Qual Life Res (2015) 24:115–124
DOI 10.1007/s11136-014-0722-8

QUANTITATIVE METHODS SPECIAL SECTION

Using quantitative methods within the Universalist model framework to explore the cross-cultural equivalence of patient-reported outcome instruments

Antoine Regnault · Michael Herdman

Accepted: 16 May 2014 / Published online: 4 June 2014

Abstract
Purpose The cross-cultural equivalence of patient-reported outcome (PRO) instruments is critical when they are used in international settings. The Universalist model of equivalence was proposed as a framework to investigate cross-cultural equivalence. The purpose of this paper was to illustrate how quantitative methods can be used to investigate cross-cultural equivalence within this framework.
Methods The six types of equivalence of the Universalist model were reviewed from a statistical perspective and statistical techniques allowing addressing the underlying question were identified. These methods are described and examples are provided of how they can be applied. An integrated pragmatic approach to the exploration of cross-cultural equivalence was developed based on these methods.
Results The statistical techniques identified were factor analysis to explore conceptual equivalence, differential item functioning to explore semantic and item equivalence, and comparison of measurement properties for the measurement equivalence. The statistical techniques addressing operational equivalence were found to be diverse and highly specific to the operational aspect under investigation. Functional equivalence involves a comprehensive appraisal of the potential impact of the results of the other equivalences on the conclusions of the research. This structured appraisal of functional equivalence offers a framework for a comprehensive, but flexible, approach for the efficient application of statistical analyses to explore cross-cultural equivalence of PRO instruments.
Conclusion The different types of equivalence of the Universalist model can be investigated using quantitative methods. An integrated approach, which could be used in a variety of settings, was developed to allow the whole notion of cross-cultural equivalence to be comprehensively and efficiently addressed.

Keywords Questionnaires · Cross-cultural equivalence · Universalist model of equivalence · Confirmatory factor analysis · Differential item functioning

A. Regnault (✉)
Mapi, HEOR and Strategic Market Access, 27, rue de la Villette, 69003 Lyon, France
e-mail: aregnault@mapigroup.com

M. Herdman
Insight Consulting and Research SL, Barcelona, Spain

Introduction

The globalization of clinical research has led to increasing use of patient-reported outcome (PRO) instruments in international settings. For this use to be appropriate, PRO measures should be adapted to this international context. In particular, their adequacy to a cross-cultural setting should be demonstrated.

The notions and concepts involved in cross-cultural equivalence in the PRO field [1–4] were largely inspired by research in cross-cultural psychology [5]. Reviews highlighted a large number of different types of equivalence in the literature [6] with authors occasionally defining the same type of equivalence in different ways [7]. In order to provide a framework for exploring issues related to cross-cultural equivalence, a model based on the Universalist approach developed in cross-cultural psychology [8] was proposed [9]. It suggests that six types of cross-cultural equivalence need to be addressed for an instrument to be

claimed as cross-culturally valid, those being conceptual equivalence, item equivalence, semantic equivalence, operational equivalence, measurement equivalence, and functional equivalence. This model has been largely cited and used as a theoretical framework for cross-cultural PRO research since it has been introduced.

From an operational perspective, methods to adapt PRO instruments into different languages for use in different cultural settings are now well defined and largely used [10–13]. These linguistic validation methods consist of a series of steps including forward translations, backward translation, and patient testing. The aim of this process is to achieve equivalent versions of the instruments which will allow for pooling and comparison of data across countries. Nevertheless, these adaptations should only be the first step towards achieving and testing cross-cultural equivalence. Once data have been collected using different language versions of an instrument, quantitative methods can also provide information on the degree to which equivalence has been achieved. Hence, quantitative methods to test cross-cultural equivalence have been implemented in the PRO field [14–23]. However, contrary to the linguistic validation methods, which are now fairly standardized, the quantitative approaches to cross-cultural equivalence are diverse, generally focus on one specific aspect of cross-cultural equivalence and use a heterogeneous set of statistical techniques. This apparent complexity reflects the multifaceted notion of cross-cultural equivalence. The quantitative assessment of cross-cultural equivalence of PRO measures would therefore likely benefit from being encompassed in a theoretical framework. Hence, the different aspects of cross-cultural equivalence, with the hypotheses to be tested and the available statistical tools to address them, would be clearly identified and researchers intending to quantitatively assess cross-cultural equivalence of a PRO measures would have a guide in this endeavour.

The Universalist model offers a well-defined theoretical framework of cross-cultural equivalence that could constitute an appropriate basis for the organization of quantitative approaches to cross-cultural equivalence. The objective of this paper is to show how quantitative methods can be used to explore the different types of equivalence in the Universalist model and to propose a pragmatic approach for its application as a guide for quantitative assessment of cross-cultural equivalence of PRO measures. To do this, we review each type of equivalence in the Universalist model from a statistical perspective and discuss quantitative methods that could be applied to explore that type of equivalence. We provide practical examples of the approaches proposed and provide suggestions as to how to proceed if results indicate that particular types of cross-cultural equivalence may not have been achieved. It is hoped that this will provide useful strategies for investigators planning cross-cultural research and help avoid some potential pitfalls.

