EU-Project FamWork
"Family Life and Professional Work: Conflict and Synergy"

A joint project of the Universities of
Munich (D), Fribourg (CH), Graz (A), Nijmegen (NL), Porto (P), Mons (B) and Palermo (I)

Research Report FamWork-04-P/01

Methodological Aspects of Cross-Cultural Research:
Measurement Challenge

Anne Marie Fontaine, Cláudia Andrade, Marisa Matias & Jorge Gato

University of Porto
Faculty of Psychology and Education
Rua do Campo Alegre, 1021/1055
4169 – 004 Porto – Portugal

http://sigarra.up.pt/fpceup/web_page.inicial

CONTENTS

INTRODUCTION: MEASUREMENT CHALLENGES IN CROSS-CULTURAL RESEARCH

One of the most important problems in cross-cultural research is the difficulty in ruling out alternative explanations. The main causes for this concern are the allocation of subjects and the lack of "experimental control" over cultural conditions. In studies with already existing groups of subjects, as is the case of cross-cultural research, the allocation of subjects is not random (quasi-experimental studies). In cross-cultural psychology, there is no control over the treatment administered to the subject, because the effect of cultural factors extends over a long period of time; therefore, the effect of a postulated cultural factor is inferred *post hoc* from differences between groups on a measurement (Van de Vijver and Leung, 1997).

Which controls are then available to diminish this bias?

According to Berry, Poortinga, Segall and Dasen (1995), there are four kinds of measure to reduce alternative explanations of differences between cultural groups:

1. A first group of measures refers to *a priori* selection of the cultural groups according to their position on the independent variable.
2. A second strategy is available when the dependent variable may be expressed in function of two or more separate scores.
3. A third group of measures includes the elimination of irrelevant variables effects by means of statistical analysis (covariance analysis and regression).
4. A final strategy refers to the use of more than just a measurement method: self-report scales, interviews, life histories analysis, etc. Doing so we are extending the data base from which we draw inferences, making them more valid.

It is well established that in cross-cultural research the challenges presented by methodological aspects are greater than in other fields of research. As Brislin (1976) stated:

> "(…) the goals of methodology in cross-cultural research are not different from those of other psychological research: reliability, validity, representativeness of experimental tasks, their generalization to behaviour outside of research studies...but, obtaining this goals is hard since there is often an unfamiliar language with which the researcher must work, and the research is done among people for whom the methods of empirical research in psychology (...) are unfamiliar and sometimes alien to their way of life (p. 216)."

According to Van de Vijver and Leung (1997), the major threat in cross-cultural studies is the bias involved in all three stages of the study: 1) theoretical conceptualisation and the formulation of research hypotheses, 2) design of the study and 3) data analysis. The purpose of this paper is not to analyse all these three aspects; instead, it focuses on problems associated with the measurement of variables in cross-cultural studies.

1. Equivalence

Two closely related concepts deserve attention: equivalence and bias. In the process of research, there is a constant need to establish equivalence at several levels in order to diminish the bias (Van de Vijver & Leung, 1997; Allen & Walsh, 2000). ***Equivalence*** is associated with the measurement level, where scores obtained in different cultural groups can be compared. ***Bias*** indicates the presence of factors that challenge the validity of cross-cultural comparisons.

There are several taxonomies distinguishing levels of equivalence and bias (e.g. Van de Vijver & Leung's, 1997; Canino & Bravo, 1994; Usunier, 1998; Allen & Walsh, 2000). Van de Vijver and Leung's classification (1997), distinguishes four levels of equivalence: construct equivalence, structural equivalence, measurement equivalence and scalar equivalence.

1.1. Construct equivalence

The word "construct" usually relates to a concept with several underlying dimensions. This concept can be measured quantitatively through the identification of its various dimensions. Construct equivalence can be understood regarding the following question: do the concepts/constructs under study have similar meaning across the social units studied (Usunier, 1998)?

1.2. Structural equivalence

Structural equivalence is present if an instrument administered in different cultural groups shows similar internal structures (such as factor structure) and similar relationships with other variables (Van de Vijver & Leung, 2001).

1.3. Measurement equivalence

This type of equivalence is shown if measurement scales have the same measurement unit across cultures, but different origins. In this case, the means cannot be directly compared; this is only possible in scalar equivalence (Van de Vijver & Leung, 2001).

1.4. Scalar equivalence or full score comparability

This equivalence can be achieved when (a) the measurement instrument is on the same ratio scale in each cultural group (same measurement unit) and (b) when scores on an instrument have the same interval scale across cultural groups. The presence of item bias makes scalar equivalence questionable (Van de Vijver & Leung 1997).

Thus, equivalence is both a function of the characteristics of an instrument and of the cultural group involved. According to Van de Vijver and Leung (1997), the level of equivalence is usually unknown in empirical studies. Equivalence cannot be assumed but should, instead, be established.

On the opposite side of equivalence is bias: scores are equivalent when they are unbiased.

2. Bias

Bias refers to all nuisance factors threatening the validity of cross-cultural comparisons and to a lack of similarity of psychological meaning of test scores across cultural groups (Van de Vijver, 2000). An overview of the biasing factors, their possible sources and of the procedures to detect and overcome these types of threats, will be presented (Van de Vijver & Leung, 1997; Van de Vijver, 2000).

There can be considered three types of bias: construct bias, method bias and item bias. These three do not have the same influence on test scores: construct and method bias have a global influence, whilst item bias has a local influence.

2.1. Construct bias

Very often, studies export their theories and concepts derived within a cultural background to another very dissimilar background, in which these theories and conceptions do not present relevance or usefulness. Consequently the construct measured may not be identical across cultural groups.

Construct bias can, thus, occur when (a) the measured construct or the behaviours from which items are sampled are not identical across cultures, and (b) there is a poor domain sampling in the instrument (underrepresented construct).

2.1.1. Procedures to detect construct bias (structure oriented techniques)

There are, at least three kinds of structure oriented techniques for detecting construct bias: factor analytic approaches, confirmatory factor analytic techniques and structural equation modelling.

2.1.1.1. Factor analytic approaches

In order to establish the cross-cultural validity of an adapted instrument, the set of relationships, both internal and external, must be demonstrated not to vary across cultural groups (Allen and Walsh, 2000). These relationships include (a) the relation within the factor structure of the assessment instrument, and (b) the relationship of the instrument with external correlates associated with the construct which the instrument is supposed to tap.

