

Michael Eduardo Reichenheim

Claudia Leite Moraes

Operationalizing the cross-cultural adaptation of epidemiological measurement instruments

ABSTRACT

The objective of the article was to offer an operational framework to assess cross cultural adaptation processes of instruments developed in other linguistic, social and cultural contexts. It covers the need for using robust measurement tools; the importance of 'universal' instruments that permit cross cultural fine-tuning; and stresses the need for adapting existent instruments rather than developing new ones. Existing controversies and proposals for different procedures in current literature are reviewed and a model for adapting instruments is presented. This synthesis covers the operational steps involved in evaluating concepts, semantic and operational items, and presents psychometric analysis guidelines that underlay an evaluation of measurement equivalence. Finally, the need for adequately controlling the quality of information presented in epidemiological studies, including a meticulous cross-cultural adaptation of research agendas, is reinforced.

KEY WORDS: Cross-cultural comparison. Semantic differential. Translating. Epidemiologic measurements. Diagnostic techniques and procedures. Validation studies.

INTRODUCTION

Epidemiological studies with explicative intentions (determinants, risk or protective factors, etiological factors and others) tend to use multi-dimensional questionnaires. These questionnaires often consist of different modules covering one or more constructs (dimensions*) of the theoretical model to be tested. In this sense, each construct involves an epidemiological instrument that needs to be incorporated into the main questionnaire.

The first task in setting up a modular questionnaire consists of a detailed literature review involving examination of research programs relating to the available instruments for each of the constructs of interest. A historical review of each possible instrument should include close scrutiny of the level of previous use and in particular, an evaluation of the stage of development of the research program. For this, it is crucial to examine all the available evidence regarding the adequacy and sufficiency of the psychometric development of each instrument. This scrutiny indicates to the researcher whether there really are satisfactory instruments for picking up the object to be studied. If instruments developed and consolidated in other cultural contexts are identified, it is also important to investigate whether they have already undergone a formal cross-cultural adaptation (CCA) process.

Departamento de Epidemiologia. Instituto de Medicina Social. Universidade do Estado do Rio de Janeiro. Rio de Janeiro, RJ, Brasil

Correspondence:

Michael Eduardo Reichenheim
R. São Francisco Xavier, 524, 7º andar
20559-900 Rio de Janeiro, RJ, Brasil
E-mail: michael@ims.uerj.br

Received: 1/8/2007
Approved: 4/23/2007

* *Constructs* and *dimensions* are distinguished from each other through the understanding that a construct may be composed of several dimensions. By extension, it must be understood that the empirical representation of a dimension is its *scale* and, in turn, the underlying numerical ordering of the scale is its *score*.

The researcher can then decide, for a given construct, whether it worth unconditionally accepting the instrument; or, if what was identified is insufficient or nonexistent, whether there is a need to construct a new instrument;⁵³ or further, whether it is necessary to start a complementary research program for CCA regarding one or more of the instruments deemed to be of merit. Differences in definitions, beliefs and behavior relating to a construct to be used in epidemiological research make it necessary for instruments developed in other cultural contexts to be preceded by meticulous evaluation of the equivalence between the original and its adapted version. The need to adapt measurement tools is not restricted to situations involving different countries and languages. Local and regional adjustments also require attention. It is difficult to decide whether fine-tuning has been achieved or not with regard to the population among which the translated version will be applied. The decision regarding the need for adaptation should take into consideration how much can be gained from the cultural adaptation and how much will be lost in terms of generalization and comparability. In countries with heterogeneous cultural roots like Brazil, proposals for typical colloquial terms that are well accepted and understood in one region or state of the country may not be pertinent in another. Moreover, cross-cultural adaptations are not necessarily restricted to space. Linguistic changes occur in the same population over time and, therefore, temporal adaptations are possible and sometimes necessary.

The lack of rigor in using measurement tools developed in other settings is a problem to be addressed. It is not uncommon for a researcher to informally translate an instrument or alter the number and content of its constituent items. While possibly well intentioned, failure to fine-tune the terminological choices to the target population, or inclusion of new items and/or deletion of others without subsequently implementing rigorous tests, may overly compromise the quality of the information.⁴¹ At worst, it may even prevent comparison of samples and studies on the same subject.