Conceptual equivalence

According to the Universalist model, it cannot be assumed that the concept(s) of interest exist, are equally relevant, and share the same structure across different cultures. Testing conceptual equivalence therefore means evaluating whether a concept, such as well-being for example, exists in all of the cultures of interest and whether it is constructed in the same way across those cultures. In developing the Universalist model for application in the PRO field, the authors noted that qualitative research would be of vital importance in answering these questions [9]. Nonetheless, statistical methods may also be useful in assessing conceptual equivalence, particularly those focusing on the scale structure of the measurement instrument. Indeed, the scale structure (i.e. the number of domains of the questionnaire and how the items of the questionnaire are grouped to assess these domains) of an instrument is designed to reflect the underlying concepts of interest and the relationships among them. So, if this structure appears to be similar in the different cultures then it implies that the underlying conceptual patterns are shared across the different populations. Factor analysis can provide insights into whether the underlying structure of a questionnaire is maintained across different cultures and thus can help elucidate whether there may be similarities in the underlying concepts in the different cultures. Using multi-sample confirmatory factor analysis (CFA) may be especially useful in evaluating conceptual equivalence [24–27]. This approach provides tools to test the extent to which theoretical models fit the observed data. The theoretical models tested can be specified to examine the similarity between cultures of the pattern linking items to dimensions, of the loadings of items to dimensions, and of the relationships between measured concepts. After the model is specified, it is applied to the observed data and a set of statistical indicators can be examined to make a decision regarding the appropriateness of the model.

If a similar theoretical model does not provide an adequate fit for data collected in the different cultures, the analysis can be extended to an exploratory factor analysis to identify differences in conceptual structure between cultures. If the exploratory factor analysis shows important differences in the factor structure of the instrument in the different cultures, it would be inadvisable to assume that the instrument is cross-culturally equivalent and pooling or statistical comparison of results from different countries might not be warranted. Then, qualitative research can be

Fig. 1 Hypothesized factor structure of the Psychological General Well-Being Index (PGWBI) [28]. The PGWBI was designed for use in a US population; the factor structure was confirmed in the French sample from the IQOD database, but not in the Japanese sample

useful in understanding the reasons behind the lack of equivalence and, together with the results of quantitative analysis, may enable the identification of core domains which are relevant across cultures for the concept under study.

An example of this approach was the application of CFA to data from the International Quality of life Outcome Database (IQOD) which allowed us to test the applicability of the original (US English) structure (Fig. 1) of the Psychological General Well-Being Index (PGWBI) in the French and Japanese versions [28]. In the case of the French version, the model fit was acceptable (for example, root mean square of approximation was equal to 0.05), while the structure fit was poorer for the data from the Japanese version of the questionnaire (root mean square of approximation of 0.09), indicating that the original structure might not be appropriate for Japan [29]. Further analysis showed that the Japanese version was likely to include different dimensions; a preliminary exploratory factor analysis revealed a 4-factor structure, compared to the original 6 dimensions, and noticeably different item grouping. Thus, the structural model of well-being employed in the PGWBI appeared to be appropriate in France and the USA but not in Japan. Though these were

preliminary results, if confirmed they would suggest that, at the very least, results obtained with the PGWBI in Japan should not be pooled or compared with results obtained in the USA or France, and possibly that a different approach to measuring psychological well-being would be required in Japan. Qualitative research and a review of the literature on the concept of well-being in Asian culture [30] might be useful to better understand the differences in results.

Item equivalence and semantic equivalence

According to the original Universalist model, item equivalence was defined as the extent to which a given item is an appropriate measure of the concept it is assumed to measure in the different cultures and semantic equivalence involves the exploration of the item's connotations for speakers of different languages. Both are therefore related to the responses of patients to individual items. Quantitative methods allowing the detection of systematic differences in the way comparable respondents answer across different versions correspond to a single concept commonly used in cross-cultural research: differential item functioning (DIF). DIF refers to a difference in the expected response of individuals who are comparable for the construct being measured but who belong to different groups (e.g. gender, age, or culture) [31].

This approach has been the object of extensive research, and although DIF is often associated with modern test theory [32–36], it can be investigated using a range of quantitative methods [31, 37, 38] from analyses of contingency tables (Cochrane–Mantel–Haenszel test, nonparametric measures of association [14, 39] or log-linear models [14, 29, 40]), to logistic regression [41–44]. Combinations of these approaches can also be used. For example, modern test theory and logistic regression have been used in combination [45, 46]. Comparisons of the various methods available to detect DIF have been undertaken [31, 40, 41, 47, 48], but no method has been shown to be universally preferable to the others.

Once an item has been flagged as showing DIF, possible sources of DIF should be investigated. It may be due to reasons related to item equivalence (an item has a different relationship with the concept of interest in one of the cultures) or semantic equivalence (e.g. errors in translation or unsuspected connotations of terms used in the question in certain languages). A review of the item by experts (e.g. linguists, sociologists) or bilingual individuals may help to detect the source of DIF [49] and the cognitive debriefing exercises which are usually carried out when cross-culturally adapting a questionnaire could also provide information as could reviews of the meaning of keywords in dictionaries, thesauruses, and other relevant sources.

