Volume 12 • Number 4 • 2009
VALUE IN HEALTH

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

Diane Wild, MSc,[1] Sonya Eremenco, MA,[2] Isabelle Mear, MA,[3] Mona Martin, RN, MPA,[4] Caroline Houchin, MA,[1] Mary Gawlicki, MBA,[5] Asha Hareendran, PhD,[6] Ingela Wiklund, PhD,[7] Lee Yee Chong, PhD,[8] Robyn von Maltzahn, MSc,[1] Lawrence Cohen, PharmD, BCPP, FASHP, FCCP,[9] Elizabeth Molsen, RN[10]

[1]Oxford Outcomes Ltd, Oxford, UK; [2]United BioSource Corporation, Bethesda, MD, USA; [3]Mapi Research Institute, Lyon, France; [4]Health Research Associates, Inc., Mountlake Terrace, WA, USA; [5]Corporate Translations, Inc., Harford, CT, USA; [6]Global Outcomes Research, Pfizer, Ltd., Sandwich, UK; [7]Global Health Outcomes, GlaxoSmithKline, Middlesex, UK; [8]Formerly at AHP Research Ltd, Middlesex, UK; [9]Washington State University, Spokane, WA, USA; [10]International Society for Pharmacoeconomics & Outcomes Research (ISPOR), Lawrenceville, NJ, USA

[Correction added after online publication 13-November-2008: Information on FDA requirements in Section I has been updated]

ABSTRACT

Objectives: With the internationalization of clinical trial programs, there is an increased need to translate and culturally adapt patient-reported outcome (PRO) measures. Although guidelines for good practices in translation and linguistic validation are available, the ISPOR Patient-Reported Outcomes Translation and Linguistic Validation Task Force identified a number of areas where they felt that further discussion around methods and best practices would be beneficial.The areas identified by the team were as follows: 1) the selection of the languages required for multinational trials; 2) the approaches suggested when the same language is required across two or more countries; and 3) the assessment of measurement equivalence to support the aggregation of data from different countries.

Methods: The task force addressed these three areas, reviewed the available literature, and had multiple discussions to develop this report.

Results: Decision aid tools have also been developed and presented for the selection of languages and the approaches suggested for the use of the same language in different countries.

Conclusion: It is hoped that this report and the decision tools proposed will assist those involved with multinational trials to 1) decide on the translations required for each country; 2) choose the approach to use when the same language is spoken in more than one country; and 3) choose methods to gather evidence to support the pooling of data collected using different language versions of the same tool.

Key words: adaptation, linguistic validation, multinational, pooling, translation.

Introduction

The ISPOR Health Science Policy Council approved and recommended the Translation and Linguistic Validation Task Force in January 2007. It was approved by the ISPOR Board in March 2007.

Task force members are experienced and knowledgeable in translation, linguistic validation, and international measurement equivalence fields working in academia, industry, Contract Research Organizations (CROs), and as advisors to governments. They represent several countries in Europe as well as the United States.

As part of a Patient-Reported Outcome (PRO) Forum at the 2007 ISPOR 12th Annual International Meeting, an overview of initial recommendations and future direction of the task force was presented. Feedback from the forum was received.

The task force met once a month to discuss the most important issues that arise from translating and linguistically validating a PRO document. The task force divided into three subgroups that developed outlines and draft reports on their respective topic. All work received full task force review.

Once a draft version of the three-section final report was completed, it was distributed to the Patient-Reported Outcomes Special Interest Group (PRO SIG) reviewer group for a three-week review period. Substantive and constructive feedback was received and, when appropriate, incorporated into the report. In addition, the report's contents were presented at the 2008 ISPOR 13th Annual International Meeting PRO Forum. Again, comments were received and incorporated into the final report. Moreover, task force members reviewed numerous versions of the report over its 18-month development period. Once consensus was reached, the manuscript was submitted to *Value in Health*.

As a result of the increasing internationalization of clinical trials, the need to translate and adapt PRO instruments for use in countries other than that of the source language has grown rapidly and continues to develop with the increasing involvement of new countries such as India and China in clinical trials. Most instruments are developed in English-speaking countries, and therefore, need to be translated and adapted for use in other countries.

Address correspondence to: Diane Wild, MSc Oxford Outcomes Ltd, Seacourt Tower, West Way, Oxford, OX2 0JJ, UK. E-mail: diane.wild@oxfordoutcomes.com

10.1111/j.1524-4733.2008.00471.x

This task force was initiated in March 2007 with the goal of expanding on the 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 [1]. Task force members represent academia, linguistic validation services, and industry.

The task force identified three important issues not previously addressed in the earlier report or in the literature: 1) selection of languages required for translation; 2) translation methods for same language versions used in multiple countries; and 3) issues and concerns around the pooling of PRO data across countries.

SECTION 1—MULTINATIONAL TRIALS: DECISION AID TO SELECT LANGUAGES FOR TRANSLATION

Introduction

As clinical trials extend into an increasing number of countries, it becomes essential to select the language versions required for PROs effectively. A compromise should be found between the need to adequately cover the languages spoken by the target population and the need to be resource efficient.

It is important not to underestimate the time and resources needed, and complexity of implementation of PRO end points across a range of culturally different countries. Industry is wary of estimating the significance of these issues adequately (country selection, center selection, number of patients per site, need and cost of training, etc.). The more patients that can be recruited through one site, the lower the risk of noise in the data because of differences in study implementation across sites. This is especially true for developing countries where there may not be standardized health care or standardized procedures. Having fewer clinical sites with many patients will ensure greater consistency in study procedures.

