematically associated with the respondents’ country of origin or cultural background. Consequently, data sets that consist of respondents from different countries of origin are in danger of misinterpreting differences between countries of origin as substantial differences with respect to the construct under study, rather than as the result of a CSRS.

Spanish respondents answers may be much more satisfied with a hotel and have a much higher intention to return than do Chinese respondents. However, this view likely reflects a CSRS, given that respondents of Hispanic background often prefer extreme answers, whereas Chinese respondents have the opposite tendency.

Being aware of the existence and potential distorting effect of CSRS on empirical study results, knowing how to detect them and — if necessary — correcting respondents’ answers for CSRS, is essential for any empirical marketing researcher engaging in the study of cross-cultural comparisons, or any other empirical research based on multicultural data.

Currently two frequently used ways of dealing with CSRS exist: to ignore them and analyze uncorrected data, or to choose one of many correction techniques, transform the original

data accordingly and analyze the transformed data. The problem with the first approach is that CSRS will probably contaminate the raw data, which would significantly influence the results of the analyses. The problem with the second approach is that any transformation of data is based on assumptions that may not actually be assured or appropriate, in which case the transformation is likely either to introduce new systematic contamination, or eliminate content-related information that would have been needed for the cross-cultural comparison.

This paper proposes a robustness-based approach requiring four steps: (1) the selection of a set of correction techniques appropriate for the problem and data at hand, (2) the correction of raw values according to all chosen correction techniques, (3) testing of cross-cultural differences on the basis of the raw data and all transformed data sets, and (4) computing of CSRS robustness indices. These steps should enable researchers to assess which cross-cultural conclusions can be safely drawn and which are endangered by the unclear effects of CSRS. Findings therefore should either not be reported as firm, or be further evaluated in a follow-up qualitative study.

Reliability is not a substitute for validity. However, the correction for response styles does not address the problem of validity. Validity needs to be ensured by asking the correct questions and testing the constructs under investigation for equivalence. The approach does not compensate for bad survey questions; rather, the proposal aims to assess analytical robustness of findings, assuming that the questions asked were well developed and sufficiently valid to measure what they were intended to measure.

CSRS robustness indices are useful for grouping items into four categories; findings concluding differences between countries may be not significant based on both raw and corrected answers of respondents. This represents the lowest-risk situation, and attracts the assumption that the two countries' respondents do not differ in, for instance, travel motivations.

Tests could also render significant results for both the raw and corrected values. While still a low-risk situation, the only insecurity here is that one test may indicate that, for instance, French respondents are more motivated by “rest and relax,” while the other test may indicate that French respondents are less motivated by “rest and relax.” After ensuring that the direction of the difference is the same, the researcher can assume that a difference exists which is not merely the result of the contamination of data with CSRS. All remaining situations are more problematic, because one test result indicates a difference and the other states the opposite. If the proportion of questionnaire items identified as HRI is low, the analysis can proceed without drawing too strong conclusions about HRIs, and instead focus on insights based on vLRIs and LRIs. However, if the proportion of HRIs is high, the researcher might also decide to include HRIs with a high CSRS robustness index, that is, where a difference is indicated for nearly all datasets.

Regardless, researchers should describe the chosen option in detail in any report in order to ensure that readers do not overestimate the value of conclusions based on HRIs.

The empirical example based on real data from a national guest survey illustrate the gravity of potential mistakes, while simultaneously demonstrating how empirical researchers could assess which variables allow reliable conclusions, and which may be contaminated by CSRS using the proposed robustness-based approach.

Making more use of binary answer formats that are not as susceptible to CSRS as are ordinal answer formats is another option for reducing the contamination level of data. Cronbach is the most prominent proponent of this option (1950: 21): “Since response sets are a nuisance, test designers should avoid forms of items which response sets infest.” Binary data format is unsuitable for some constructs (such as the evaluation of one’s own personality), where the rater is highly familiar with the rating object, and reasonably assumes that a very precise evaluation is

possible. However, other constructs, such as intentions to visit, are much more suitable for binary scales, because a binary choice forms the underlying construct. Advantages of binary scales have been reported by numerous researchers (Dolnicar 2003; Jacoby and Matell 1971; Komorita and Graham 1965; Matell and Jacoby 1971; Mazanec 1984; Peabody 1962), and should not be discarded as an option just because ordinal (typically, Likert-scaled answer formats) were more popular in the past, given the large number of reported methodological shortcomings of ordinal sales (Kampen and Swyngedouw 2000).