Invariance in the factor structure across groups can be analysed using factor analysis or other techniques directed to the detection of the underlying structure of the instruments. One of these techniques is *replicatory factor analysis* (Ben-Porath, 1990, in Allen and Walsh, 2000). Following this procedure, data from a representative sample of the group with whom the instrument will be adopted, are analysed using the same exploratory factor analysis techniques as in the original instrument. In this new analysis, the number of factors extracted is limited to the number of factors identified with the instrument in its culture of origin. This procedure gives a test of the factorial invariance or internal structure of the instrument across cultural groups. If the two structures vary significantly in these comparisons, a qualitative change occurred in what was being measured. If this is the case, it would be advisable to adapt a different instrument or to develop a culture-specific one. If the factorial structure is invariant across samples, one may continue with instrument validation (Allen & Walsh, 2000).

According to Van de Vijver and Leung, (1997), the most popular techniques for addressing construct equivalence are exploratory factor analysis, followed by target rotations and evaluation of factorial agreement/ congruence coefficient across samples.

2.1.1.2. Confirmatory Factor Analytic Techniques (CFA)

An extension of exploratory factor analysis methodology has been developed – confirmatory factor analytic techniques (CFA). This technique can also be included as structural equation modelling. CFA allows the testing of a theoretical model regarding factor structure, loadings of variables on factors and factor correlations in the adopted instrument, as found in the sample where it was first developed, using the data of the new group (Van de Vijver & Leung, 1997). Usually, the evaluation of model fit in the new sample follows a procedure of progressive constraints: commonly the researcher starts with a hypothesis of an equal number of factors across groups, followed by a test of the hypothesis of equal factor loadings. If this model shows fit, equality of factors covariance (correlations) can be added as another constraint; in a final step, the fit of the model can be used specifying equality of factor variances. It is also possible to begin the search with a highly restrictive model and diminish the equality constraints in the next models. At the end, if the two matrices are similar, the factor structure identified in the original research fits the data well in the new cultural group. Thus, it is possible to continue studying the external correlates of the test with the new group. If the two matrices are very dissimilar, the fitness of the model identified in previous research to the new data is poor, indicating that the factor structure varies between samples; in this case, it would be advisable to select a different instrument for possible adaptation or to develop a new culture-specific instrument (Allen and Walsh, 2000; Van de Vijver and Leung, 1997).

An extension of CFA refers to collecting data on multiple groups simultaneously. This is more suitable when an instrument is to be used with different cultural groups. The procedure, in this case, requires constraining all common factor loadings to be invariant across groups.

CFA limitations

By fitting the data to the factor structure prespecified in the sample of origin, CFA avoids some problems of exploratory factor analysis (EFA). However, it also presents some limitations (Allen & Walsh, 2000):

1 - Most CFA procedures assume multivariate normality. Nonetheless, the distribution of data obtained with instruments applied in new cultural groups can change considerably. In order to consider non-normality when present, some

estimation procedures that do not require this assumption and test statistics that allow for correction, have been developed.

2 – CFA estimation models have been developed for covariance matrices. When they are applied to correlations, some problems can occur, such as inaccurate standard errors for estimates of parameters and inaccurate fit indicators. (Cudek, 1989 in Allen & Walsh, 2000). Therefore, the covariance matrix from the original factor should be used whenever possible.

3 – To obtain stable parameter estimates large sample sizes are recommended. This can be difficult, as most cross-cultural studies are made with small population sizes.

2.1.1.3. Structural Equation Modelling – SEM (analysis of covariance structures)

SEM can be seen as a set of versatile data analytic tools with components of both regression and factor-analytic models. Some of these applications are path analysis, multidimensional scaling techniques, INDSCAL and cluster analysis. The most commonly used computer programs for SEM are LISREL and EQS (Van de Vijver and Leung, 1997).

Path analysis is of special importance when we are discussing causal models. This analysis enables the study of multiple dependent variables and direct and indirect effects of one set of variables on a dependent variable.

Multidimensional scaling techniques attempt to reproduce a matrix of distances between stimuli (questions of an inventory) in a small number of dimensions that can be meaningfully interpreted. This scaling is analogous to factorial analysis as far as rotational axes are concerned (distances between stimuli are not affected by orthogonal rotations of the axes); therefore, different rotations may be carried out in order to evaluate the solutions agreement. To carry out multidimensional scaling, there is a set of procedures named PINDIS: Procrustean Individual Differences Scaling, which allows the application of this technique to various cultural groups, optimising the agreement between the solutions. A computer program for this model is Matchals.

INDSCAL allows simultaneous modelling of cross-cultural similarities and differences. In this model, a number of cross-culturally identical dimensions are assumed to underlie a data set, although the weights (salience) of these dimensions may vary across cultures (Van de Vijver & Leung, 1997).

Cluster analysis is aimed at the classification of multivariate data in a limited set of non-overlapping categories; each category has some common characteristic that is not shared by members of another category. After cluster analyses for each cultural group have been performed separately. A dichotomous matrix may be built by comparing each cluster matrix, indicating which variables belong to which cluster. These dichotomous matrices can be compared across cultural groups, using common agreement indices for nominal data, such as Cohen's kappa (Van de Vijver & Leung, 1997).

2.1.2. Procedures to detect construct bias (level oriented techniques)

While structure orientation techniques focus on relationships among variables and attempt to identify similarities and differences in these relationships across cultures, level oriented techniques focus on differences in magnitude of variables across cultures.

The choice between an analysis of variance, a t-test or a regression analysis depends on the measurement level of the independent variables. Nominal and ordinal-level independent variables are analysed in an analysis of variance. Interval-level variables are usually analysed with a regression model (Van de Vijver & Leung, 1997).

2.1.2.1. Analysis of variance and t-test

In both techniques the null hypothesis specifies that there are no differences across cultural groups. The t-test is used to compare two cultural groups; analysis of variance is used when we are studying data from more than two groups. For both full score comparability is required. In these analyses a cultural group is the independent variable; the score on a psychological instrument constitutes the dependent variable. The aim is to study the main effect of culture, reflected in different ways, on the dependent variable, in the studied cultures.