In addition, there may be good reason to consider recruiting relatively large numbers of patients within a given country (e.g., 100 patients). First, this will make psychometric evaluation of the instrument more feasible. Second, there may be country-specific regulatory rules in which a certain number of patients within a country need to participate in clinical trials in order for the product to be approved for use within that country. Regulatory issues should also be taken into consideration when determining how many patients to recruit within each country.

The objective for this section of the report was to identify a process for selecting the number of languages required to adequately cover global trial populations.

Methods

The Section 1 Subgroup focused its efforts on developing a tool that would assist in determining which languages were required for countries selected to participate in a global clinical trial. The objective was to gather essential information and suggest a decision-making process that would guide users toward optimal versus definitive solutions.

The literature search revealed very little on the subject. Once it was completed, a grid was developed as a first draft tool, which required the following information:

1. Population analysis—to determine the nature of the population and its potential impact on the languages spoken in that country.
2. Disease prevalence in the country—whether it is a factor impacting language choice.
3. Analysis of language inclusion necessity—to provide both study information that may influence the selection of languages and language information that should be taken into account when making the inclusion decision.

In order to assess the ease of use, the consistency in terms of results and the clarity of the instructions, three linguistic validation companies used the grid to select languages for the following countries: Argentina, Chile, South Africa, the United States, India, and Singapore. The results were consistent among the three companies.

A post-use grid debriefing indicated that the instructions were clear, but certain changes were necessary to strengthen and clarify the requested information. In addition, a decision tree (Fig. 1) was developed to demonstrate the way in which the selection process should be conducted, once the grid was completed. See final grid (Table 1).

The grid should be completed by someone with data regarding the languages spoken in the target country, and for part 3, section 1 (analysis of language inclusion necessity), with knowledge of the study's design. The biggest issue encountered in language selection was and is data accuracy and availability. All information should be obtained from reliable sources (if possible, from up-to-date government Web sites). Any information included in the grid should be cross-referenced and clearly mentioned at the beginning of the document. Ideally, three different sources should be used and either the most reliable figure retained or an average figure calculated, if no reliable government Web site is available.

Results

Decisions regarding language selection are driven by many factors, e.g., whether a language is official (i.e., a language which has a legal status). Typically, an official language is used in government, in courts, and for all administrative matters. In most cases, an official language is selected, unless the number of speakers is limited. However, a percentage may be relative with respect to the whole population of the country. The 3% threshold shown in Figure 1 does not represent an absolute level; instead, it provides an indication of the level below which it may not be worth including the language (according to consensus from the pre-test). This level could be lower or higher depending on the number of speakers it represents, the disease prevalence in the population, and the ease of or difficulty in finding patients reflecting the intended study population of the trial. The authors do not recommend creating translations for languages spoken as few as 3% of the population apart from in the case of rare conditions.

Language selection for country

Section 1 Conclusion

Determining which language versions are needed for each country participating in a clinical trial remains a difficult issue for which there is no magic recipe. The use of tools shared by all decision-makers can greatly facilitate the decision-making process and develop a better rationale for this purpose.

SECTION 2—SAME LANGUAGE, DIFFERENT COUNTRY

Introduction

After establishing which languages will be needed for each country participating in a clinical trial, the question arises as to which approach is most suitable for producing the required

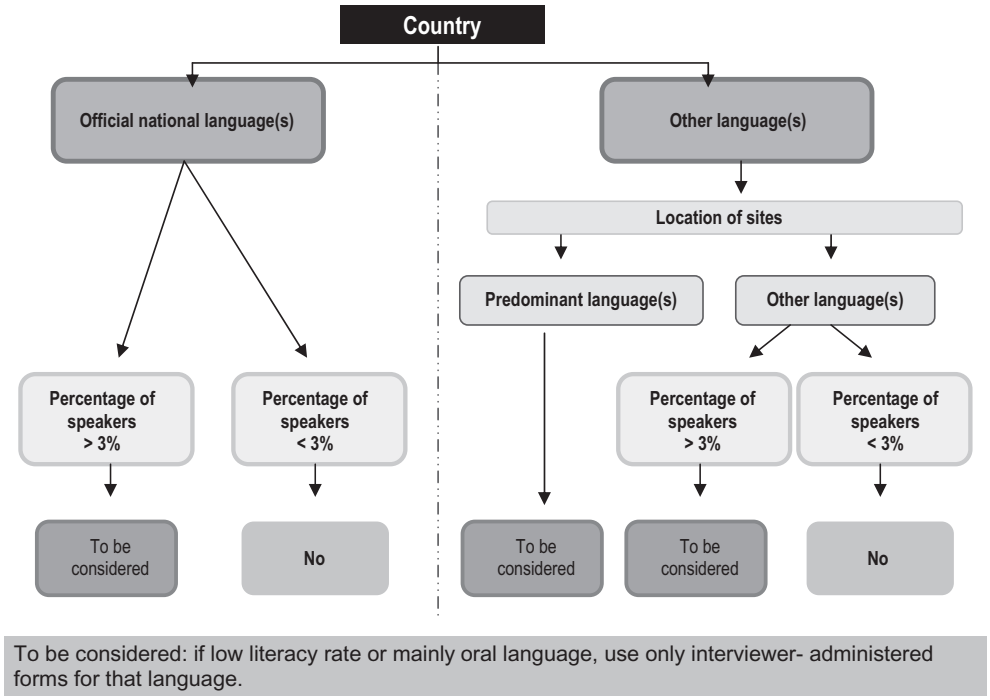

Figure 1.

translations. Consensus guidelines for the translation of PRO instruments in general have been proposed [1,2]. However, these guidelines focus on the overall process of developing a linguistically validated translation regardless of the language in question, and do not address the issue of how to approach the problem of creating an instrument in the same language for use in different target countries. The objective of this section is to present possible approaches for handling this situation and to identify the corresponding advantages and disadvantages of each approach so that a more informed decision can be made. It is beyond the scope of this article to address regional differences that can occur within a country. For the purposes of this article, we will use the term "country" to refer to different locations where the same language may be in use.