If the potential cause of the DIF is found, the item could be modified to address it or, if this is impossible, the item could be removed from the questionnaire. However, the decision to delete an item from a questionnaire should not be taken lightly. Items have usually been included in a questionnaire for a good reason and eliminating an item can affect content validity and possibly psychometric performance. The explicit approval of the developer of the original questionnaire should be sought and the consequences on content validity, psychometric performance, and conceptual equivalence considered. Deleting an item also has technical implications. In particular, it will impact the calculation of scores: the scoring algorithm of the modified instrument will require careful adjustment. In this context, the use of modern test theory models has substantial advantages because they allow comparable measures to be obtained from different sets of items.

As an example of how semantic and item equivalence can be explored, a DIF detection procedure based upon ordinal log-linear models was applied to the French and US English versions of the PGWBI from the IQOD. This analysis flagged item 22 of the PGWBI as an item with DIF in the IQOD data [29]. The item explores stress and pressure perceived by the respondent. So at a given level of anxiety, French respondents gave much higher responses to item 22 than American respondents. Whether this was due to a difference in the way people interpreted the item (semantic equivalence) or to a difference in the relationship between the notion of stress and the general notion of anxiety (French people expressing a higher level of stress at a given level of anxiety than Americans regardless of the formulation of the question) was unfortunately not investigated in the original research.

Another very rich example of how DIF could be used to investigate semantic and item equivalence can be found in a series of paper investigating the items of the EORTC QLQ-C30 [19, 20, 50]. This research investigated DIF between 13 language versions of the questionnaire using logistic regression applied to a huge dataset (more than 28,000 observations). Importantly decisions on whether an item was functioning differentially did not rely solely on statistical significance but also on the magnitude of differences and on qualitative insight from interviews with bilingual individuals. Also the expected impact of the differential responses between different versions of the questionnaire on the final study results was considered in the interpretation of the DIF results. In addition, the comparison was performed both between language versions of the questionnaire and between cultures (i.e. grouping different language versions into homogeneous cultural groups), hence allowing an interesting discussion about the separation of semantic and item equivalence. Thus, this body of research can be seen as an extensive investigation

of the item and semantic equivalence of the EORTC QLQ-C30 questionnaire.

Operational equivalence

Operational equivalence refers to whether the methods used to actually collect the data are equally appropriate in different cultures. In the original paper, it was defined as referring to 'the possibility of using a similar questionnaire format, instructions, mode of administration, and measurement methods'. These are therefore mainly related to technological issues (e.g. use of electronic devices in different cultures) and normative issues (e.g. openness with which topics are discussed or ways opinion are given).

The method used to document a specific aspect of operational equivalence naturally needs to be adapted to the aspect in question: assessing the impact of cultural norms does not require the same methods as assessing differences in the impact of mode of administration between cultures. Because of the diversity of aspects encompassed when assessing operational equivalence, the methods required are therefore heterogeneous and cannot be easily characterized. Moreover, for a single aspect of operational equivalence, various approaches may be possible.

For instance, with regard to response scales, which is perhaps one of the most widely studied aspects of operational equivalence, various methods have been used. For instance, the Thurstone scaling method was applied to compare 13 translations of the response choice labels of the SF-36 in the framework of the IQOLA project [16] while other authors have focused on the response style of individuals from different cultures to various response scales [51–53]. While the former approach tended to support the comparability of the response scales of the SF-36 across the different versions, the latter showed some relationships between cultural orientations and response styles (e.g. masculinity, one of the cultural orientation studied, is statistically significantly associated with extreme response style or Uncertainty avoidance is associated with Acquiescence) [51]. The two approaches are related as they focus on the interaction of respondent and response scale but they used different perspectives and different quantitative tools.

Thus, the methods used to study the impact of instrument format, mode of administration, and other operational methods to collect a response will depend on the design and aim of the study. Both qualitative and quantitative methods could be used in such studies; the most important criterion will be ensuring that the methods used are appropriate to the goals and context of the study.

Measurement equivalence

The fifth type of equivalence in the Universalist model, measurement equivalence, relates to whether the different versions of the questionnaire have acceptable measurement properties. Measurement equivalence should also refer to the comparability of the measurement properties of the instruments in the different language versions. The statistical methods to address this type of equivalence are probably the most straightforward as they pertain to the assessment of commonly used psychometric properties, i.e. reliability, validity, and ability to detect change over time. Consequently, this type of equivalence has seen considerable attention [54].