In more complex designs, called factorial designs, in addition to culture, one or more independent variable, such as gender or age are included. Interaction effects between these new variables and culture can be seen through analysis of variance (Van de Vijver & Leung, 1997).

2.1.2.2. Regression approaches

When considering the adoption of an instrument into a new culture, it is also important to investigate the relation between test scores from the instrument and direct measures of cultural variables. Regression analysis is used for this purpose.

Regression encompasses a set of statistical techniques that allow the assessment of the influence of one or more independent/predictor variables on a dependent variable. The goal of regression is to produce an equation that provides the best prediction of a dependent variable from a group of variables. This technique provides regression coefficients for each independent variable that express the strength of the relationship between each one of the independent variables studied and the dependent variable. This statistic gives an overall evaluation of the success of the independent variables in predicting variation in the dependent variable (Allen & Walsh, 2000; Van de Vijver & Leung, 1997).

Regression has been used in cross cultural research to test bias, through the *Cleary Rule*: "a test developed for use in measurement of a construct is equivalent if it has the same regression equation with some external correlate of behaviour in the new cultural (…) group, as with the group with which it was developed" (Allen & Walsh, 2000, pp. 74). However, this rule contains some problematic assumptions: it assumes that the distribution of scores from the external behaviour correlate in the new cultural group is similar to the original group; it holds that the groups are matched on relevant third variables, such as socio-economic status and it assumes that the external correlate of behaviour is equivalent across groups (Nunnally & Bernstein, 1994 cited in Allen & Walsh, 2000).

This kind of methodology is used to determine the existence or absence of cultural variables influencing the results in terms of test scores, and in terms of the latent construct. It is important to know if the construct under assessment is equally considered in the same way across cultures, and if not, what variables may be altering construct equivalence. The goal is not simply to prove instrument validity by the equivalence of test scores, but in the case of construct inequivalence, to determine the factors that influence it and study them. These factors should not necessarily be treated as nuisance variables.

2.1.2.3. Item response theory approaches

This approach assumes a relationship between the person's response and the trait, attribute or attitude that is being investigated. It is mainly used with one-dimension instruments that promote dichotomic responses.

2.1.3. Other procedures to detect construct bias

In order to address construct bias, before data are gathered, two broad strategies can be used: decentring and use of local informants/bilingual subjects. Decentring consists in the development of the same instrument by a group of researchers from all intended cultures. Through this procedure it is likely that construct bias will be detected, even before the instrument is applied.

Local informants or bilingual subjects can be used to get some information about the target culture and also to informally administer the instrument, so that perceptions about the instruments can be held (Van de Vijver & Leung, 1997).

It is very important to assure that the domain of behaviour under study is identical cross-culturally. Thus, when making inferences, we can assure they are valid and not biased by different coverage of the domain in question. When the domain of behaviour is not identical cross-culturally, there is no valid comparison.

2.1.4. Limitations to a construct approach to equivalence

One of the most important limitations respecting this approach is that the methods exposed do not establish cross-cultural equivalence in the underlying construct; instead, they provide evidence to disconfirm the equivalence (logic of hypothesis disconfirmation), by assessing incompatibilities with construct equivalence. Besides construct equivalence, it is also important to establish linguistic and metric equivalence. Finally, it is also necessary to consider the context in which a test will be used (Allen & Walsh, 2000).

Thus, when there is a suspicion that the construct is not identical across cultures it is advisable to perform a local survey asking informants to describe the construct and its characteristics (Usunier, 1998; Van de Vijver & Leung, 1997).

2.2. Method Bias

Method bias refers to the presence of nuisance variables due to methodology related factors. Method bias can be further subdivided into three types: sample bias, instrument bias and administration procedures bias (Van de Vijver, 2000).

2.2.1. Sample bias

Differences between samples in test relevant characteristics, such as level of education, level of motivation or knowledge of testing language, can mislead the results and induce errors of interpretation (Van de Vijver, 2000).

If the primary goal is to verify differences between cultures, it is preferable to look for cultures with similar characteristics, therefore reducing the number of alternative explanations. If the goal is to look for universal patterns, then cultures as different as possible should be included, in order to obtain more information (Van de Vijver & Leung, 1997).

Next, some sampling procedures will be explored (according to Van de Vijver and Leung, 1997). In *convenience sampling* researchers select a culture because of its convenience: the investigators are from that culture or they are acquainted with local collaborators. Though this choice is not theoretically based, it is a procedure which involves lower costs.

In *systematic sampling*, cultures are selected in a theory-guided fashion: the cultures selected should present different values on a theoretical continuum. In order to make the most of this procedure, cultures far apart in one theoretical dimension should be selected. Thus, the chance of detecting cultural differences is maximized. Some studies which use this procedure are theory driven or generalizability studies.

Random sampling involves the sampling of a large number of cultures in a random way. This is the most desirable strategy for generalizability studies, although, in its true form, it is impractical, because of limits of time and resources. What commonly happens is finding as many collaborators as possible in several different cultures.

Another type, *stratified random sampling*, can be used to control demographic differences across cultural groups, thus limiting the alternative explanations of the results found. It can be made by matching the subjects: only the subjects with a certain demographic profile will be sampled for the study. This strategy can also be hard to employ because of limitations in subject availability, or because the groups under analysis are too different in their demographic variables. When this is the case, a

statistical control approach can be used, involving the measurement of the major demographic variables upon which the cultural groups vary. These variables will then be treated as covariates and controlled when cultural comparisons are made (Van de Vijver & Leung, 1997).

2.2.2. Instrument bias

This kind of bias refers to instrument characteristics that induce cross-cultural score differences, which bear no relation to the construct under study. This can arise from response biases such as differential response styles (acquiescence and extreme ratings); differential social desirability and from stimulus (un)familiarity.

According to Paulhus (1991), "A response bias is a systematic tendency to respond to a range of questionnaire items on some basis other than the specific item content. When the individual displays the bias consistently across time and situations, the bias is named *response style"* (p.17). Next, some kinds of response bias will be described.

2.2.2.1. Socially Desirable Responding (SDR)

The tendency to give answers that make respondents look good has been named *socially desirable responding*. This effect can be controlled or measured and taken into account when analysing the results.