Table 1 Language selection grid

Sources of information:					
1. Population analysis					
Aim: to determine the nature of the population and its impact on languages spoken in the country					
% of Population born in Country	% of Immigrants not born in Country	Origin of Immigration	Literacy Rate	National Official Language(s)	Language of Governments and Media
X	X	X	X	X	X
2. Disease prevalence					
	Ethnicity	**Age range**		**Socioeconomic group**	
	X	X		X	
3. Analysis of language inclusion necessity					
Aim: to deterine the likelihood of the language's need for inclusion according to the criteria below					
Country XXXX					
Location of trial sites					
To be completed by study personnel		City or Region		City or Region	City or Region
Targeted number of patients per site					
To be completed by study personnel		Language A speakers		Language B speakers	Language C speakers
		X		X	X
Associated predominant language(s)					
Main language(s) classified by # of speakers		**Language**		**Language**	**Language**
		X		X	X
Official language (Y/N)		X		X	X
Oral/Written language		X		X	X
% & number of people who speak it as a mother tongue		X		X	X
Primary location of speakers (indicate any specific location)		X		X	X
Conclusion: language needed or not needed		X		X	X

X indicates where interviewer would fill in information in the grid.

Background

Two major reasons for the same language being spoken in more than one country are colonialism and boundary changes. More recently, increased immigration has led to situations where the country of origin language is still spoken by immigrants to the extent that services for immigrant populations need to be provided in their original language, (e.g., the Turkish population in Austria). As a result of the geographical dispersion of the different language groups, spoken language has evolved to the point that pronunciation and vocabulary differ, (e.g., Portuguese in Portugal versus Brazil). Written language tends to evolve more slowly. Different linguistic populations are more likely to understand the same written language compared to the spoken language, which makes the use of questionnaires written in the same language possible across different countries.

Methods

A literature review was conducted to investigate existing guidelines and approaches used to address the issue of translations, which are required in the same language for use in different countries. Few publications address this question [2,3]. In many instances, publications provided recommendations intended for a specific instrument, such as the FACIT [4,5], EORTC [6], EQ-5D (EuroQOL guidelines), and the Nottingham Health Profile [7], but were not recommendations for PRO instruments in general. The results of this literature review confirmed the need for more information in this area.

Possible Approaches

Based on the literature and our practical experience, we have identified three primary approaches to address the issue of languages spoken in more than one country. A fourth approach, developing a translation for use in one country and assuming that it will be acceptable for use in all other countries sharing the same language with no further work, will not be discussed in this article. We consider this approach to be inadvisable with the exception of rare cases where the sample size from a country is so small that the additional work is not economically feasible. For the purposes of this article, we shall assume that English is the source language of the PRO instruments to be translated.

The approaches are as follows:

1. *Country-specific approach*: different versions of a translation are developed for each major country or subpopulation within a country.
2. *Same language adaptation approach*: a language version of an instrument exists for one country; it is adapted for use in new countries or populations.
3. *Universal approach*: the translation is intended for multiple locales from the outset. Translators from different countries of origin reach a compromise in order to achieve a translation understood by all.

In some cases, a combined approach can be taken in which the translation is initiated using the universal approach, and if resolution of differences is impossible, small variations for each country are implemented. It is also possible to reverse the process from country-specific to universal. However, this can be more difficult because the translation is focused only on one group from the outset.

We have found that, in practice, many variations on each of these approaches can be taken depending on the instrument in question and its translation history. We have developed a set of scenarios to better illustrate how these approaches can be applied in real-world situations. These scenarios present the approaches as a continuum from completely country-specific on one end to completely universal on the other end, with multiple variations of the approaches, including same language adaptation, in between the two extremes. These scenarios also include ranges of the numbers of translators involved and a relative ranking of the time and cost involved (based primarily on the number of translators and the type and number of steps involved) as other ways to compare the approaches.

Table 2 presents scenario 1, where there are multiple Spanish-speaking target countries and no existing translations available. Table 3 presents scenario 2, where a translation for France exists, and multiple translations for French-speaking target countries are required.

Additional scenarios exist beyond the scope of this article. They are the following:

1. One year after the validations described in Scenario 1 have been completed, a Peruvian Spanish version is required.
2. Spanish for Mexico, United States, Guatemala, Colombia, Venezuela, and Argentina are required. No current Spanish translations exist.
3. English for Canada, UK, Australia, India, and South Africa are required. The original version of the instrument is US English.
4. Chinese for Taiwan and Hong Kong are required. No current translations exist.

Note: Space constraints do not allow for the reproduction of all scenarios within this article. If you would like to receive a full version of the scenario tables, please contact the authors of this section.