PROCESS OF CROSS-CULTURAL ADAPTATION

Historically, the adaptation of instruments developed in another culture and/or language was limited to a simple translation from the original, or exceptionally, to literal comparison of the original with a back-translation. For some time, researchers working in different fields have been suggesting that semantic evaluation constitutes only one of the steps needed for CCA.^{2,5,7,8,24,27,43} They have recommended that this process should be a combination of a literal translation of words and sentences from one language to another and a meticulous process of fine-tuning that takes into consideration the cultural context and lifestyle of the target population of the translation.^{4,24,28}

There are several articles in the literature with excellent accounts of theoretical approaches and practical proposals, which on the whole have this expanded vision.^{3,4,9,19,24,28,35,46,59} Nonetheless, there is no consensus regarding the strategies to use, which makes operational synthesis a mosaic of procedures originating from diverse sources. However, driven by the present article authors' own practices, choices are made using one of the possible models as a guide.^{27,28}

The proposal of Herdman et al,²⁸ which was developed and refined in relation to quality-of-life tools, is based on an interesting review in which the authors identify a plethora of terminology found in the literature and the confusion that the consequent overlapping of this terminology generates among researchers in this field.²⁷ The first of these two important articles also points out four perspectives that tend to govern CCA research programs.²⁷ The first, termed "naïve," is based only on a simple and informal process of translation of the original instrument. The second, termed "relativist," maintains that it is impossible to use standardized instruments in different cultural contexts and proposes that only those developed locally should be used. In this case, the notion of equivalence is not pertinent and, by extension, there is no possibility for interlocution. The third perspective, termed "absolutist," assumes that culture has a minimal impact on the constructs to be measured and that these do not change in different contexts. Methodologically, the emphasis is all on the process of translation and back-translation of the instrument. The last perspective, termed "universalist", does not assume, *a priori*, that the constructs are the same in different cultural contexts. Thus, it is first necessary to investigate whether the concept effectively exists or whether it is interpreted similarly in the new culture, so as to later establish its cross-cultural equivalent through suitable methodology.

In a subsequent article published in 1998, Herdman et al²⁸ proposed a basic guide. Assuming the "universalist" stance, they presented an evaluation model for the CCA process that included an assessment of the equivalence between the original instrument and the adaptation. Definitions and details are offered with respect to six types, namely, conceptual, item, semantic, operational, measurement and functional equivalence.

Following this, the present article suggests an operational system for using instruments developed in other linguistic-sociocultural contexts. This article was motivated by the perspective that there is an interest in comparing epidemiological profiles and findings from research conducted in different locations and cultures. Another reason for the present study was the relative lack of structured texts in Portuguese regarding "what and how to" carry out CCAs. This is a real gap given the recent, yet growing presence of studies of this

Table. Main stages involved in evaluating the cross-cultural equivalence of measurement instruments.

Aspect evaluated	Evaluation strategy
Conceptual equivalence	<ul style="list-style-type: none"> • Literature review involving publications on the culture of the original instrument and the target population • Discussion with specialists • Discussion with target population
Item equivalence	<ul style="list-style-type: none"> • Discussion with specialists • Discussion with target population
Semantic equivalence	<ul style="list-style-type: none"> • Translations • Back-translations • Evaluation of the semantic equivalence between the back-translations and the original • Discussion with target population • Discussion with specialists for final adjustments • Pretest of the translation
Operational equivalence	<ul style="list-style-type: none"> • Evaluation by research group regarding the pertinence and adequacy of: <ul style="list-style-type: none"> - Layout and format of the questions/instructions - Application setting - Application mode - Categorization mode
Measurement equivalence	<ul style="list-style-type: none"> • Psychometric studies <ul style="list-style-type: none"> - Focus 1: Evaluation of dimensional validity and adequacy of component items - Focus 2: Evaluation of reliability - Focus 3: Evaluation of the construct validity and criterion validity
Functional equivalence	<ul style="list-style-type: none"> • Provided by the equivalencies identified in the other evaluation stages

type in the Brazilian public health literature. A search in the Scientific Electronic Library Online (SciELO) using the key words [*questionnaire or instrument*] and [*adaptation or translation or reliability or validity*] and filtering for titles involving CCA of instruments to be directly applied (verbally), 121 Brazilian articles were identified. Of these, 36 were in journals specific to public health* and, with the exception of one from 1999, all were published in the current decade. Some of these Brazilian experiences have been guided by the proposal of Herdman et al, either totally^{39,40,47,51,52} or partially.^{1,11,16,21-23,26,45,54,64}