However, comparisons have usually been limited to a qualitative comparison of the obtained values (e.g. [55, 56]), whereas techniques exist that would enable a more rigorous comparison of psychometric properties. When evaluating the reliability of the instrument, Cronbach's alphas could be compared between different language versions of the instruments using a statistical test [57, 58] and intra-class correlation coefficients used for the assessment of test–retest reliability can also be quantitatively compared across cultures using the linear mixed model framework [59]. As for the analysis of validity of the instrument, many analyses may actually be related to the questions addressed in the conceptual equivalence part since validity is related to the question of the relationship of the measure to the concept being assessed. However, it can be imagined that culture be a confounding factor tested in the analyses investigating the relationships between the scores and other parameters. For instance, when instrument scores are compared across different severity groups to determine clinical validity, culture could be included as a covariate in the comparison to determine whether the relations between the score and the severity are similar across cultures (then an ANCOVA would be performed instead of an ANOVA). Finally, the effect sizes characterizing the ability to detect change can be easily compared across versions since they are designed to assess the magnitude of change on a common metric. More rigorous comparisons of measurement properties should be encouraged because such an approach provides evidence on the comparability of properties, and therefore on measurement equivalence. If, for a particular questionnaire, it can be shown, on a sample large enough to provide sufficient statistical power, that there are no statistically significant differences between measurement properties for all different language versions, then those versions can be definitely considered as having achieved measurement equivalence.

Functional equivalence

The final level of the model is functional equivalence. The aim of this level is to answer the overall question: Does the instrument do what it is supposed to do equally well in the different cultures? The idea of including functional equivalence in the original model was to draw together the different types of equivalence described above and make an overall assessment of the results obtained in the various analyses performed to assess the five other types of equivalence, and to evaluate the potential impact of findings on the final conclusions of the study.

The level of functional equivalence required will depend on the context and objectives of the use of the instrument. For instance, the expected level of functional equivalence of an instrument used in an observational study specifically designed to compare a concept between cultural groups will be more demanding than that of an instrument used in a clinical trial where what is at stake is the change over time regardless of the cultural group, generally after randomization: in the former, the measurement should be directly comparable and the cultural group plays a key role in the analysis while in the latter, culture is more of a covariate and only the change in the measurement needs to be compared.

The notion of functional equivalence therefore allows the development of a pragmatic process that encompasses all types of equivalence. This process to investigate functional equivalence is schematically represented in Fig. 2. The first functional equivalence decision appears at the conceptual equivalence level and is straightforward: if the dimensional structure of the instrument is different between cultures, functional equivalence cannot possibly be supported. This rule is particularly meaningful since in such a case the construct being measured itself is different from one culture to another. This reminds us of the crucial importance of conceptual equivalence in the Universalist model. Then, a critical appraisal of the potential impact of the results should be made at all stages of the Universalist model. First, the implications of the results obtained by DIF detection techniques should be studied. Indeed, some authors have shown that the impact of DIF may be weak when aggregated at the test level [60, 61]. It can therefore be relevant to study to what extent the items flagged as functioning differentially could eventually bias the results of the full instrument. If there is no (or minor) risk of changing the study conclusions, then corrective actions for DIF would not be mandatory. Then, if the measurement methods used in the study are shown to affected by culture at the operational equivalence step, again the impact on the study conclusions should be evaluated and if this difference jeopardizes the results of the study, alternative methods that would be less sensitive to culture and that would allow reliable conclusions to be made should be searched for. Finally, the impact of potential differences in measurement properties on the study conclusions should be appraised. For instance, a difference in the ability to detect change would be of limited importance for cross-sectional comparisons but would be critical if the purpose of the study is to compare the change over time in the concept between the cultures.

In conclusion, the assessment of functional equivalence consists of a comprehensive appraisal of the statistical results obtained in the analysis of the other types of equivalence, bearing in mind the context of use of the instrument.

Discussion

The aim of this paper was to explore in a structured fashion the different statistical techniques that can be used to investigate equivalence between different language versions of the same questionnaire. We also aimed to highlight some of the situations that can arise in this context and provide recommendations to deal with them. By doing so, we aim to offer researchers who want to quantitatively assess the cross-cultural equivalence of PRO measures with useful operational guidance underpinned by solid theory.

Using a comprehensive theoretical model of equivalence as framework for the application of statistical methods has a number of advantages. Firstly, it reminds researchers using adapted versions of PRO instruments across cultures of the different types of equivalence that need to be checked. Indeed, a considerable amount of cross-cultural research in the PRO field has focused exclusively on a simple comparison of measurement properties. A literature review ascertained that measurement equivalence (i.e. the comparison of psychometric properties) was the only level of the model that had been extensively explored and deplored the lack of assessment of other types of equivalence of the Universalist model [54]. Secondly, having a well-defined theoretical framework helps to organize efficiently the various statistical methods available to address cross-cultural issues and emphasizes how these methods complement one another by addressing different aspects of cross-cultural equivalence. Thirdly, it shows the potential for a mixed methods approach to cross-cultural issues, combining quantitative and qualitative methods to address questions related to cross-cultural equivalence. For example, even if conceptual equivalence is primarily a qualitative issue addressed with qualitative methods, quantitative methods could help to formally validate the results of the qualitative work. In contrast, DIF detection is a quantitative method, but a qualitative interpretation of the results can help determine the reasons for the DIF and possibly how to