References

Arce-Ferrer A.J. and J.J. Ketterer, 2003, "The effect of scale tailoring for cross-cultural application on scale reliability and construct validity" in *Educational and Psychological Measurement*, 63(3):484-501.

Bachman J.G. and P.M. O'Malley, 1984, "Yea-saying, nay-saying and going to extremes: Black-white differences in response style" in *Public Opinion Quarterly*, 48:491-509.

Bartram D., 1996, "The relationship between ipsatized and normative measures of personality" in *Journal of Occupational and Organizational Psychology*, 69:25-39.

Baumgartner H. and J.B.E.M. Steenkamp, 2001, "Response styles in marketing research: A cross-national investigation" in *Journal of Marketing Research*, 38(2):143-156.

Bhalla G. and L.Y.S. Lin, 1987, "Cross-cultural marketing research: A discussion of equivalence issues and measurement strategies" in *Psychology and Marketing*, 4(4):275-285.

Chan W., 2003, "Analyzing Ipsative Data in Psychological Research" in *Behaviormetrika*, 30(1):99-121.

Chen C., S. Lee and H.W. Stevenson, 1995, "Response style and cross-cultural comparison of rating scales among East Asian and North American students" in *Psychological Science*, 6(3):170-175.

Cheung G.W. and R.B. Rensvold, 2000, "Assessing Extreme and Acquiescence Response Sets in Cross-Cultural Research using Structural Equation Modeling" in *Journal of Cross-Cultural Psychology*, 31(2):187-212.

Chun K.T., J.B. Campbell, and J.H. Yoo, 1974, "Extreme Response Style in Cross-Cultural Research" in *Journal of Cross-Cultural Psychology*, 5(4):465-480.

Cronbach L.J., 1950, "Further Evidence on Response Sets and Test Design" in *Educational and Psychological Measurement*, 10:3-31.

Das J.P. and T. Dutta, 1969, "Some correlates of extreme response set" in *Acta Psychologica*, 29(1):85-92.

Dolnicar S., 2003, "Simplifying three-way questionnaires — Do the advantages of binary answer categories compensate for the loss of information?" ANZMAC CD Proceedings 2003.

Drasgow F., 1987, "Study of the Measurement Bias of Two Standardized Psychological Tests" in *Journal of Applied Psychology*, 72(1):19-29.

Fischer R., 2004, "Standardization to Account for Cross-Cultural Response Bias — A Classification of Score Adjustment Procedures and Review of Research" in *Journal of Cross-Cultural Psychology*, 35(3):263-282.

Greenleaf E.A., 1992, "Improving Rating Scale Measures by Detecting and Correcting Bias Components in Some Response Styles" in *Journal of Marketing Research* (May), 29:176-188.

Greenleaf E.A., 1992, "Measuring Extreme Response Style" in *Public Opinion Quarterly*, 56(3):328-351.

Huang C.D., A.T. Church and M.S. Katigbak, 1997, "Identifying cultural differences in items and traits: Differential item functioning in the NEO personality inventory" in *Journal of Cross-Cultural Psychology*, 28(2):192-218.

Hui C.H., and H.C. Triandis, 1989, "Effects of Culture and Response Format on Extreme Response Style" in *Journal of Cross-Cultural Psychology*, 20(3):296-309.

Jacoby J. and M.S. Matell, 1971, "Three-Point Likert Scales Are Good Enough" in *Journal of Marketing Research*, 8:495-500.

Johnson T., Y.I. Cho and S. Shavitt, 2005, "The Relation Between Culture and Response Styles — Evidence From 19 Countries" in *Journal of Cross-Cultural Psychology*, 36(2):264-277.

Kampen J. and M. Swyngedouw, 2000, "The Ordinal Controversy Revisited" in *Quality and Quantity*, 34(1):87-102.

Komorita S.S. and W.K. Graham, 1965, "Number of scale points and the reliability of scales" in *Educational and Psychological Measurement*, 25(4):987-995.

Kozak M., E. Bigne and L. Andreu, 2003, "Limitations of cross-cultural customer satisfaction research and recommending alternative methods" in *Journal of Quality Assurance in Hospitality and Tourism*, 4(3/4):37-59.

Kruskal J., 1977, "The relationship between multidimensional scaling and clustering" in J.V. Ryzin (ed.) *Classification and Clustering*, Academic Press, Inc., New York, pp. 17-44.

Leung K. and M.H. Bond, 1989, "On the empirical identification of dimensions for cross-cultural comparisons" in *Journal of Cross-Cultural Psychology,* 20(2):133-151.