PROPOSAL FOR PUTTING CCA INTO OPERATION

A synthesis of the CCA evaluation process is summarized in the Table. Each of the steps needed for appreciating the different aspects of equivalence is detailed in the following. There is no elaboration on the functional equivalence, since this represents a synthesis of what came before, as defined by the proponents of the model. This synthesis depicts whether the efficiency of an instrument is equally acceptable across two or more cultures.²⁸

Conceptual and item equivalence

An evaluation of conceptual equivalence consists of exploring the construct of interest and the weights given to

its different constituent domains in the location (country, region, city) of origin and in the setting of the target population where the instrument will be used. As shown in the Table, in general, this stage involves a discussion with a group of specialists. This has the aim of exploring whether the various domains covered by the original instrument in defining the concepts of interest would be relevant and pertinent to the new context for which it is being adapted. In this process, the pertinence of the items for picking up each of the domains is evaluated. The discussions take place in the light of a literature review that prioritizes publications on the processes involved in developing the source-instrument and the bibliographic material available in the local context. Selected members and individuals representative of the target population should be involved, either through individualized open interviews or through collective activities such as focus groups.^{14,32}

Semantic equivalence

Evaluation of semantic equivalence involves the capacity to transfer the meaning of concepts contained in the original instrument to the translated version, thereby giving rise to a similar effect among respondents in both cultures. The evaluation guide for this aspect of equivalence should involve several steps.²⁸ Returning to the Table, the process begins with a translation of the original instrument into the language of the target culture. It is suggested that two or more versions should

* Cadernos de Saúde Pública (20), Revista de Saúde Pública (15) and Revista Brasileira de Epidemiologia (1)

be obtained independently, so that ideally there are more options for defining the terms to be used in the translation that is to be tested. These versions are then translated back to the original by other translators, also independently. The profile of the translators also matters. Some authors recommend that the translation process should be conducted by professionals whose native language and culture are those of the place for which the translation is being done.^{24,28,46} For example, in the context of an instrument developed in England to be adapted for use in Brazil, the translations of the original to Portuguese should be conducted by Brazilians with a good understanding of English and the back-translations should be carried out by English people with a good understanding of Portuguese.

Next, a new bilingual translator formally evaluates the equivalence between the back-translations and the original instrument. In addition to being independent, this evaluation should also be blinded in relation to the translators and back-translators. Preferably, the forms that are presented to the professional should not indicate which “vignette” refers to the back-translation and which to the original. One way of achieving this is to randomize the order in which they appear. In the case of simultaneous evaluation of more than one back-translation, in addition to the form for each pair (containing the original and a back-translation), forms with pairs of back-translations should also be submitted, so that the evaluator is unable to identify the original “vignette” among the group. Obviously, these forms are not actually analyzed; they just serve as “decoys.”

While Herdman et al²⁸ brought out various types of linguistic meaning for consideration, two deserve to be mentioned. The first point concerns an evaluation of the equivalence between the original and each of the back-translations with regard to the referential (denotative) meaning of the constituent words/terms, i.e. the ideas or subjects in the world to which one or several words refer. If the referential meaning of a word in the original and its respective translation are the same, it is presumed that a literal correspondence exists between them. The second point concerns the general (connotative) meaning of each item of the instrument, contrasting the original with what was picked up in the translation to the target language. This correspondence transcends the literalness of the words and also encompasses more subtle aspects, such as the impact that a term has within the cultural context of the target population. This assessment is necessary because the literal correspondence of a term does not imply that the same emotional or affective reaction is evoked in different cultures. It is absolutely necessary to fine-tune the instrument to achieve correspondence of perception and impact among respondents. This matter is particularly relevant in instruments created to empirically pick up concepts that are culturally constructed, since

a word or statement used with a given intention in its original context may not produce the same effect in the target population for the new version. Substitution of another term may enable full recovery of the desired equivalence. At this point, it is useful to return to the target population so that the subtleties brought out by the various translation proposals can be managed and debated. This fourth step can be achieved, for example, by going back to the focus groups.^{14,32}

The fifth step of the semantic evaluation involves the same group of specialists that took part in the conceptual and item equivalence evaluation stage and seeks to identify and address the problems from each of the previous activities. If possible, the team should be complemented by at least one of the translators involved previously, and preferably the one who was in charge of the formal comparison between the back-translations and the original instruments described above. With these obstacles overcome, a synthesis of the translations is proposed, which may incorporate items that arose from one of the translations produced or may select certain items that have been modified so as to meet the above criteria better.