A summary of advantages and disadvantages of the country-specific approach follows:

Advantages

- Allows for more colloquial and idiomatic usage. There will be much less risk that patients will misunderstand wording used, including terms used for cultural references, such as education. This might be especially relevant for older patients who are accustomed to more traditional language with fewer loan words from other countries.
- Timelines may be shorter than with the universal approach because time is not spent waiting for consensus from all country representatives.

Disadvantages

- Major variations in translations of the same language may introduce bias and reflect stylistic differences of translators rather than true differences between the languages as spoken in those countries. It is difficult to provide evidence for what is an essential change based on true linguistic/cultural issues and what is the personal opinion of the translators involved.
- Sponsor costs for Case Report Forms (CRF) printing may be higher.

Advantages and disadvantages of the same language adaptation approach follow:

Table 2 Scenario 1

Scenario 1
Spanish for multiple target countries: Spain, US, Mexico, and Argentina
No current Spanish translations exist.

Key: 1 = Least; 5 = Most

Option No.	Option Description	PROS*	CONS	Relative Cost	Relative Time Required
1 **Country-specific**	Create four separate translations independently of each other. Debrief in each country. **Resources Required** 8–16 translators/2–4 back translators **Result** 4 country-specific translations	CSP1. All translations tailored specifically for each country. CSP2. Technological terms, educational/health-care systems, demographic items, product names etc. would not require further adaptation. CSP3. Translators from different countries not required to compromise in order to reach agreement. CSP4. Translators from different countries not biased by the vocabulary and sentence structure of a translation created for another country.	CSC1. Maximum potential for stylistic or otherwise unnecessary differences in final translations. CSC2. Users required to maintain maximum number of versions. CSC3. Potential for bias because of differently worded questions.	5	1
2 **Combination of country-specific and same language adaptation**	Create a translation for Spain and debrief in Spain. Independently create a second translation for any of the other three countries. Then ask translators from the remaining two countries to adapt the initial translation for their target country. Debrief separately in each of the three countries. **Resources Required** 6–8 translators/4 back translators **Result** 2 country-specific translations/2 same language adaptations	CSLAP1. Translation for Spain and second translation tailored specifically for those countries. CSLAP2. Adaptations linguistically and culturally suitable for each remaining country. CSLAP3. Opportunity to adapt technological terms, educational /health-care systems, demographic items, product names etc. CSLAP4. Fewer stylistic or otherwise unnecessary differences between various versions because they were adapted rather than created independently. CSLAP5. Eliminates problems associated with adapting all other versions from a version with known dissimilarities in sentence structure and usage.	Same as CSC2 and CSC3 plus: CSLAC1. Translators performing adaptations may be biased by original translation and/or fail to make necessary cultural changes. Should this happen, some wording may not be completely linguistically suitable in the translation used.	4	2
3 **Combination of country-specific and same language adaptation**	Create a translation for Spain. Provide the translators for the other three target countries with the Spanish-Spain translation and ask translators to adapt it for each of their target countries. Debrief separately in each of the four countries. **Resources Required** 5–10 translators/2–4 back translators **Result** 1 country-specific translation/ 3 same language adaptations	Same as CSLAP2, CSLAP3, CSLAP4 plus: CSLA2P1. Translation for Spain tailored specifically for that country.	Same as CSC2, CSC3, and CSLAC1, plus: CSLA2C1. More difficult to create adaptations from version for Spain because of dissimilarities in sentence structure and usage.	3	2
4 **Combination of country-specific and universal**	Create a translation for Spain and debrief in Spain. Translators from Mexico, the US, and Argentina work together to create a single translation that they agree would be acceptable in all three target countries. Debrief separately in each of the three countries. **Resources Required** 5–8 translators/2–4 back translators **Result** 1 country-specific translation/1 universal translation	Same as CSLA2P1 plus CSUP1. Users required to maintain fewer versions. CSUP2. Even fewer stylistic or otherwise unnecessary differences because second version developed universally.	Same as CSC3 and CSLAC1 plus CSUC1. More likely that some wording will not be completely suitable, linguistically and culturally, for all countries. CSUC2. Less opportunity to adapt technological terms, educational/health-care systems, demographic items, product names etc. CSUC3. Translators may compromise on wording in order to reach agreement on final translation. CSUC4. Four unique versions may still be necessary depending on debriefing results.	2	1
5 **Universal**	Translators from Spain, Mexico, the US and Argentina work together to create a single translation that they agree would be acceptable in all four target countries. Debrief separately in each of the four countries. **Resources Required** 4 translators/1–2 back translators **Result** 1 universal translation	UP1. Users required to maintain a single version. UP2. Single version eliminates the problem of stylistic or otherwise unnecessary differences between versions. UP3. Less potential for bias because same question is asked to all.	Same as CSUC3 and CSUC4 plus: UC1. Most likely that some wording will not be completely suitable, linguistically and culturally, for all countries. UC2. No opportunity to adapt technological terms, educational /health-care systems, demographic items, product names etc. UC3. Most difficult to reach consensus on wording. UC4. Suitability of wording based on viewpoint of a single translator in each country.	1	2

*In order to reduce repetition of text in the scenarios, each Pro or Con is labelled with the initials of the Option, the letter P or C for pro or con, and the number listed. For example, the first Pro under country-specific is labeled CSP1.

Table 3 Scenario 2

Scenario 2
Multiple target countries: France, Belgium, Canada, and Switzerland
A fully harmonized and debriefed translation of Instrument A exists for France.