Measurement of SDR:

Social desirability measures are very helpful to support the discriminant validity of an instrument; they are also useful for covariate and target rotation techniques. Some of the measures are (Paulhus, 1991):

- *Edwards Social Desirability Scale* (Edwards, 1997)
- *Marlowe- Crowne Social Desirability Scale* (Crowne & Marlowe, 1960)
- *MMPI Lie Scale* (Hathaway & Mc Kinley, 1951)
- *MMPI K Scale* (Meehl & Hathaway, 1946)
- *Balanced Inventory of Desirable Responding* (Paulhus, 1984, 1988)
- *RD -16* (Schuessler, Hittle & Cardascia, 1978)
- *Children's Social Desirability Scale* (Crandall, Crandall & Katkovsky, 1965)

Control of the influence of SDR:

Control of the influence of SDR can be made through rational techniques, factor analytic techniques, covariate techniques, demand reduction and stress minimization.

(a) *Rational techniques*: the self-report instruments can be built in a way that prevents the subject from responding in a socially desirable way. A forced-choice format, in which the two statements are equated for social desirability may be used. As far as single statements are concerned, they should be neutral regarding this aspect (Paulhus, 1991).

(b) *Factor analytic techniques*: if a component appears to represent SDR, it can be deleted before the factors are rotated. If a measure of SDR was administered, then the above mentioned component can be rotated to the SDR measure (Cheung & Chan, 2002; Paulhus, 1991).

(c) *Covariate techniques*: a measure of SDR can be administered along with content measures; then, SDR may be partialed out of correlations between two content scales to control for spurious correlation. An alternative is to adjust the raw score obtained, by regressing the content score on SDR: the residual then represents the content score corrected for SDR.

(d) *Demand reduction* is achieved through methods that reduce situational pressure for desirable responding. The most obvious is to assure respondents` anonymity. This can be achieved by physically separating respondents, insisting that they put no identifying marks on the questionnaire, or saying beforehand that the questionnaire will be sealed in an envelope and dropped in a box on the way out. Another strategy, when a match across two administrations is necessary, is to ask respondents to give their birthdays or use a consistent pseudonym for all administrations. Another method is the “bogus popeline” – a pseudo lie-detector. In its most aggressive form, respondents are hooked to electronic equipment that the researcher claims can assess their attitude directly through physiological measures. This technique can also be applied in less aggressive versions, such as warning the respondents that in the questionnaire there are methods to detect faking.

In face-to-face interviews a randomised response method can be applied: a desirability loaded question is presented along with a neutral question, the respondent is asked to flip a coin and answer the first question if tails, and the second if heads. This way, the respondent is under less pressure to respond desirably, as the researcher did not know, beforehand, what the item selected could be. The evidence

suggests that even though these techniques increase the report of sensitive behaviours and attitudes it has some limitations.

Another technique refers to questioning acquaintances of the target subject about the target behaviour (Paulhus, 1991).

(e) *Stress minimization*: refers to the reduction of tension during the test administration (Paulhus, 1991).

2.2.2.2. Acquiescence

Acquiescence is the tendency to agree rather than disagree with propositions in general. This tendency is assumed to be the result of subjects' uncertainty. There can be the yea-sayer, who tends to agree with statements or say, "yes" to questions; and the naysayers, who tend to disagree with statements or say "no" to questions. Two types of acquiescence can be distinguished: agreement acquiescence and acceptance acquiescence. In the former, a person tends to agree with all types of items, even an item and its own negation. In the latter, a person tends to endorse all qualities as true for him/herself, even apparently contrary ones (Cheung & Chan, 2000; Paulhus, 1991).

Ways of controlling acquiescence (Paulhus, 1991):

- Balance the scoring key: usually half the items are keyed positively and half the items are keyed negatively. This controls the agreement acquiescence: in order to obtain a high score, the respondent must agree and also disagree with the same number of items.
- To address acceptance acquiescence it is necessary to add conceptual opposites that are also worded as assertions.

2.2.2.3. Extremity Response Bias (ERB)

It is the tendency to use extreme choices on a rating scale. The problem arises from the difficulty in distinguishing whether an extreme rating indicates a strong opinion or a tendency to use the extremities of rating scales. ERB can be controlled by putting questions in multiple-choice format. Reducing the options to two also eliminates ERB but it strongly reduces the sensitivity of the measure (Paulhus, 1991).

To control the bias due to the divergence in response styles (item-non response pattern, median response style or extreme response style) data should be standardized in national data sets (standardized scores are adjusted so that they are normally

distributed with zero mean and unit standard deviation). The use of standarized scores offers a common metric for the various data sets. A monotrait-multimethod matrix can be applied when the data are collected with more than one method and results are compared across methods. Cross-cultural differences in ERB will be enhanced through comparison of the effect size of the method across cultures. Dissimilar effect sizes point to different influences of response procedures across cultures (Van de Vijver & Leung 1997).

This procedure has the strong disadvantage of eliminating possibly significant differences in mean, across national samples. Therefore, it is advisable to use this procedure only when there is suspicion of inequivalence, especially in response style. Another technique to control response bias involves the interviewee rating his familiarity with the response procedure applied and with the stimulus characteristics. A statistical correction for cross-cultural differences in familiarity is then possible to achieve, in an analysis of covariance (Paulhus, 1991).

For measuring metric equivalence, one can compare reliabilities across national data sets; or test for equality of measurement error variances (δ) using multiple groups with LISREL (Usunier, 1998).

2.2.3. Administration Bias

Administration bias has its origins in the personal characteristics of the researcher and in his/her interaction with the interviewee (problems in communication, different interviewing skills). It can also stem from different administration conditions: noise in the environment, the presence of other people besides the interviewer/interviewee or language problems. Another factor that bears some relevance refers to self-disclosure. There seem to be relevant differences between cultural groups in the tendency to disclose private information to strangers (as the researcher) (Van de Vijver, 2000; Van de Vijver & Leung, 1997).

According to Van de Vijver and Leung (1997), some of the most important problems in the administration of instruments in cross-cultural contexts are*:* tester/interviewer associated problems, testee/interviewee and interaction between these two.