The sixth and last step of the semantic equivalence evaluation stage involves a pretest. The compiled version of the instrument is applied to groups of individuals from the target population for a thorough evaluation of its acceptability, understanding and emotional impact. One technique to be used in the pretest is to ask the respondents to paraphrase each item, while the interviewer makes a note regarding whether the respondents understood the item referred to or not. A “series” of n interviews (e.g. 30 to 40) is conducted until a preestablished percentage adjustment (understanding) is achieved for all of the items (e.g. $\geq 90\%$). These interim evaluations can be conducted by the research team itself or, even better, by a group of experts brought together for this purpose. From the evidence found in this pretest, the final semantic adjustments to the compiled version are made, so that it can then be tested.

Operational equivalence

Operational equivalence refers to comparison between the characteristics of using an instrument in the target and source populations, such that there is efficacy even if the *modus operandi* is not the same. It is important to scrutinize the possible influences of certain characteristics of the instruments, such as the layout and format of the questions/instructions (e.g. on printed paper or in electronic format); the application setting (e.g. within a hospital or at home); and the way it is applied (e.g. face-to-face interviews or self-applied). The specification equivalence of the “outcome space”,⁶⁶ i.e. the scalability of each item, should also be taken into account. Therefore, it is important to note how the

item is categorized and the possible repercussions from choosing particular modifications. For example, one perfectly appropriate modification in situations where an instrument is applied with others in a multidimensional questionnaire, but in which the time envisaged for its application has to be short, would be to transform items that were originally proposed as five levels (Likert⁶² scale) into dichotomized items (0/1).

Operational changes often result from the circumstances in which the instrument should or can be used, and are not of the researcher's choosing. Thus, from an action perspective, evaluations of operational equivalence between application situations found when conceiving the instrument in the source-culture and those predominating in the target culture require an initial, eminently qualitative evaluation regarding the possibilities of success. The groups of experts that were brought into previous stages can also be consulted in these discussions.

Once a consensus regarding the viability and adequacy of one or more action strategies has been established, these strategies are incorporated into the study that will underlie the psychometric analyses to be implemented in the measurement equivalence stage. In this respect, it is the "hard" evidence explored in the psychometric analysis and the possible discrepancies between competing operational proposals that will either corroborate or refute the adaptation premises initially proposed by the specialists. Clearly, evidence of psychometric equivalence between the original and the version under scrutiny also positively attests to the operational adequacy of the instrument and, by extension, affirms its operational equivalence.

Measurement equivalence

As previously mentioned, measurement equivalence is based on investigation of the psychometric properties of the translated instrument. At first glance, nothing differentiates the execution of a psychometric study from any classic epidemiological study, since both require the same rigor in their processes.^{49,50} In particular, it is of great interest to identify the domain of the population picked up in the study, thereby catching a glimpse of the extent to which the findings may apply to the population among which the instrument will actually be used.

Once the field stage has been planned and implemented, data analysis is the next stage. In the same way as could be proposed in evaluating the development of a new instrument, three psychometric focuses can also be suggested: evaluation of the dimensional structure, including adaptation of the component items; evaluation of information reliability, through a process using the scales under test; and evaluation of the validity of these scales in their diverse nuances.⁵³

The perspective of the present article nonetheless differs a little from what is vested in the process of creating an instrument. Bearing in mind that if it is really sought to establish (measurement) equivalence, the central focus is not so much on the magnitude of the psychometric estimates as such, but on systematic comparison between these and what was obtained in previous studies in their original language/culture. For example, by considering some aspect of reliability as a pointer towards the adequacy of the measurement equivalence, it is less important to consider the absolute values of an interclass correlation coefficient⁵⁶ than whether this comes close to what was found in the studies on which the original instrument was based. As indicated at the beginning of the text, a relatively high value would obviously be expected, since choosing a particular instrument for adaptation presupposes a positive psychometric history.

As shown in the Table, the first task is to explore the dimensional structure of the instrument and the adequacy of the component items. Multivariate methods are at the heart of the process. The dimension pattern uncovered in previous studies can be initially accessed through exploratory factor analysis (EFA),^{12,20,30,34,48,55,58} even though in a certain way a structure has already been previously postulated in terms of dimensionality and component items. Bearing in mind that the term "exploratory" describes the technique itself more than a substantive perspective in this case, the EFA is a good start for ensuring that confirmatory factor analysis (CFA)^{31,34,36,58} can subsequently be implemented on a firm basis. An EFA followed by a CFA not only helps in effectively exploring whether the conjectured multidimensional structure exists, but also allows exploration of item behavior *vis-à-vis* the foreseen scales.