Key: 1 = Least; 5 = Most

Option No.	Option description	PROS	CONS	Relative cost	Relative time required
1 **Country-specific**	Create translations for Belgium, Canada, and Switzerland independently of the existing translation for France. Debrief in each of the three countries. **Resources Required** 6–12 translators/3–6 back translators **Result** 4 translations	Same as Scenario 1, Option 1.	Same as Scenario 1, Option 1.	4	1
2 **Same language adaptation**	Provide the translators for the other three target countries with the French-France translation and ask the translators from each of the remaining countries to adapt the French translation for their target country independently. Debrief separately in each of the three countries. **Resources Required** 3–6 translators/0–3 back translators **Result** 1 translation/3 adaptations	Same as CSLAP2, CSLAP3, CLSAP4.	Same as CSC2, CSC3, and CSLAC1 plus: SLAC1. More difficult to create Canadian adaptation because of dissimilarities in word usage.	3	4
3 **Same language adaptation by region**	Group countries according to "linguistic-similarity" creating regional translations—one for Europe and one for Canada. Ask translators from Belgium and Switzerland to adapt the French-France translation for use in either country. Translators from Canada would be asked to adapt the French-France version for use in Canada only. Debrief separately in Europe using a larger sample size so that both target countries are covered. Debrief as usual in Canada. **Resources Required** 3–6 translators/0–2 back translators **Result** 1 translation/2 adaptations (moving toward universal)	Same as CSLAP4 and CSU1.	Same as CSC3, CSU1, CSU2, CSU3, CSU4.	2	2
4 **Universal**	Translators from France, Belgium, Canada, and Switzerland work together to create a single translation that they agree would be acceptable in each of the four target countries. Debrief separately in each of the countries (including France again, if resulting translation is different from original French translation). **Resources Required** 4–8 translators/1–2 back translators **Result** 1 translation	Same as Scenario 1, Option 5.	Same as Scenario 1, Option 5.	1	3

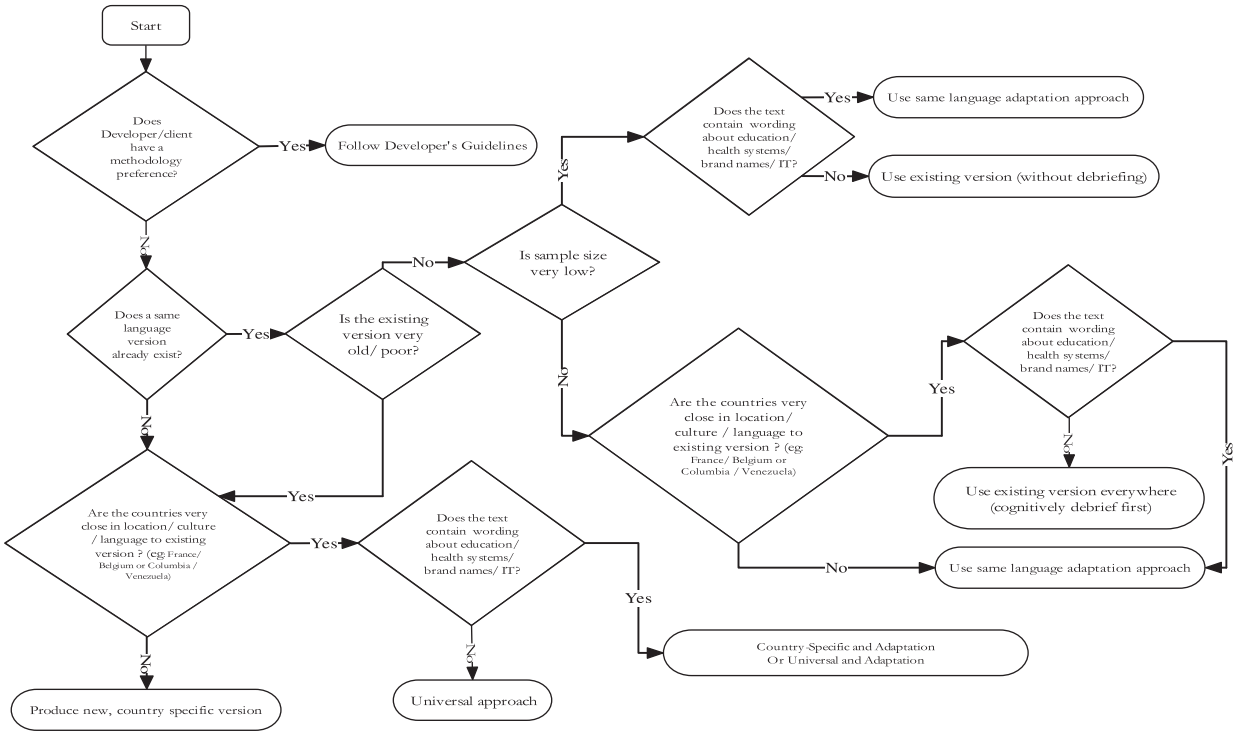

Figure 2 Decision tree.

Advantages

- When used for a specific country, advantages are similar to the country-specific advantages above. It also allows for terminology specific to that country while retaining parts of the previous translation.

Disadvantages

- In cases where fewer translators are involved, there may be an increased influence of personal opinion over linguistic necessity.

Advantages and disadvantages of the universal approach follow:

Advantages

- The same question is asked to all groups; therefore, there is less likelihood of bias from asking differently worded questions.
- There may be practical and logistical benefits in only having one language version available.