2.2.3.1. Interviewer effect

The most widely reported interviewer effect refers to the presence of a culturally different person (the interviewer) that affects the respondent's behaviour: reluctance to answer (when respondents feel that the interviewer is intruding upon their privacy) or conscious biasing of the answers (when respondents fear the researcher's opinion).

To overcome this undesirable effect, there are two techniques: (a) the interviewer should be formerly introduced to the respondent and be alerted to these effects; (b) *a posteriori* technique refers to the measurement of the interviewer's characteristics (in this way a statistical correction can be performed).

Cross-cultural studies often involve dissimilar groups of subjects, with several background differences; thus, another administration problem can arise from sample incomparability. To deal with this problem, lengthy instructions, including various examples and exercises, should be given beforehand. The instrument can also be previously applied in a pilot study, without the objective of information gathering, in order to analyse the instrument. The observer should assure him/herself that the respondent understands all the questions. Only after this should the sample be collected.

A posteriori procedures include measurement of relevant background characteristics and their statistical correction.

2.2.3.2. Interviewee/interviewer interaction problems

This interaction problem refers mainly to ambiguous communication. In order to improve the accurateness of communication, when there's a different background, the interviewers, besides knowing how to administer the instrument, must also be masters in intercultural communication (Hambleton & Kanjee, 1997). Another procedure involves the measurement of the interviewer's communication skills so that they can be used as covariates in *a posteriori* statistical analysis.

2.2.4. Procedures to deal with method bias (Paulhus, 1991; Van de Vijver & Leung, 1997)

To deal with method bias, common strategies involve:

- Use of fixed scoring rules;
- Standardized administration of the instrument;

- Training of test administrators in usual interview skills, instructions that should be transmitted and in intercultural communication skills;
- Detailed protocol to perform interpretations;
- Use of collateral information, such as the behaviour of the respondent during the assessment;
- Strategies to control background variables that may affect the score.

The measurement of context variables is a design adaptation that can be made to reduce the number of alternative explanations. It can encompass sampling matching (previously explained) and statistical control. Context variables can be related with the subject (age, gender, psychological characteristics) or related with the culture (educational status, health institutions, demographic variables...). This measurement is suitable when group differences are large. The process goes as follows: context variables are measured at individual and group level; then, an analysis of covariance or hierarchical regression procedure is applied; the impact of context variables on the observed score differences of cultural groups is evaluated and statistically corrected (Van de Vijver, 2000).

Other design adaptations include (Van de Vijver & Leung, 1997):

- Repeated test administrations, in order to compare score changes across cultural groups;
- Assessment of social desirability of the items and of response styles;
- Pilot studies to examine precarious aspects of the instrument and its administration;
- Triangulation to enhance validity (application of diverse measures in order to capture the same construct). It assumes that if convergent results are obtained with different measures, bias has no influence on those results. This technique is most useful when statistical techniques for detecting biases cannot be applied.

2.3. Item bias

An item is said to be biased if persons with the same trait, but coming from different cultures are not equally likely to endorse the item. Some reasons for such differential response patterns may be due to differences in appropriateness of the item content, inadequate item formulation or inadequate translation. This bias is also known as differential item functioning.

Item bias can be seen in several ways: as an indicator that an instrument is inadequate for cross-cultural comparison, as an indicator of important cross-cultural differences (such as cultural idiosyncrasies), and as an undesirable phenomenon that must be removed, because only unbiased items constitute a solid base for cross-cultural comparison.

2.3.1. Procedures to deal with item bias

In order to target *item bias*, there are 2 procedures, a judgmental and a statistical one. The former refers to linguistic and psychological analysis; the latter refers to differential item analysis (Van de Vijver, 2000; Berry *et al.*, 1995).

2.3.1.1. Judgmental procedure

In judgmental item bias procedures, a content analysis of an item is made using the help of experts in the target group for which the bias is examined. Experts give their opinion about the content of the stimulus, evaluating if it belongs or not to the domain of the target behaviour. They also evaluate if it implies specific knowledge or more experience in one culture than in other. As these judges should have very close knowledge of the cultures involved in the study, and also be masters in the theories and notions underlying the instrument, the easiest way to perform this method is by consulting colleagues from the cultures involved and asking their opinions (Berry *et al.*, 1995).

The process of instrument translation and adaptation (translation equivalence attainment) can be included in these procedures. This process is also one of the most widely reported problems in the literature about equivalence in cross-cultural research. As a consequence special attention will be given to this topic.

2.3.1.1.1. Enhancement of the validity in multilingual studies – Translation Equivalence

According to Usunier (1998, pp.49), there are at least three elements in language that influence the research process:

> *"Words, as they signal specific meaning; Words as they are assembled in sentences and text through grammar ad syntax and work as codes that must in some way be "translated" into other codes; and language, in general, provides the speaker with a particular world view."*

For most researchers, the first problem arises with the choice of an assessment tool that, in most cases, is in English. The challenge is to assure that the several translated and adapted versions of the instrument are equivalent to the original one. Only when achieving this equivalence, will it be possible to compare the results (Canino & Bravo, 1994).

As Ellis (1989) stated, though language is a defining characteristic of a culture, when cultural differences and similarities are under investigation, language differences become a serious measurement problem that may inhibit valid inferences from the results obtained. Thus, for an instrument to produce comparable scores, it is necessary for it to be capable of identifying similar phenomena or psychological constructs, regardless of the language in which it is presented.

Obtention of translation equivalence is only completely attained when the following equivalence subcategories are considered: lexical equivalence, (which is provided by dictionaries), idiomatic equivalence (refers to idiosyncratic expressions from some cultures that are difficult to translate), grammatical-syntactical equivalence (how words are ordered in a language, sentences are constructed and meaning is expressed) and experiential equivalence (what words and sentences mean for people in their everyday experience) (Usunier,1998).

Carlson et al. (2000) mention that besides grammatical translation difficulties, the assumption that concepts are universally held across cultures and that they are readily transferable is a serious error that can undermine all the process of adapting instruments.

Another area of concern relates to possible changes in the psychometric properties of the instrument after translating/adapting it.

In multilingual studies, there are three available options to attain translation equivalence, from a psychological point of view (Van de Vijver & Leung, 1997):

1) Direct application or one-way translation (Carlson *et al.*, 2000): an instrument is applied through literal translation, when it is thought to be psychologically and linguistically appropriate in all groups under study. This is the option that involves less effort and costs and that allows direct comparisons with other studies that have used the same instrument. The main disadvantages relate to construct inequivalence, lower reliability and lower validity.