In cases of scales formed by dichotomous and ordinal items, the psychometric properties of these items and the scales that they form are better accommodated in item response theory (IRT) models.^{10,15,18,25,37,57,62,63,65} These are a nonlinear form of CFA, from the perspective of generalized latent variable models.⁵⁸ In addition to the usual focus on item loadings, some other key questions should also be taken into account in each of the dimensional scales constituting the instrument. It is of interest to ratify the presence of combined scalability of the items and the discriminatory capacity of each item. The absolute and relative positioning along the continuum of the latent variable (factor/dimension) that the scale of the instrument aspires to pick up requires scrutiny to identify the presence (undesirable) or absence (desirable) of information gaps along the spectrum. Likewise, evaluating the level of information provided by the set of items across the scale range and the precision of the information along the continuum of the latent variable are also matters to be compared with the original instrument.

A simpler alternative when the items are dichotomous or ordinal is to use, respectively, tetrachoric and polychoric correlation matrices. These are obtained by transforming Gaussian matrices before factor analysis, which could be either exploratory (EFA) or confirmatory (CFA).^{17,*} The inadvertent use of Gaussian correlation matrices in these situations tends to result in model misspecification, which could lead to spurious results^{20,29,55} and a false judgment regarding the absence or presence of equivalence.

The second psychometric focus involves formal evaluations of the reliability of the scales (internal consistency, stability and intra or inter-observer reproducibility^{13,33,41,42,62}). The objective is to evaluate to what extent the scores of an instrument (i.e. of the component scales) are free from random error.⁴⁴ This serves not only to provide robustness regarding the quality of the study relating to the CCA process, but also as a further stage in the adaptation process. Over the long term, a series of studies using the instrument and consistently revealing good measurement (information) reliability, ends up also attesting to the inherent quality of the instrument.

Even if it were possible to sanction the dimensional structure, adequacy of the items (via IRT, for example) and reliability of the process using the adapted instrument, there would be no effective guarantee that the CCA had been successful without explicit evaluation of validity-related matters. In this respect, the perspective provided by Streiner & Norman⁶² should be emphasized. In this, the validity of an instrument is ultimately established through determining the adequacy of the theory that supports it.

Various strategies have been used. Construct validity studies are frequently used when there is no reference instrument (gold standard) for contrast. The relationships between the dimensions supposedly picked up by the different scales of the instrument are evaluated, as are the relationships with other concepts, attributes and characteristics connected with the general theory within which the construct under scrutiny is situated. Finding associations that were predicted or fine-tuned using previous evidence corroborates and reinforces the validity of the instrument and, with regard to the present article, the adequacy of the CCA process. The opposite may also be important if it is found that there is no relationship between the theoretical concepts manifested by the adapted scales and others that are recognizably outside the scope of the general theory involving the phenomenon of interest.

While not necessarily ruling out an investigation of construct validity when a reference instrument, examination or test is available to contrast with the instrument

under scrutiny, it is also appropriate in such cases to evaluate the criterion validity (concurrent or predictive). From the principal viewpoint of the CCA for an instrument, evaluations of the ability of instrument scales to discriminate may be extremely enlightening. Knowing that an instrument applied in epidemiological studies not only picks up the continuum of an underlying latent variable, but also is substantially “linked” to what a reference examination or instrument would find, is clearly beneficial and attractive.

While the primary focus remains on measurement equivalence, problems and discrepancies that transcend the CCA may still show up in the process, thereby leading to discussions regarding a broader plan for instrument quality *per se*. Discerning these two issues is not always possible. For example, in the CCA for the Conflict Tactics Scales: Parent-Child,⁵² the item regarding the act of “slapping/punching the child on the face, head, or ears” was found to be much more related to acts of physical violence than to corporal punishment, although the latter was the dimension within which the authors of the instrument placed the item.^{60,61} It remains to be determined whether the measurement equivalence was effectively compromised because of this discrepancy in the connotative meaning between the two cultures, or whether the inadequacy was in the instrument originally proposed. In this latter case, the item in question would pertain to the dimension of physical violence, independent of the culture in question.