Disadvantages

- Wording may not sound as natural to patients, and the language may be culturally and linguistically bland (however, in most cases, patients are able to understand by using their passive vocabulary, which refers to understanding terms even if not part of the patient's everyday usage).
- If a universal language version has been created, the translation becomes less straightforward than any individual approach, especially if at a later date a translation is needed for another country speaking the same language. See scenario 2 for options for how to handle this situation.

Factors to Consider when Deciding on an Approach

There is consensus that no single approach is superior to another, with each having its own advantages and disadvantages. Our recommendation is that each situation and study be evaluated on a case-by-case basis. We have developed a decision tree (Fig. 2) to assist with this process and provide guidance regarding instances in which a clear preference can be made for one approach over the other. The factors to take into account are the languages and countries involved as well as the content of the measure and developer guidelines.

General Guidelines

1. The decision as to the most appropriate approach to use relates partly to the cultural similarities and differences between the countries in which the target languages are spoken. For example, the similarities are often greater between two Spanish-speaking South American countries than those between Spain and any of the South American countries. The linguistic similarity between French for France and French for Belgium is greater than that between French for France and French for Canada.
2. Some subject matter cannot be translated universally:
   - educational systems;
   - some demographics (such as income level and ethnicity);
   - some sensitive or dietary content may be highly culturally linked (e.g., sexual performance or items referring to alcohol);

- health resource utilization referring to specific institutions (e.g., Meals on Wheels);
- government institutions and programs like health insurance, disability; and
- computer terminology, now required for some ePRO translations.

The appropriateness of the measure to the method is something that must always be considered and never assumed.

Methods for Carrying Out These Translation Approaches

After deciding upon the approach or combination of approaches, the following general processes are recommended:

Same language adaptation method

- Start from an existing language version.
- Two native speakers (professional translators or key in-country investigators) from the target population review the existing translation and identify terms or usage that are not acceptable in the target country, or that would be misunderstood there.
- After comparing the reviewers' comments, a native-speaking reviewer (new person or one of the above) confirms the acceptability of the revised instrument in the target country or identifies terms that would be misunderstood there.
- Perform back-translation of revised items from target language to English, to ensure accuracy.
- Implement changes.
- Proofread revised instrument.
- Cognitive debriefing with patients in each target country to confirm that the new version is well understood.

Universal translation recommendations applied to standard translation methodology

- Forward translators should be from multiple countries of origin, especially the target countries, for the study. For example, for German, one forward translator could be from Germany and the other forward translator from Austria. If many other countries need to be included, more than two forward translations are suggested.
- Reconciler should be familiar with usage in different countries to be able to reconcile the differences in a way that is not slanted toward one country or another.
- After back-translation, multiple reviewers or clinicians provide input on the reconciled translation and identify problems for their country or region.
- Translators work together to find solutions acceptable to all the regions. In cases where such resolution is not possible, different final translations can be produced which maintain most of the same wording but include the country-specific variations required.
- Proofreading (by representatives of the different countries if universal version is maintained).
- Cognitive debriefing with patients in the different countries is recommended to confirm that the universal translation is understood and acceptable. The subject pool needs enough people from each target country to capture feedback, in most cases a minimum of five per country or region. If the subjects find problems with the universal translation and make suggestions for changes, a same language adaptation of the universal version could then be created based on this feedback with only the problematic terms or items altered.

Country-specific translation method. The country-specific translation of a new measure would follow the ISPOR good practice guidelines [1], with translators, reconcilers, and in-country representatives from a single country being involved in the process.

Section 2 Conclusion

When addressing the issue of translations for the same language, different country, the situation is complex. With a lack of literature on this issue, we relied on practical experience. We explored several well-known approaches: country-specific, same language adaptation, and universal. Developing the scenarios and decision tree confirmed that no single approach is better or more appropriate in every situation. Our recommendation is that each study be assessed on a case-by-case basis. Furthermore, the language, countries involved, and instrument content need to be evaluated to determine the most suitable approach.

More research to compare the results of these approaches is necessary. For example, one possible empirical study could compare the data collected with a universal translation from two or more countries using differential item functioning to assess if the same translation performed differently in the countries being compared. Another possible study would compare data collected from different countries using country-specific versions or same language adaptations to assess whether the different translations show bias. Studies such as these would contribute empirical data on different translation approaches and provide insight into methods to assess pooling of data from different translations.

SECTION 3—ISSUES AROUND THE AGGREGATION OF DATA FROM MULTIPLE LANGUAGES AND CULTURES

Introduction

The increasing inclusion of patient-reported outcome (PRO) measures in large multicountry trials has introduced many new methodological challenges to the analysis and interpretation of data from the trials. PROs are often developed in English and translated into the various languages needed to support these global trials.

Current literature describes standard linguistic validation methods for developing high quality new language versions of PROs [1,2]. Most of these methods, however, focus on the quality of the individual language versions to ensure semantic equivalence, and do not extend to addressing the appropriateness of combining and analyzing data derived from multiple language versions.

Clinical trials are conducted in many countries to enhance the variability and number of subjects included for evaluation. Increased sample size has a positive effect on the power and representativeness of the statistical analyses [8]. However, if data is inappropriately pooled, inherent differences among the different data sources can be hidden, yielding misleading results and inferences and, ultimately, invalid findings [9].

Successful aggregation of multinational data sets is critical to achieve the benefits associated with an increased sample size. Issues around the appropriateness of data pooling are important. Differences in clinical data across countries could result from nuances of clinical practice standards of care, site personnel, procedures, and variations in case mix. Differences from cultural and linguistic preferences in the subjective expression of outcome variables present additional challenges for PROs.