2) An instrument is adapted for use in a different cultural context. There occurs a literal translation of a set of items and a change in wording or contents of other items.

3) When it is assumed that the original instrument is inadequate in the new context, a new instrument will be developed to capture the construct more adequately. This process is called assembly.

When the presumed impact of bias is low, a literal translation and application of the instrument in the new context should be the option. If the impact of construct and method bias is expected though not so large as to invalidate the whole instrument (only a few items are expected to show cultural idiosyncrasies) the option should be to adapt the instrument, because the composition of a new instrument will limit the opportunity of cross-cultural comparison. If construct bias is expected to play a significant role, a new instrument should be assembled (Van de Vijver & Leung, 1997).

A very important question refers to the *translatability of items* and of the instrument as a whole. A text is said to be poorly translatable, when the loss of salient characteristics cannot be avoided in translation. Globally, these characteristics refer to denotations, connotations, and language-specific meanings (Van de Vijver & Leung, 1997).

Brislin (1986, in Van de Vijver & Leung, 1997), presented some guidelines for the writing of new items and modifying existing ones:

> "1) Use short, simple sentences of fewer than 16 words.
>
> 2) Employ the active rather than the passive voice, because the former is easier to comprehend (...)
>
> 3) Repeat nouns instead of using pronouns (…)
>
> 4) Avoid metaphors and colloquialisms.
>
> 5) Avoid the subjunctive form, with words like could and would (…)
>
> 6) Add sentences to provide context for key ideas. Redundancy is not harmful for communicating key aspects of the instrument.
>
> 7) Avoid verbs and prepositions telling "where" and "when" that do not have a definite meaning. (For example: often, refers to how many times exactly.)
>
> 8) Avoid possessive forms where possible, because it may be difficult to determine the ownership (…)
>
> 9) Use specific rather than general terms. Who is included in "members of your family" strongly differs across cultures (…) (p. 38)."

2.3.1.1.2. Instrument translation/adaptation procedures

Instrument translation and adaptation procedures aim to establish linguistic equivalence before instrument administration. These include: translation/back translation, cultural decentering of the instrument and the committee approach.

The most commonly referred procedure and also the most frequently applied is *translation/back translation*. In this technique the original version is translated by an interpreter to the target language; a second interpreter or a group of interpreters who don't know the original version, translate the text back to its original language. The accuracy of the translation is evaluated by comparing the original and back translated versions. Differences between the two versions point to translation problems that must be dealt with. A limitation of this procedure relates to the fact that literal translations may not capture the meaning of the item. Moreover, possible differences in readability, comprehensibility and natural flow of the test for both versions may be ignored.

According to Van de Vijver (2000), a good translation requires (a) the combined expertise of linguistic experts who take care of the linguistic equivalence, and (b) psychological experts who ensure psychological equivalence as well as psychometric adequacy. Translators can also be given instructions regarding inference, wording and phrasing and by emphasizing test "adaptation" over test "translation" (Carlson *et al.*, 2000).

An alternative to this procedure is *cultural decentring of the instrument*. In this case an instrument is retrospectively changed in order to assure translatability (words and concepts that are difficult to translate or that are culture-specific are removed from the original version). A pre-test of the translated research instrument in the target culture is necessary, until satisfactory levels of reliability on conceptual and measurement equivalence are attained. The disadvantage of this procedure is the labour and effort it requires, as it encompasses a multilingual work group with expertise in the construct under study (Usunier, 1998; Van de Vijver & Leung, 1997).

A third alternative to obtain linguistic equivalence refers to a *committee approach*. Committees of bilingual subjects are contacted to translate or adapt the instrument. The advantage of this approach resides in the cooperative effort that can enhance the translation quality. The main disadvantage is the absence of an independent evaluation of the translation adequacy (Carlson *et al.*, 2000; Van de Vijver & Leung, 1997).

Besides judgemental procedures (used before instrument administration, some statistical procedures can also be used to detect accuracy in the translated version. In these procedures, item statistics obtained in several linguistic versions are compared. When these item statistics are dissimilar, inequivalence due to poor translation, inadequate item content must be considered. Item bias techniques can also be used to assess this linguistic equivalence.

After having a translated/adapted instrument it is advisable to pre-test it. This pre-test can be done in two ways. In the first, a translation test is made by asking experts to review the translated version for clarity and linguistic appropriateness. The second way, and the most reliable, relates to the pre-testing of the instrument in a sample similar to the one that will be the target of the investigation. That is, individuals from the target population are asked individually to read or to listen to the questions and to paraphrase their understanding of it. The responses should resemble the original language version of the instrument. Discrepancies can be reviewed and analysed for misinterpretations. After this verbal reflexion, an equivalence test on a small group can be performed. Here, it is also possible to test the psychometric properties of equivalence, reliability and score distribution (Carlson *et al.*, 2000).

Vallerand (1989, in Banville, Desrosiers & Genet-Volet, 2000) presented an illustrative schema of the steps involved in the process of translation/adaptation of instruments. This schema entails some of the recommendations previously made (Schema1).