On finding inconsistencies, various possibilities from different perspectives should be considered. First, the quality of the adaptation should be questioned and faults in one or more stages of the process should be sought. Nonetheless, it is necessary to bear in mind some details in the interpretation. Focusing on reliability, for example, lower estimates than those found in the original do not necessarily indicate problems. Reliability is a circumstantial indicator that reflects both the quality of the measurement (presence or absence of measurement error) and the variability of an event in the population base studied.³⁸ Issues relating to study domains as objects of comparison should also be debated. Specific population differences among studies, such as the respondents’ level of schooling, gender, and age range may interfere in the performance of an instrument. Psychometric discrepancies do not necessarily mean that there is an important failure in the adaptation process as such, and the results warrant case-by-case debate.

FINAL CONSIDERATIONS

In closing their text on the pillars that sustain the validity of epidemiological studies, Reichenheim & Moraes⁴⁹

* Uebersax JS. The tetrachoric and polychoric correlation coefficients [monograph on internet]. [S.l.]; 2006. Available at: <http://ourworld.compuserve.com/homepages/juebersax/tetra.htm>.

suggested that many of the potential methodological obstacles that dominate such studies go completely beyond the idea of “objective truth.” The authors therefore proposed the notion of “constructed truth” as a basis on which to affirm knowledge. They pointed out how the “construction” of knowledge becomes evident when the extent to which the specification validity of a statistical model needs a theoretical framework for its implementation is perceived. In turn, this theoretical model tends to grow and gradually consolidate over the course of an iterative process of theory and experimentation. Rigor in methods is at the core of this argument when planning and carrying out an epidemiological study.

Likewise, it is perhaps not an overstatement to suggest that the quality of information is the central tie between theory and empirical, and that for this reason the rigor adopted in measurement processes should occupy a privileged position in epidemiological investigative practices. And this is not only a matter of being careful with data collection, even though this stage concludes all the preparation. Attentive examination of the instrument is equally prudent, while always keeping the perspective that there should be investment in formal adaptations of instruments drawn up in other contexts. As previously mentioned, this question becomes essential if there is an interest in comparing results from epidemiological research conducted in different settings and cultures.