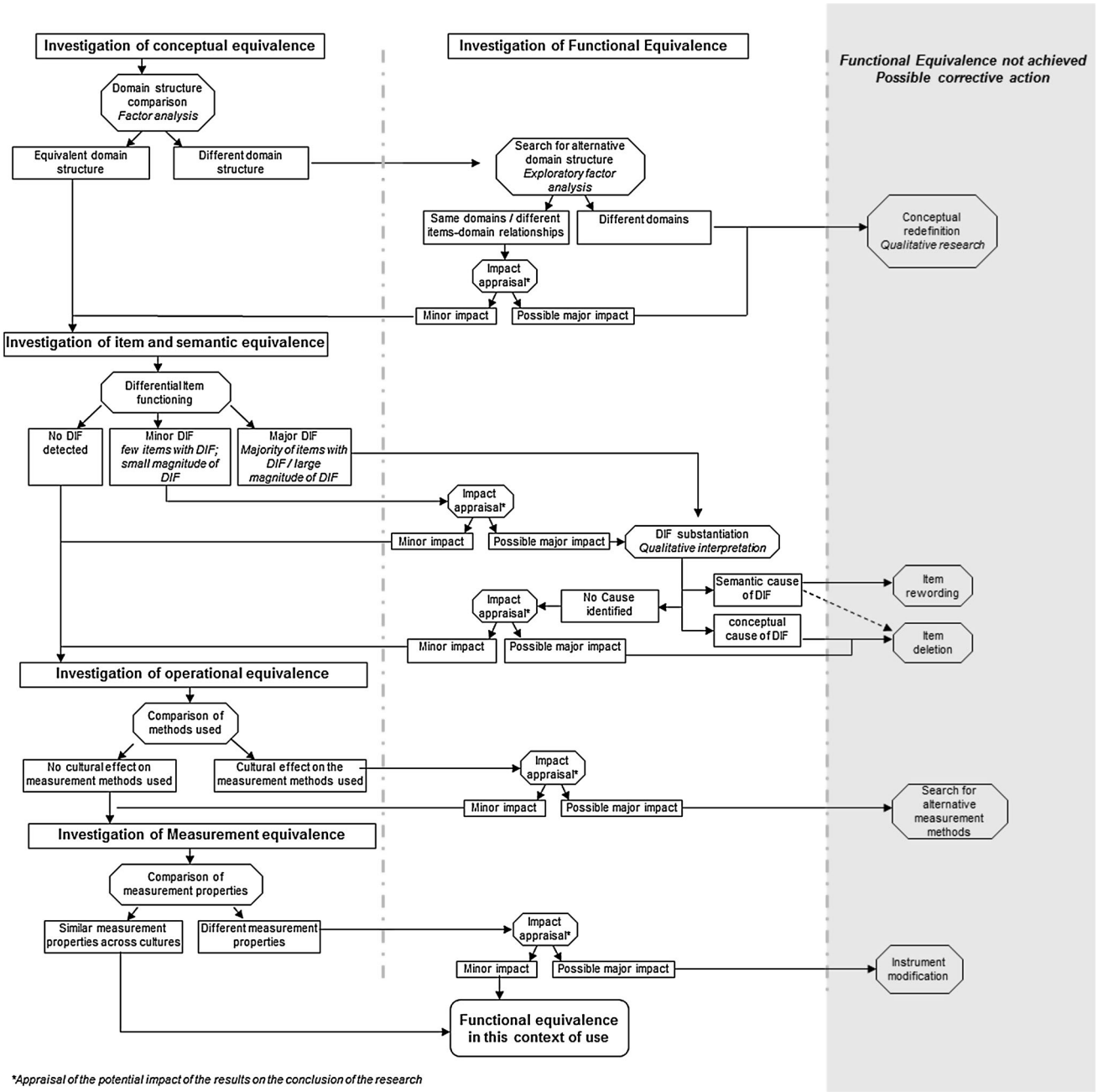

Fig. 2 Integrated approach for the application of statistical methods within the Universalist model of equivalence

address it. Finally, using a theoretical framework in this way can give a broad picture of how suitable a given instrument is for research in different cultural settings given specific study objectives.

One of the strengths of our approach is its flexibility. Indeed, it offers by design the possibility to be adjusted to the context and objective of the research. In particular, the phase of assessment of functional equivalence leaves some room for adaptation of the decisions made according to the context. These decisions should be based on a critical appraisal of the statistical results and their potential impact on the use of the instrument, in the specific context of interest. First, the objective of the study in which the questionnaire is used is certainly a critical element to consider when assessing the results of any step of the model. The aim of cross-cultural equivalence assessment can be to decide whether data from different language versions could be pooled to support treatment effect analyses in a multinational clinical trial; whether it is valid to compare the measured concepts between two different

cultures using the questionnaire in an international observational study; or to compare newly developed country norms to those of other countries that have been previously produced. The level of equivalence needed for these different purposes might well differ. Similarly, the study design can be a critical criterion in the appraisal. For instance, in a multinational clinical trial, if DIF is observed for several items of an instrument but only for a language version corresponding to a very small country sample, the impact is likely to be very limited, and this should be taken into account in the choice of the method used and interpretation of study results. It should be noted that, despite its versatility, the approach we propose still relies on statistical analyses, which require sufficiently large samples to be applied reliably. The techniques used could certainly be adapted to small samples but any decision made on samples with less than 30–50 patients per cultural group will have to be treated with caution, and larger samples are likely needed to support definitive decisions on the cross-cultural equivalence of PRO measures.