Marin G., R.J. Gamba, and B.V. Marin, 1992, "Extreme Response Style and Acquiescence among Hispanics — The Role of Acculturation and Education" in *Journal of Cross-Cultural Psychology*, 23(4):498-509.

Matell M.S. and J. Jacoby, 1971, "Is there an optimal number of alternatives for Likert scale items? Study 1: Reliability and validity" in *Educational and Psychological Measurement*, 31:657-74.

Mazanec J.A., 1984, "How to detect Travel Market Segments: A Clustering Approach" in *Journal of Travel Research*, 23(1):17-21.

Paulhus D.L., 1991, “Measurement and control of response bias” in J.P. Robinson, P.R. Shaver and L.S. Wrightsman (eds.), *Measures of Personality and Social Psychological Attitudes*, Academic Press, San Diego, pp. 17-59.

Paunonen S.V. and M.C. Ashton, 1998, “The structured assessment of personality across cultures” in *Journal of Cross-Cultural Psychology*, 29(1):150-170.

Peabody D., 1962, “Two components in bipolar scales: direction and extremeness” in *Psychological Review*, 69(2):65-73.

Roster C.A., R. Rogers and G. Albaum, 2003, “A cross-cultural/national study of respondents’ use of extreme categories for rating scales,” Proceedings of the Ninth Annual Cultural Research Conference 2003.

Sekaran U., 1983, “Methodological and theoretical issues and advancements on cross-cultural research” in *Journal of International Business Studies*, 14(2):61-73.

Shiomi K. and R. Loo, 1999, “Cross-cultural response styles an the Kirton adaptation-innovation inventory” in *Social Behaviour and Personality*, 27(4):413-420.

Si S.X. and J.B. Cullen, 1998, “Response categories and potential cultural biases: effects of an explicit middle point in cross-cultural surveys” in *International Journal of Organizational Analysis*, 6(3):218-230.

Smith A.M. and N.L. Reynolds, 2002, “Measuring cross-cultural service quality: A framework for assessment” in *International Marketing Review*, 19(4/5):450-481.

van Herk H., Y.H. Poortinga and T.M.M. Verhallen, 2004, “Response Styles in Rating Scales — Evidence of Method Bias in Data From Six EU Countries” in *Journal of Cross-Cultural Psychology*, 35(3):346-360.

Watson D., 1992, "Correcting for Acquiescent Response Bias in the Absence of a Balanced Scale: An Application to Class Consciousness" in *Sociological Methods and Research*, 21(1):52-88.

Yates J.F., L.W. Lee and J.G. Bush, 1997, "General knowledge overconfidence: cross-national variations, response style and reality" in *Organizational Behaviour and Human Decision Processes*, 70(2):87-94.

Table 1 CSRS Standardization-Based Correction Techniques

Category	Alternatives	Comments
Unit	Within-subject	Across variables for each individual, assumes no content-related differences between respondents.
	Within-group	Across individuals for each variable, assumes overall average scores and/or variance is comparable over groups.
	Within-culture	Across variables and individuals for each culture, assumes equality of response styles within culture.
	Double	Within-subject followed by within-group for each culture.
Adjustment using	Means	Removes ARS. Leads to ipsatization
	Dispersion indices	Removes ERS. Needs a balanced design with negative and positive items.
	Means and dispersion indices	See the separate discussion of means and dispersion indices.
	Covariates	Assumes that correlation between covariates and other items is due to response style.

Table 2 Raw and Standardized Mean Answers of Respondents by Region of Origin and Overall Average

	Raw					Corrected				
	France	Italy	USA	Vienna	Average	France	Italy	USA	Vienna	Average
Rest and relax	0.76*	0.66	0.51*	0.76*	0.69	0.56	0.54	-0.04*	0.75*	0.51
Comfort	0.54*	0.42*	0.48	0.50	0.48	-0.04	-0.16	-0.14	0.01*	-0.07
Sports	0.41	0.39	0.36	0.42	0.40	-0.40	-0.29	-0.47*	-0.17*	-0.31
Excitement	0.44*	0.31*	0.67*	0.29*	0.40	-0.28	-0.47*	0.43*	-0.52*	-0.28
Creativity	0.29	0.29	0.36*	0.26*	0.29	-0.70*	-0.55	-0.48*	-0.61	-0.59
Culture	0.70*	0.64*	0.69*	0.40*	0.58	0.39*	0.47*	0.53*	-0.24*	0.22
Fun	0.53	0.53	0.61*	0.44*	0.51	-0.05*	0.16*	0.26*	-0.09*	0.05
Good company	0.58	0.63*	0.62*	0.44*	0.55	0.08	0.42*	0.29*	-0.10*	0.14
Unspoiled nature	0.86*	0.77	0.76	0.75*	0.78	0.80	0.85*	0.69	0.70	0.76
Health and	0.32	0.33	0.26*	0.45*	0.36	-0.60*	-0.44	-0.76*	-0.13*	-0.43