REFERENCES

1. Avanci JQ, Assis SG, Santos NC, Oliveira RVC. Escala de violência psicológica contra adolescentes. *Rev Saude Publica*. 2005;39(5):702-8.
2. Badia X, Alonso J. Re-scaling the Spanish version of the Sickness Impact Profile: an opportunity for the assessment of cross-cultural equivalence. *J Clin Epidemiol*. 1995;48(7):949-57.
3. Beaton DE, Bombardier C, Guillemin F, Ferraz MB. Guidelines for the process of cross-cultural adaptation of self-report measures. *Spine*. 2000;25(24):3186-91.
4. Behling O, Law KS. Translating questionnaires and other research instruments: problems and solutions. Thousand Oaks: Sage; 2000.
5. Berkanovich E. The effect of inadequate language translation of Hispanics' responses to health surveys. *Am J Public Health*. 1980;70(12):1273-6.
6. Bowling A. Research Methods in Health. Investigating Health and Health Services. Buckingham: Open University Press; 1997.
7. Bravo M, Canino GJ, Rubio-Stipec M, Woodbury-Farina M. A cross-cultural adaptation of a psychiatric epidemiologic instrument: the diagnostic interview schedule's adaptation in Puerto Rico. *Cult Med Psychiatry*. 1991;15(1):1-18.
8. Bucquet D, Condon S, Ritchie K. The French version of the Nottingham Health Profile. A comparison of items weights with those of the source version. *Soc Sci Med*. 1990;30(7):829-35.
9. Bullinger M, Anderson R, Cella D, Aaronson N. Developing and evaluating cross-cultural instruments from minimum requirements to optimal models. *Qual Life Res*. 1993;2(6):451-9.
10. Cella D, Chang CH. A discussion of item response theory and its application in health status assessment. *Med Care*. 2000;38 (9 Supl):1166-72.
11. Chor D, Griep RH, Lopes CS, Faerstein E. Medidas de rede e apoio social no Estudo Pró-Saúde: pré-testes e estudo piloto. *Cad Saude Publica*. 2001;17(4):887-96.
12. Comrey AL, Lee HB. A first course in factor analysis. Hillsdale: Lawrence Erlbaum; 1992.
13. Cronbach LJ. Coefficient alpha and the internal structure of tests. *Psychometrika*. 1951;16(3):297-334.
14. Dawson S, Manderson L, Tallo VL. The Focus Group Manual. Methods for Social Research in Tropical Diseases. Geneva: World Health Organization; 1992.
15. De Boeck P, Wilson M. Explanatory item response models: a generalized linear and nonlinear approach. New York: Springer-Verlag; 2004.
16. Alves MGM, Chor D, Faerstein E, Lopes CS, Werneck GL. Versão resumida da job stress scale: adaptação para o português. *Rev Saude Publica*. 2004;38(2):164-71
17. Divgi DR. Calculation of the tetrachoric correlation coefficient. *Psychometrika*. 1979;44(2):169-72.
18. Embretson SE, Reise SP. Item response theory for psychologists. New Jersey: Lawrence Erlbaum Associates; 2000.
19. Eremenco SL, Cella D, Arnold BJ. A comprehensive method for the translation and cross-cultural validation of health status questionnaires. *Eval Health Prof*. 2005;28(2):212-32.
20. Gorsuch RL. Factor analysis. 2. ed. Hillsdale: Lawrence Erlbaum; 1983.
21. Grassi-Oliveira R, Stein LM, Pezzi JC. Tradução e validação de conteúdo da versão em português do Childhood Trauma Questionnaire. *Rev Saude Publica*. 2006;40(2):249-55.
22. Griep RH, Chor D, Faerstein E, Lopes CS. Apoio social: confiabilidade teste-reteste de escala no Estudo Pró-Saúde. *Cad Saude Publica*. 2003;19(2):625-34.
23. Griep RH, Chor D, Faerstein E, Werneck GL, Lopes CS. Validade de constructo de escala de apoio social do Medical Outcomes Study adaptada para o português no Estudo Pró-Saúde. *Cad Saude Publica*. 2005;21(3):703-14.
24. Guillemin F, Bombardier C, Beaton D. Cross-cultural adaptation of health-related quality of life measures: literature review and proposed guidelines. *J Clin Epidemiol*. 1993;46(12):1417-32.
25. Hambleton RK, Swaminathan H, Rogers HJ. Fundamentals of item response theory. Newbury park: Sage; 1991.
26. Hasselmann MH, Reichenheim ME. Adaptação transcultural da versão em português das "Conflict Tactics Scales Form R" (CTS-1) usada para aferir violência no casal: equivalências semântica e de mensuração. *Cad Saude Publica*. 2003;19(4):1083-93.
27. Herdman M, Fox-Rushby J, Badia X. "Equivalence" and the translation and adaptation of health-related quality of life questionnaires. *Qual Life Res*. 1997;6(3):237-47.
28. Herdman M, Fox-Rushby J, Badia X. A model of equivalence in the cultural adaptation of HRQoL instruments: the universalist approach. *Qual Life Res*. 1998;7(4):323-35.
29. Jöreskog KG, Sörbom D. LISREL 8 User's Reference Guide. Chicago: Scientific Software International; 1996.
30. Kline P. An easy guide to factor analysis. New York: Routledge; 1994.
31. Kline RB. Principles and practice of structural equation modeling. 2. ed. New York: The Guilford Press; 2005.
32. Krueger R. Focus Groups: a Practical Guide for Applied Research. 2. ed. London: SAGE; 1994.
33. Kuder GF, Richardson MW. The theory of estimation of test reliability. *Psychometrika*. 1937;2(3):151-60.
34. Loehlin JC. Latent variable models. An introduction to factor, path and structural equation analysis. 4. ed. Mahwah: Lawrence Erlbaum Associates; 2004.
35. Maneesriwongul W, Dixon JK. Instrument translation process: a methods review. *J Adv Nurs*. 2004;48(2):175-86.
36. Maruyama GM. Basics of structural equation modeling. Thousand Oaks: Sage; 1998.