There are currently no established criteria concerning study design or analytical requirements to assure the appropriateness of

the aggregation of data derived from multiple languages and cultures. The assumption that different language versions of the same instrument (if adapted using appropriate methodology) have equivalent psychometric properties, and perform similarly in the different language groups, has been used to support multinational pooling of trial data [10]. This is, however, an untested assumption. Luo et al. [11] have suggested that the data pooling step can only be carried out confidently if the different language versions are measuring the same construct with the same metric.

Many types of equivalence can be considered for pooling PRO data from multiple sources. These include basic issues of the descriptive characteristics of the population including literacy levels, cognitive equivalence of the concepts used in the data collection instruments, equivalence in the methods used to derive the multiple language versions of the measures, equivalence in the ability of the measures to demonstrate acceptable psychometric properties (psychometric equivalence), and metric (measurement) equivalence in the response scales and scoring of the collected data.

For pooling PRO data from multicountry studies, initial steps include ensuring equivalence in study design and methods. The next step is to ensure linguistic and cognitive equivalence of the PRO concepts being measured (using appropriate linguistic and cultural adaptation methods, see section 2). Methods for these have been discussed in the literature [1,2,12]. In addition, equivalence of the concepts being measured could be tested initially in focus groups or one-on-one interviews in the new countries/cultures to ensure that the concept is valid.

This section provides a discussion of methods used in order to assess measurement equivalence, which should be considered for pooling of data derived from different languages and cultures.

Measurement Equivalence

Measurement equivalence is a complex concept with many different existing definitions.

Horn and McArdle [13] and Mullen [14] view measurement equivalence as reflecting the degree to which a measurement instrument and its corresponding data collection protocol can yield reliable and valid data about some phenomena of interest across different populations. Measurement equivalence is deemed present if, under different conditions for observing study phenomena, the measurement operations yield the same attribute [9].

Luo et al. [11] define measurement equivalence as different language versions of the same instrument yielding similar scores at the item and scale levels with identical levels of Health Related Quality of Life (HRQoL) expressed by the respondents. Drasgow (1984) further recognizes measurement equivalence when the relationship between observed test scores and the attribute measured are identical across subpopulations in the data set [15]. Therefore, the primary aim of measurement equivalence is to ensure that the only reason for differences in scores between contributing populations is due to actual differences between the groups in the construct being measured, and not other issues that could affect the pooled data.

Simply, measurement equivalence is an analytic process for assessing instruments' resulting measurement properties when used in a trial across the contributing populations.

Methods for Demonstrating Measurement Equivalence

A diverse range of methods are employed in order to assess measurement equivalence, including classical test theory, factor analysis, structural equation modeling (SEM) and differential item functioning (DIF).

Classical Test Theory (CTT). Classical test theory is based on the idea that a test consists of a series of items, each of which is one attempt to measure the psychological characteristic being assessed. Each attempt at measuring the trait (item) produces a slightly different result because each test item is not perfectly reliable. Classical test theory, when used with measurement equivalence, ensures that each language version has similar theoretical values (such as mean, variance, etc.) within a similar population.

Scott-Lennox et al. [16] used the classical test theory elements of reliability and validity to test the equivalence of different language versions of the MOS in an HIV population. Data relating to the MOS and the SCL-57 were collected using five different language versions, in addition to the self-reported data collected from 363 HIV positive outpatients. The intracultural and transcultural psychometric adequacy of the translations was assessed to determine the appropriateness of the Medical Outcomes Study (MOS) HIV translations. The internal consistency, item discrimination, and item convergence, as well as floor and ceiling effects of the different language versions, were assessed. The authors found that, in general, the five translations had similar psychometric properties to the US English version.

Stahl et al. [17] examined the cross-sectional and longitudinal correlations between translations of the Asthma Quality of Life Questionnaire (AQLQ) and clinical assessments. They concluded that the consistency of the cross-sectional correlations between the AQLQ and the clinical scales across countries support the validity of the translations.

Factor analysis. Factor analysis relies on the notion that equivalence is achieved when the relationships between observed scores and latent variables are equal across comparison groups. The lack of equivalence can be due to two reasons: 1) mean group differences for the latent variable, or 2) a lack of equivalent measurement across groups. The first reason is easy to test; the second is more serious as it means that the measure functions differently across groups [16,18–20]. Strauss and Carpenter [21] used factor analysis to assess the factor structure of the French translation of the Strauss and Carpenter revised outcome criteria scale (SCOCS-R). They found that the factor structure, as well as the inter-rater reliability and convergent validity of the French translation, mirrored that of the original English version of the scale.

Structural Equation Modeling (SEM). SEM is a statistical technique for testing and estimating causal relationships using a combination of statistical data and causal assumptions. SEM can be used to assess a series of nested measurement models [22] to determine whether there is measurement equivalence between language groups.

Differential Item Functioning (DIF). DIF occurs when people from different groups (e.g., languages/cultures) with the same latent trait have a different probability of giving a certain response on a measure. It is a feature of the Item Response Theory (IRT) approach and has been utilized to assess measurement equivalence between translations. Scott et al. [23] compared 13 translations of the European Organisation for Research and Treatment of Cancer (EORTC) Quality of Life Questionnaire-C30 (QLQ-C30) across 22 countries. Most languages showed similar results to English. However, at least one instance of significant DIF was found for each translation. In another study conducted by Pagano and Gotay [24], DIF was found on several of the items of the EORTC QLQ-C30 between Caucasian, Filipino, Hawaiian, and Japanese groups although the overall QoL scores were the same across groups.