While the outcome of the impact appraisal can simply result from a careful interpretation of the results by experts (statisticians, linguists, sociologists), it would be more reliable if it can be supported by research demonstrating the potential impact of the results. For example, simulation or sensitivity analyses that would inform the potential impact of the differences observed between cultures on the final conclusion of the study would be very helpful to support the findings on cross-cultural equivalence. Finally, it is also important to be aware that the conclusions of this approach regarding functional equivalence of the instrument are valid in the specific context of use under scrutiny and cannot automatically be translated to other contexts of use (i.e. an HRQoL instrument judged appropriate for use in a multinational clinical trial may not be appropriate for an observational study designed to compare HRQoL across different countries).

The rapid growth in the use of PRO instruments in multicultural settings makes it important to try to justify their relevance and appropriateness for use in different cultural contexts. We hope that the present article will help to encourage the application of quantitative methods in the assessment of cross-cultural equivalence of PRO instruments by providing researchers in this context with a structured framework for their research.

Conclusion

Considering the Universalist model from a statistical perspective offers a clarification of the potential role of quantitative methods to explore cross-cultural equivalence. A variety of statistical methods are available to assist the assessment of cross-cultural equivalence. The use of a clear theoretical framework involving complementary information derived from both qualitative work and empirical quantitative methods will allow interpretation and identification of potential problems to be addressed and thereby enable better evaluation of the overall cross-cultural equivalence of different versions of a given instrument.

Acknowledgments We would like to thank Christine de la Loge for having initiated this work and for her participation in the earliest phase of this work.

References

1. Anderson, R. T., Aaronson, N. K., & Wilkin, D. (1993). Critical review of the international assessments of health-related quality of life. *Quality of Life Research, 2*(6), 369–395.
2. Bullinger, M., Anderson, R., Cella, D., & Aaronson, N. (1993). Developing and evaluating cross-cultural instruments from minimum requirements to optimal models. *Quality of Life Research, 2*(6), 451–459.
3. Hays, R. D., Anderson, R., & Revicki, D. (1993). Psychometric considerations in evaluating health-related quality of life measures. *Quality of Life Research, 2*(6), 441–449.
4. Schmidt, S., & Bullinger, M. (2003). Current issues in cross-cultural quality of life instrument development. *Archives of Physical Medicine and Rehabilitation, 84*(4 Suppl 2), S29–S34.
5. Hui, C. H., & Triandis, H. C. (1985). Measurement in cross-cultural psychology a review and comparison of strategies. *Journal of Cross-Cultural Psychology, 16*(2), 131–152.
6. Johnson, T. P. (2006). Methods and frameworks for crosscultural measurement. *Medical Care, 44*(11 Suppl 3), S17–S20.
7. Herdman, M., Fox-Rushby, J., & Badia, X. (1997). 'Equivalence' and the translation and adaptation of health-related quality of life questionnaires. *Quality of Life Research, 6*(3), 237–247.
8. Berry, J. W. (2002). *Cross-cultural psychology: Research and applications*. Cambridge: Cambridge University Press.
9. Herdman, M., Fox-Rushby, J., & Badia, X. (1998). A model of equivalence in the cultural adaptation of HRQoL instruments: The Universalist approach. *Quality of Life Research, 7*(4), 323–335.
10. Acquadro, C., Conway, C., Giroudet, C., & Mear, I. (2004). *Linguistic validation manual for patient-reported outcomes (PRO) instruments*. Lyon: Mapi Research Institute.
11. Behling, O., & Law, K. S. (2000). *Translating questionnaires and other research instruments: Problems and solutions*. Thousand Oaks: Sage.
12. McKenna, S. P., & Doward, L. C. (2005). The translation and cultural adaptation of patient-reported outcome measures. *Value Health, 8*(2), 89–91.
13. Wild, D., Grove, A., Martin, M., Eremenco, S., McElroy, S., Verjee-Lorenz, A., et al. (2005). Principles of good practice for the translation and cultural adaptation process for patient-reported outcomes (PRO) measures: Report of the ISPOR task force for translation and cultural adaptation. *Value in Health, 8*(2), 94–104.
14. Bjorner, J. B., Kreiner, S., Ware, J. E., Damsgaard, M. T., & Bech, P. (1998). Differential item functioning in the Danish translation of the SF-36. *Journal of Clinical Epidemiology, 51*(11), 1189–1202.
15. Bullinger, M., Alonso, J., Apolone, G., Leplege, A., Sullivan, M., Wood-Dauphinee, S., et al. (1998). Translating health status