37. Mellenbergh GJ. Generalized linear item response theory. *Psychol Bull.* 1994;115(22):300-7.
38. Miettinen O. Design options in epidemiologic research. An update. *Scand J Work Environ Health.* 1982;8(Supl1):7-14.
39. Moraes CL, Hasselmann MH, Reichenheim ME. Adaptação transcultural para o português do instrumento "Revised Conflict Tactics Scales (CTS2)" utilizado para identificar a violência entre casais. *Cad Saude Publica.* 2002;18(1):163-75.
40. Moraes CL, Reichenheim ME. Cross-cultural measurement equivalence of the Revised Conflict Tactics Scales (CTS2) Portuguese version used to identify violence within couples. *Cad Saude Publica.* 2002;18(3):783-96.
41. Nunnally J, Bernstein I. *Psychometric theory.* 2. ed. New York: McGraw-Hill; 1995.
42. Osburn HG. Coefficient alpha and related internal consistency reliability coefficients. *Psychol Methods.* 2000;5(3):343-55.
43. Patrick DL, Sittampalam Y, Somerville SM, Carter WB, Bergner M. A cross-cultural comparison of health status values. *Am J Public Health.* 1985;75(12):1402-7.
44. Pedhazur EJ, Schmelkin LP. *Measurement, design, and analysis: an integrated approach.* Hillsdale: Lawrence Erlbaum Associates; 1991.
45. Pereira LSM, Marra TA, Faria CD, Pereira DS, Martins MAA, Dias JMD, et al. Adaptação transcultural e análise da confiabilidade do Southampton Assessment of Mobility para avaliar a mobilidade de idosos brasileiros com demência. *Cad Saude Publica.* 2006;22(10):2085-95.
46. Perneger TV, Leplège A, Etter JF. Cross-cultural adaptation of a psychometric instrument: two methods compared. *J Clin Epidemiol.* 1999;52(11):1037-46.
47. Pesce RP, Assis SG, Avanci JQ, Santos NC, Malaquias JV, Carvalhaes R. Adaptação transcultural, confiabilidade e validade da Escala de Resiliência. *Cad Saude Publica.* 2005;21(2):436-48.
48. Pett MA, Lackey NR, Sullivan JJ. *Making sense of factor analysis: the use of factor analysis for instrument development in health care research.* Thousand Oaks: Sage; 2003.
49. Reichenheim ME, Moraes CL. Alguns pilares para a apreciação da validade de estudos epidemiológicos. *Rev Bras Epidemiol.* 1998;1(2):131-48.
50. Reichenheim ME, Moraes CL, Hasselmann MH. Equivalência semântica da versão em português do instrumento *Abuse Assessment Screen* para rastrear a violência contra a mulher grávida. *Rev Saude Publica.* 2000;34(6):610-6.
51. Reichenheim ME, Moraes CL. Adaptação transcultural do instrumento "Parent-Child Conflict Tactics Scales (CTSPC)" utilizado para identificar a violência contra a criança. *Cad Saude Publica.* 2003;19(6):1701-12.
52. Reichenheim ME, Moraes CL. Psychometric properties of the Portuguese version of the Conflict Tactics Scales: Parent-child Version (CTSPC) used to identify child abuse. *Cad Saude Publica.* 2006;22(3):503-15.
53. Reichenheim ME, Moraes CL. Desenvolvimento de instrumentos de aferição epidemiológicos. In: Kac G, Schieri R, Gigante D, organizadores. *Epidemiologia Nutricional.* Rio de Janeiro: Fiocruz; 2007. No prelo.
54. Rothman KJ, Greenland S. *Modern Epidemiology.* 2. ed. Philadelphia: Lippincott-Raven; 1998.
55. Rummel RJ. *Applied Factor Analysis.* 4. ed. Evanston: Northwest University Press; 1988.
56. Shrout PE, Fleiss JL. Intraclass correlations: uses in assessing rater reliability. *Psychol Bull.* 1979;86(2):420-8.
57. Sijtsma K, Molenaar IW. *Introduction to nonparametric item response theory.* Thousand Oaks: Sage; 2002.
58. Skrondal A, Rabe-Hesketh S. *Generalized latent variable modeling: multilevel, longitudinal, and structural equation models.* Boca Raton: Chapman & Hall/CRC; 2004.
59. Sperber AD. Translation and validation of study instruments for cross-cultural research. *Gastroenterology.* 2004;126(1Supl 1):S124-8.
60. Straus MA, Hamby SH, Finkelhor D, Moore DW, Runyan D. Identification of child maltreatment with Parent-Child Conflict Tactics Scales: development and psychometric data for a national sample of American parents. *Child Abuse Negl.* 1998;22(4):249-70. Errata em: *Child Abuse Negl.* 1998;22(11):1177.
61. Streiner DL, Norman GR. *Health measurement scales. A practical guide to their development and use.* 3. ed. Oxford: Oxford University Press; 2003.
63. Van der Linden WJ, Hambleton RK. *Handbook of Modern Item Response Theory.* New York: Springer; 1996.
64. Vilete LMP, Coutinho ESF, Figueira ILV. Confiabilidade da versão em Português do Inventário de Fobia Social (SPIN) entre adolescentes estudantes do Município do Rio de Janeiro. *Cad Saude Publica.* 2004;20(1):89-99.
65. Wilson M. *Constructing measures. An item response modeling approach.* Mahwah: Lawrence Erlbaum Associates; 2005.