beauty										
Surroundings	0.84*	0.57*	0.76*	0.72	0.72	0.74*	0.27*	0.71*	0.61	0.57
Free and easygoing	0.76*	0.49*	0.69	0.73*	0.67	0.53*	0.03*	0.47	0.64*	0.43
Entertainment	0.44*	0.37	0.47*	0.28*	0.37	-0.31	-0.31	-0.13*	-0.55*	-0.36
Atmosphere	0.43	0.37	0.47*	0.37*	0.40	-0.34	-0.32	-0.11*	-0.34	-0.29
Locals	0.69*	0.58	0.69*	0.47*	0.59	0.32*	0.27	0.48*	-0.06*	0.21
Sun and water/snow	0.47	0.43	0.31*	0.48*	0.43	-0.23	-0.16	-0.61*	-0.05*	-0.22
Cozyness	0.48*	0.53	0.43*	0.65*	0.54	-0.20*	0.15	-0.27*	0.41*	0.08
Organized	0.29	0.26	0.34*	0.29	0.29	-0.71*	-0.65	-0.54	-0.56	-0.61
Child care	0.17*	0.12	0.06*	0.12	0.12	-1.05	-1.02	-1.39*	-1.02*	-1.10
Maintain nature	0.83*	0.70	0.64*	0.70	0.72	0.71*	0.63	0.34*	0.55	0.57
Safety	0.84*	0.67*	0.79	0.77	0.76	0.75	0.55*	0.77	0.77	0.71

* Significantly different from the overall average at the p=0.05 level.

Table 3 Robustness of Cross-Cultural Findings

For HRIs the number of significant (S) to insignificant (I) differences are indicated in brackets by S:I. The seven different correction techniques are used: raw, within-subject and within-culture using means and/or standard deviations.

	France	Italy	USA	Vienna
Rest and relax	HRI (3:4)	vLRI	**LRI**	**LRI**
Comfort	HRI (1:6)	HRI (3:4)	vLRI	HRI (2:5)
Sports	vLRI	vLRI	HRI (4:3)	HRI (4:3)
Excitement	HRI (1:6)	**LRI**	**LRI**	**LRI**
Creativity	HRI (4:3)	vLRI	HRI (6:1)	HRI (2:5)
Culture	**LRI**	**LRI**	**LRI**	**LRI**
Fun	HRI (1:6)	HRI (4:3)	**LRI**	**LRI**
Good company	vLRI	**LRI**	**LRI**	**LRI**
Unspoiled nature	HRI (4:3)	HRI (1:6)	HRI (2:5)	HRI (3:4)
Health and beauty	HRI (5:2)	vLRI	**LRI**	**LRI**
Surroundings	**LRI**	**LRI**	HRI (5:2)	vLRI
Free and easygoing	**LRI**	**LRI**	HRI (2:5)	**LRI**

Entertainment	HRI (3:4)	vLRI	**LRI**	**LRI**
Atmosphere	vLRI	vLRI	**LRI**	HRI (3:4)
Locals	**LRI**	vLRI	**LRI**	**LRI**
Sun and water/snow	vLRI	vLRI	**LRI**	**LRI**
Cozyness	**LRI**	vLRI	**LRI**	**LRI**
Organized	HRI (4:3)	vLRI	HRI (3:4)	vLRI
Child care	HRI (3:4)	HRI (2:5)	**LRI**	HRI (1:6)
Maintain nature	**LRI**	vLRI	**LRI**	vLRI
Safety	HRI (3:4)	**LRI**	HRI (2:5)	HRI (2:5)

Figure 1 Classification of Variables Based on Robustness of Test Results

		Results from analyses based on corrected data	
		Significant difference between cultures	No significant difference between cultures
Results from analyses based on uncorrected data	Significant difference between cultures	Low-risk items (LRI): items that reliably discriminate between cultures	High-risk items (HRI): items that could be misinterpreted on the basis of systematic data contamination
	No significant difference between cultures	High-risk items (HRI): items that could be misinterpreted on the basis of systematic data contamination	Very low-risk items (vLRI): items that reliably do not discriminate between cultures