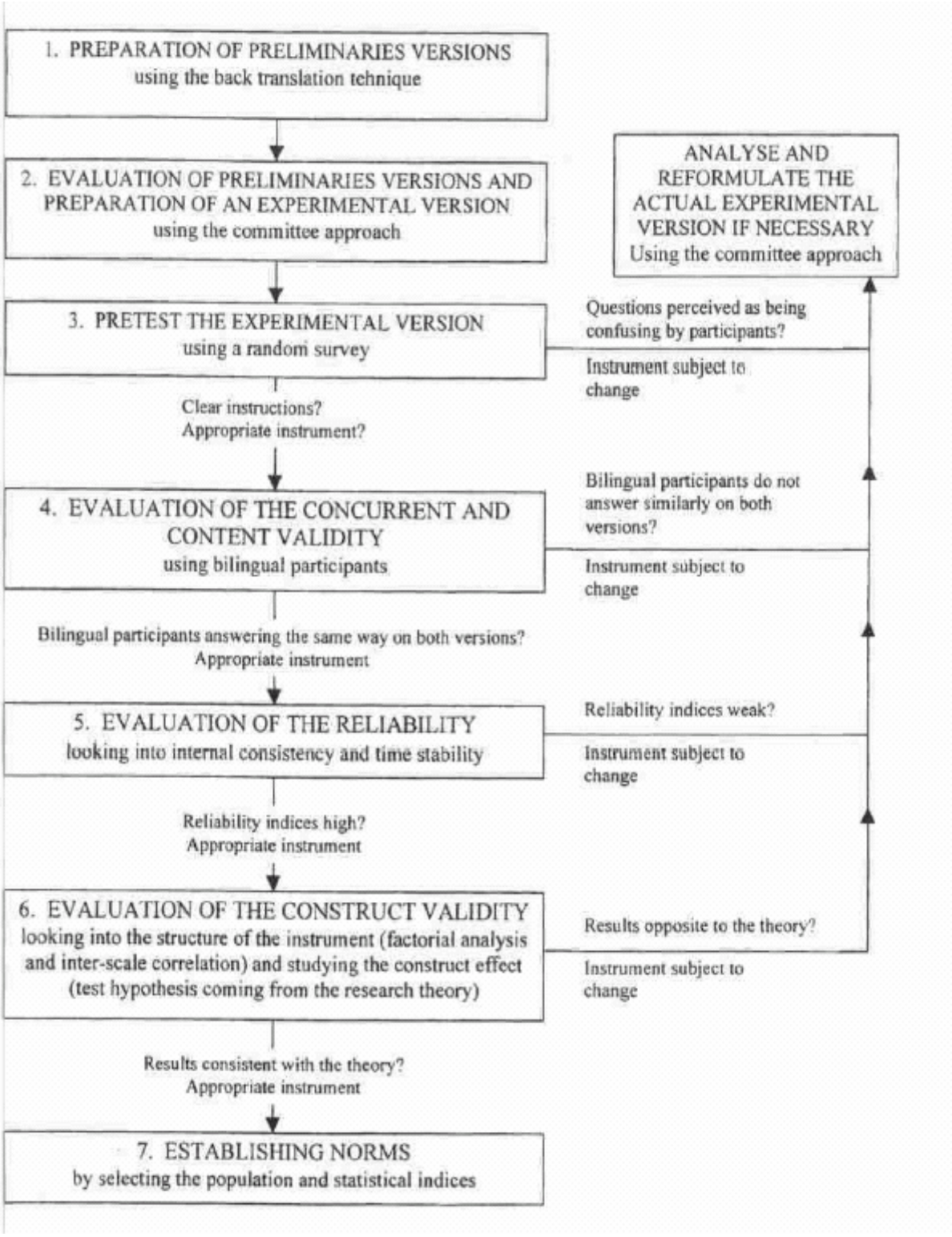
2.3.1.2. Differential item analysis

The previous procedures concerning item bias detection and solution are *a priori* proceedings. *A posteriori* procedures relate to statistical analysis, also known as differential item functioning techniques. An item, as previously stated, is an unbiased measure of a theoretical construct. It is expected that persons with an equal standing on the theoretical construct underlying the instrument should have the same score on the item, independently of group membership (Van de Vijver and Leung, 1997).

There are three methods to identify bias: (1) analysis of variance, a technique mostly used for ratio and intervalar level data; (2) Mantel-Haenszel statistics for dichotomous data; and (3) item response theory, mostly applied with dichotomous variables, though it can also be applied with polychotomus variables.

As Van de Vijver and Leung (1997) stated “Item Response Theory (IRT) proposes that item responses can be related to a latent trait by means of a logistic curve, usually specified by three parameters. The probability of a positive response to an item is defined as a function of an individual standing on the latent trait that the item assesses (p. 74)”. Before an IRT analysis is carried out it is necessary to establish the unidimensionality of the scale. If the scale is multidimensional, each one-dimensional subscale should be examined separately. Factor analysis can be used for this purpose.

Schema1: Steps included in the cross-cultural translation technique (Vallerand, 1989)



IRT is very useful for cross-cultural research because the estimates of item parameters do not depend on the standing of a group on the latent trait studied, and because fit tests can be applied to evaluate the extent to which empirical data conform

to the theoretical model. An item is said to be biased if one or more of its parameters differ significantly across cultural groups. Two approaches have been applied using IRT models: parameter and model–based comparisons (Van de Vijver & Leung, 1997).

Parameter-based procedures for detecting item bias include the following steps (Van de Vijver & Leung, 1997; pp.77, 78):

1. "An item response theory model with the appropriate number of parameters is selected to fit the data in each culture. The two–parameter model is used for attitudinal data and personality measures.

2. The parameters identified for each cultural group are equated on the same metric through an iterative linking procedure.

3. Biased items are deleted and eliminated with the aid of item characteristic curves and a chi-square test. The parameters are equated again with the linking procedure applied to unbiased items only; this procedure stops when no biased items are deleted.

4. The biased items identified are eliminated from the scale before cross-cultural comparisons are made."

In the *model-comparison approach*, two models are compared. The compact model assumes that there is no difference between the item parameters for all items across the two groups. A chi-square statistic reflecting how well the compact model fits the data, is then obtained. The augmented model assumes that the parameters of the item being tested are different across groups; a chi-square statistic is obtained. The two chi-square statistics are compared, and if the difference is significant the item shows bias. Regardless of its usefulness to cross-cultural research, IRT has some important limitations, namely the criteria that have to be met and the assumption that there is no transfer between item responses. It is also necessary to have large sample sizes in order to obtain stable estimates, which it is not always possible (Van de Vijver & Leung, 1997)

A final procedure to promote validity of assessment in cross-cultural research entails the application of differential norms to promote fairness and accuracy of the decisions based on the test results. Similar scores obtained by individuals from different cultures may have a different meaning. Differential norms can be used to

correct for social inequality and unequal opportunities in society for various cultural groups.

Concluding Remarks

Some highlights about the measurement challenges have been presented in this paper. The focus was on the researcher's needs at the level of instrument assembly and administration, and preliminary data analysis. As far as data interpretation is concerned, there are some precautions to be addressed as well as some strategies designed to increase accurateness of inferences. However, this will be the target of another paper. In Appendix A, some general guidelines concerning the translation and adaptation of instruments can be found.