questionnaires and evaluating their quality: The IQOLA project approach. International quality of life assessment. *Journal of Clinical Epidemiology, 51*(11), 913–923.
16. Keller, S. D., Ware, J. E., Jr, Gandek, B., Aaronson, N. K., Alonso, J., Apolone, G., et al. (1998). Testing the equivalence of translations of widely used response choice labels: Results from the IQOLA project. International quality of life assessment. *Journal of Clinical Epidemiology, 51*(11), 933–944.
17. Ravens-Sieberer, U., Auquier, P., Erhart, M., Gosch, A., Rajmil, L., Bruil, J., et al. (2007). The KIDSCREEN-27 quality of life measure for children and adolescents: Psychometric results from a cross-cultural survey in 13 European countries. *Quality of Life Research, 16*(8), 1347–1356.
18. Robitail, S., Ravens-Sieberer, U., Simeoni, M. C., Rajmil, L., Bruil, J., Power, M., et al. (2007). Testing the structural and cross-cultural validity of the KIDSCREEN-27 quality of life questionnaire. *Quality of Life Research, 16*(8), 1335–1345.
19. Scott, N. W., Fayers, P. M., Bottomley, A., Aaronson, N. K., de Graeff, A., Groenvold, M., et al. (2006). Comparing translations of the EORTC QLQ-C30 using differential item functioning analyses. *Quality of Life Research, 15*(6), 1103–1115.
20. Scott, N. W., Fayers, P. M., Aaronson, N. K., Bottomley, A., de Graeff, A., Groenvold, M., et al. (2007). The use of differential item functioning analyses to identify cultural differences in responses to the EORTC QLQ-C30. *Quality of Life Research, 16*(1), 115–129.
21. Skevington, S. M. (2002). Advancing cross-cultural research on quality of life: Observations drawn from the WHOQOL development. World health organisation quality of life assessment. *Quality of Life Research, 11*(2), 135–144.
22. Ware, J. E., Jr., Kosinski, M., Gandek, B., Aaronson, N. K., Apolone, G., Bech, P., et al. (1998). The factor structure of the SF-36 Health Survey in 10 countries: Results from the IQOLA project. International quality of life assessment. *Journal of Clinical Epidemiology, 51*(11), 1159–1165.
23. Wild, D., Eremenco, S., Mear, I., Martin, M., Houchin, C., Gawlicki, M., et al. (2009). Multinational trials-recommendations on the translations required, approaches to using the same language in different countries, and the approaches to support pooling the data: The ISPOR patient-reported outcomes translation and linguistic validation good research practices task force report. *Value Health, 12*(4), 430–440.
24. Mullen, M. R. (1995). Diagnosing measurement equivalence in cross-national research. *Journal of International Business Studies, 26*(3), 573–596.
25. Singh, J. (1995). Measurement issues in cross-national research. *Journal of International Business Studies, 26*(3), 597–619.
26. Vandenberg, R. J., & Lance, C. E. (2000). A review and synthesis of the measurement invariance literature: Suggestions, practices, and recommendations for organizational research. *Organizational Research Methods, 3*(1), 4–70.
27. Meredith, W., & Teresi, J. A. (2006). An essay on measurement and factorial invariance. *Medical Care, 44*(11 Suppl 3), S69–S77.
28. Dupuy, H. J. (1984). The psychological general well-being (PGWB) index. *Assessment of quality of life in clinical trials of cardiovascular therapies*, pp. 170–183.
29. Regnault, A. (2007). *Méthodes quantitatives pour l'évaluation de la validité interculturelle des instruments de mesure subjective évaluée par les patients*. Université Claude Bernard Lyon 1.
30. Spencer-Rodgers, J., Peng, K., Wang, L., & Hou, Y. (2004). Dialectical self-esteem and East-West differences in psychological well-being. *Personality and Social Psychological Bulletin, 30*(11), 1416–1432.
31. Holland, P. W., & Wainer, H. (1993). *Differential item functioning*. Newbury Park (NJ): Lawrence Erlbaum Associates.
32. Cohen, A. S., Kim, S. H., & Wollack, J. A. (1996). An investigation of the likelihood ratio test for detection of differential item functioning. *Applied Psychological Measurement, 20*(1), 15–26.
33. Lord, F. M. (1980). *Application of item response theory to practical testing problems*. Hillsbaum, NJ: Lawrence Erlbaum.
34. Raju, N. S. (1988). The area between two item characteristic curves. *Psychometrika, 53*(4), 495–502.
35. Raju, N. S. (1990). Determining the significance of estimated signed and unsigned areas between two item response functions. *Applied Psychological Measurement, 14*(2), 197–207.
36. Teresi, J. A., Kleinman, M., & Ocepek-Welikson, K. (2000). Modern psychometric methods for detection of differential item functioning: Application to cognitive assessment measures. *Statistics in Medicine, 19*(11–12), 1651–1683.
37. Teresi, J. A. (2006). Different approaches to differential item functioning in health applications. Advantages, disadvantages and some neglected topics. *Med Care, 44*(11 Suppl 3), S152–S170.
38. Teresi, J. A., & Fleishman, J. A. (2007). Differential item functioning and health assessment. *Quality of Life Research, 16*(Suppl 1), 33–42.
39. Petersen, M. A., Groenvold, M., Bjorner, J. B., Aaronson, N., Conroy, T., Cull, A., et al. (2003). Use of differential item functioning analysis to assess the equivalence of translations of a questionnaire. *Quality of Life Research, 12*(4), 373–385.
40. van der Flier, H., Mellenbergh, G. J., Ader, H. J., & Wijn, M. (1984). An iterative item bias detection method. *Journal of Cross-Cultural Psychology, 13*, 267–298.
41. Hidalgo, M. D., & Lopez-Pina, J. A. (2004). Differential item functioning detection and effect size: A comparison between logistic regression and Mantel-Haenszel procedures. Educational and Psychological Measurement, 64(6), 903.
42. Regnault, A., Marfatia, S., Louie, M., Mear, I., Meunier, J., & Viala-Danten, M. (2009). Satisfactory cross-cultural validity of the ACTG symptom distress module in HIV-1-infected antiretroviral-naive patients. *Clin Trials, 6*(6), 574–584.
43. Zumbo, B. D. (1999). *A handbook on the theory and methods of differential item functioning (DIF): Logistic regression modeling as a unitary framework for binary and likert-type (ordinal) item scores*. Directorate of Human Ressources Research and Evaluation, Department of National Defense.
44. Scott, N. W., Fayers, P. M., Aaronson, N. K., Bottomley, A., de Graeff, A., Groenvold, M., et al. (2010). Differential item functioning (DIF) analyses of health-related quality of life instruments using logistic regression. *Health Qual Life Outcomes, 8*, 81.
45. Crane, P. K., Gibbons, L. E., Jolley, L., & van Belle, G. (2006). Differential item functioning analysis with ordinal logistic regression techniques. DIFdetect and difwithpar. *Medical Care, 44*(11 Suppl 3), S115–S123.
46. Crane, P. K., Gibbons, L. E., Narasimhalu, K., Lai, J. S., & Cella, D. (2007). Rapid detection of differential item functioning in assessments of health-related quality of life: The functional assessment of cancer therapy. *Quality of Life Research, 16*(1), 101–114.
47. Clauser, B. E., & Mazor, K. M. (1998). Using statistical procedures to identify differentially functioning test items. *Educational Measurement: Issues and Practice, 17*(1), 31–44.
48. Hambleton, R. K., & Rogers, H. J. (1989). Detecting potentially biased test items: Comparison of IRT area and Mantel-Haenszel methods. *Applied Measurement in Education, 2*(4), 313–334.
49. Scott, N. W., Fayers, P. M., Aaronson, N. K., Bottomley, A., de Graeff, A., Groenvold, M., et al. (2010). Interpretation of differential item functioning analyses using external review. *Expert Review of Pharmacoeconomics & Outcomes Research, 10*(3), 253–258.