References

Allen, James; Walsh, James (2000). *A construct-based approach to equivalence: methodologies for cross-cultural/multicultural personality assessment research.* In Richard H. Dana (ed.), Handbook of cross-cultural and multicultural personality assessment (pp. 63-86). Mahwah, NJ: Lawrence Erlbaum.

Banville, Dominique; Desrosiers, Pauline; Genet-Volet, Yvette (2000). *Translating questionnaires and inventories using a cross-cultural translation technique.* Journal of Teaching in Physical Education, vol. 19, pp. 374-387.

Berry, John W. ; Poortinga, Ype H.; Segall, Marshall H.; Dasen, Pierre R. (1995). *Cross-cultural psychology: research and applications.* Cambridge: Cambridge University Press

Brislin, Richard W. (1976). *Comparative research methodology: cross-cultural studies*. International journal of psychology, vol.11, n.º3, pp. 215-229

Canino, Glorisa; Bravo, Milagro (1994). *The adaptation and testing of diagnostic and outcome measures for cross-cultural research.* In International Review of Psychiatric, sep94, vol. 6, Issue 4, pp. 281-287.

Carlson, E. (2000). *A case study in translation methodology using the health promotion lifestyle profile II.* Public health nursing, vol. 17, nº. 1, pp.61-70.

Cheung, Mike; Chang, Wai (2002). *Reducing uniform response bias with ipsative measurement in multiple group confirmatory factor analysis*. Structural Equation Modeling, vol. 9, nº. 1, pp. 55-77.

Ellis, Barbara (1989). *Differential item functioning: implications for test translations*. Journal of Applied Psychology, 1989, vol. 74, n.º 6, pp. 912-921. American Psychological Association

Hambleton, R. K.; Kanjee, A. (1997). *Translation of test and attitude scales.* In John. P.Keeves - Educational Research, Methodology, and Measurement: an International Handbook, 2nd. Edition (pp-965-970). Pergamon: Cambridge, UK.

Paulhus, Delroy (1991). *Measurement and control of response bias*. In John Robinson, Philip Shaver, Lawrence Wrightsman (Eds.). Measures of personality and social psychological attitudes (pp. 17-59). San Diego: Academic Press

Usunier, Jean-Claude (1998). *International & Cross-cultural Management Research*. London: Sage Publications

Van de Vijver, Fons (2000). *The nature of bias*. In Richard H. Dana (ed.). Handbook of cross-cultural and multicultural personality assessment (pp. 87-106). Mahwah, NJ: Lawrence Erlbaum.

Van de Vijver, Fons; Leung, Kwok (1997). *Methods and data analysis for cross-cultural research*. Thousand Oaks, CA: Sage

Van de Vijver, Fons; Leung, Kwok (2001). Personality in cultural context: methodological issues. Journal of personality: vol. 69; nº. 6, pp. 1007-1031.

APPENDIX A: GUIDELINES FOR TRANSLATING AND ADAPTING PSYCHOLOGICAL AND EDUCATIONAL INSTRUMENTS

A committee representing several psychological organizations (Van de Vijver & Leung, 1997) has formulated guidelines for translating and adapting psychological and educational instruments.

Guidelines on context (defining the general background):

1 – "Effects of cultural differences that are not relevant or important to the main purposes of the study should be minimized to the extent possible"

2 – "The amount of overlap in the constructs in the population of interest should be assessed"

Guidelines on instrument development, translation and adaptation (recommended practices in designing multilingual instruments):

3 – "Instrument developers/publishers should ensure that the translation/adaptation process takes full account of linguistic and cultural differences among the populations for whom the translated/adapted versions of the instrument are intended"

4 - "Instrument developers/publishers should provide evidence that the language use in the directions, rubrics, and items themselves as well as in the handbook are appropriate for all cultural and language populations for whom the instrument is intended"

5 - "Instrument developers/publishers should provide evidence that the testing techniques, item formats, test conventions, and procedures are familiar to all intended populations."

6 - "Instrument developers/publishers should provide evidence that item content and stimulus materials are familiar to all intended populations."

This guideline does not pose that all stimulus material should be equally familiar to all individuals; instead it states that stimulus should be familiar through previous experience or should be made familiar through lengthy instruction.

7 - "Instrument developers/publishers should implement systematic judgmental evidence, both linguistic and psychological, to improve the accuracy of the translation/ adaptation process and compile evidence on the equivalence of all language versions."

8 - "Instrument developers/publishers should ensure that the data collection design permits the use of appropriate statistical techniques to establish item equivalence between the different language versions of the instrument."

9 - "Instrument developers/publishers should apply appropriate statistical techniques to (a) establish the equivalence of the different versions of the instrument and (b) identify problematic components or aspects of the instrument which may be inadequate to one or more of the intended populations."

10 - "Instrument developers/publishers should provide information on the evaluation of validity in all target populations for whom the translated/adapted versions are intended."

11 - "Instrument developers/publishers should provide statistical evidence of the equivalence of questions for all intended populations."

12 – "Non-equivalent questions between versions intended for different populations should not be used in preparing a common scale or in comparing these populations. However, they may be useful in enhancing content validity of scores reported for each population separately."

Guidelines on administration (defining issues regarding instruments administrations):

13 - "Instrument developers/publishers should try to anticipate the types of problems that can be expected and take appropriate actions to remedy these problems through the preparation of appropriate materials and instructions."

14 - "Instrument developers/publishers should be sensitive to a number of factors related to the stimulus materials, administration procedures, and response modes that can moderate the validity of the inferences drawn from the scores."

15 – "Those aspects of the environment that influence the administration of an instrument should be made as similar as possible across populations for whom the instrument is intended."

16 – "Instrument administration instructions should be in the source and target languages to minimize the influence of unwanted sources of variation across populations."

A pre-test session in which subjects can become acquainted with the testing or interview situation will also reduce cross-cultural differences in stimulus or response format familiarity.

17 – "The instrument manual should specify all aspects of the instrument and its administration that require scrutiny in the application of the instrument in a new cultural context."

18 – "The administration should be unobtrusive, and the examiner-examinee interaction should be minimized. Explicit rules that are described in the manual for the instrument should be followed"