50. Scott, N. W., Fayers, P. M., Aaronson, N. K., Bottomley, A., de Graeff, A., Groenvold, M., et al. (2009). The practical impact of differential item functioning analyses in a health-related quality of life instrument. *Quality of Life Research, 18*(8), 1125–1130.
51. Johnson, T., Kulesa, P., Cho, Y. I., & Shavitt, S. (2005). The relation between culture and response styles evidence from 19 countries. *Journal of Cross-Cultural Psychology, 36*(2), 264–277.
52. Lee, J. W., Jones, P. S., Mineyama, Y., & Zhang, X. E. (2002). Cultural differences in responses to a Likert scale. *Research in Nursing & Health, 25*(4), 295–306.
53. van Herk, H., Poortinga, Y. H., & Verhallen, T. M. M. (2004). Response styles in rating scales evidence of method bias in data from six EU countries. *Journal of Cross-Cultural Psychology, 35*(3), 346–360.
54. Bowden, A., & Fox-Rushby, J. A. (2003). A systematic and critical review of the process of translation and adaptation of generic health-related quality of life measures in Africa, Asia, Eastern Europe, the Middle East, South America. *Social Science & Medicine, 57*(7), 1289–1306.
55. Sarro, S., Duenas, R. M., Ramirez, N., Arranz, B., Martinez, R., Sanchez, J. M., et al. (2004). Cross-cultural adaptation and validation of the Spanish version of the Calgary Depression Scale for Schizophrenia. *Schizophrenia Research, 68*(2–3), 349–356.
56. Tauler, E., Vilagut, G., Grau, G., Gonzalez, A., Sanchez, E., Figueras, G., et al. (2001). The spanish version of the paediatric asthma quality of life questionnaire (PAQLQ): Metric characteristics and equivalence with the original version. *Quality of Life Research, 10*(1), 81–91.
57. Feldt, L. S., & Kim, S. (2006). Testing the difference between two alpha coefficients with small samples of subjects and raters. *Educational and Psychological Measurement, 66*(4), 589–600.
58. Fledt, L. S. (1969). A test of the hypothesis that Cronbach's alpha or Kruder-Richardson coefficient twenty is the same for two tests. *Psychometrika, 34*, 357–370.
59. Vangeneugden, T., Laenen, A., Geys, H., Renard, D., & Molenberghs, G. (2005). Applying concepts of generalizability theory on clinical trial data to investigate sources of variation and their impact on reliability. *Biometrics, 61*(1), 295–304.
60. Pae, T. I., & Park, G. P. (2006). Examining the relationship between differential item functioning and differential test functioning. *Language Testing, 23*(4), 475–496.
61. Roznowski, M., & Reith, J. (1999). Examining the measurement quality of tests containing differentially functioning items: Do biased items result in poor measurement? *Educational and Psychological Measurement, 59*(2), 248–269.