The choice of method(s) used in order to assess measurement equivalence should depend on a number of factors, but perhaps most importantly, the sample size from the countries of interest. The sample size requirements for a CTT analysis are much smaller than those required for factor analysis, SEM and DIF. When Smith et al. [18] used SEM techniques, they found that a Japanese translation of the circadian rhythm scale did not measure the same construct as the original English version of the scale. However, when the authors used a CTT approach, they found equivalence between the versions.

Challenges around interpretation. No matter which method is used to support the pooling of multicultural trial data, there is no existing gold standard to define the level of accepted similarity or variance. While measurement equivalence may be the aspired goal, practical solutions and parameters need to be developed to define the degree of measurement equivalence required to support data pooling from global trials. Alternative solutions to resolve these issues need to be explored further.

Whether it is practically attainable to establish a level of equivalence or not, some form of guidance on the suggested approach and degree of required similarity would be helpful. Ideally, this would include tests to explore whether data can be pooled, and rules for deciding when the PRO data from a country should be dropped. Smith et al. [18] and Harvey et al. [25] suggest ensuring that samples are as similar as possible at the outset of the study to reduce confounding effects.

However, it is difficult to establish a priori regarding just what mix of cross-national differences exist. In addition, there is a need to consider representativeness of the different samples within the multicultural populations (Harvey et al., [25]). As suggested earlier, the cross-cultural validity of concepts could be explored with qualitative work. In addition, this could be explored through analysis of existing trial sets, and discussion with local clinical experts.

When differences are found, it may be advisable to conduct qualitative research in the countries/cultures of interest with groups of patients and/or health professionals to determine whether there are any potential explanations for the differences. If available, other data sets using the same PROs can be explored.

Section 3 Conclusion

When assessing the acceptability of pooling data from multinational trials, evidence of equivalence can be examined at various levels including similarity of study design and the linguistic/cultural equivalence of the PRO. Examining measurement equivalence is an additional consideration. There are several analytic approaches that explore the similarity and difference of measurement equivalence across subpopulations. However, a definitive point of similarity or acceptable amount of difference that can be tolerated does not currently exist. Efforts to strengthen the quality of individual language versions as they are developed cannot provide a guarantee that the measures themselves will perform with an acceptable level of similarity across populations.

Researchers are encouraged to apply a wide variety of quantitative and qualitative techniques and to review the results of these analyses holistically to determine whether a test is adequate for a specific application [15].

Patrick et al. [26] have suggested that in order to verify basic measurement properties, such as distribution of responses, internal consistency, test–retest reliability, and validity across multiple languages, it is necessary to consider population characteristics and variability around the data, such as distribution of responses, means, standard deviations, etc., for different language groups.

Researchers are calling for greater scientific proof of the appropriateness of many accepted, but untested, methods for treating multinational trial data [27]. It is important that any recommendations derived from such proof be both appropriately rigorous to protect the quality of future trials, and sufficiently flexible to accommodate the rapidly changing nature and environment for clinical research in which PROs are increasingly used to determine the value of pharmacologic compounds.

Source of financial support: No financial support was received to fund the manuscript.

References

1 Wild D, Grove A, Martin ML, et al. ISPOR principles of good practice: the cross-cultural adaptation process for patient reported. Value Health 2005;8:94–104.
2 Acquadro C, Conway K, Hareendran A, Aaronson, N, for the ERIQA Group. Literature review of methods to translate health-related quality of life questionnaires for use in multinational clinical trials. Value Health 2008;11:509–21.
3 Eremenco S, Cella D, Arnold BJ. A comprehensive method for the translation and cross-cultural validation of health status questionnaires. Eval Health Prof 2005;28:212–32.
4 Bonomi AE, Cella DF, Hahn EA, et al. Multilingual translation of the functional assessment of cancer therapy (FACT) quality of life measurement system. Qual Life Res 1996;5:309–20.
5 Cella D, Hernandez L, Bonomi AE, et al. Spanish language translation and initial validation of the functional assessment of cancer therapy quality-of-life instrument. Med Care 1998;36:1407–18.
6 Koller M, Aaronson NK, Blazeby J, et al. Translation procedures for standardised quality of life questionnaires: The European Organisation for Research and Treatment of Cancer (EORTC) approach. Eur J Cancer 2007;43:1810–20.
7 Hunt SM, Alonso J, Bucquet D, et al. Cross-cultural adaptation of health measures. European Group for Health Management and Quality of Life Assessment. Health Policy 1991;19:33–44.
8 Thumboo J, Fong KY, Chan SP, et al. The equivalence of English and Chinese SF-36 versions in bilingual Singapore Chinese. Qual Life Res 2002;11:495–503.
9 Rungtusanatham MJ, Ng CH, Zhao X, Lee TS. Pooling data across transparently different groups of key informants: measurement equivalence and survey research. Decis Sci 2008;39:115–45.
10 Mark BA, Wan TTH. Testing measurement equivalence in a patient satisfaction instrument. West J Nurs Res 2005;27:772–87.
11 Luo N, Chew LH, Fong KY, et al. Do English and Chinese EQ-5D versions demonstrate measurement equivalence? An exploratory study. Health Qual Life Outcomes 2003;1:7.
12 Maneesriwongul W, Dixon JK. Instrument translation process: a methods review. J Adv Nurs 2004;48:175–86.
13 Horn J, McArdle J. A practical and theoretical guide to measurement invariance in aging research. Exp Aging Res 1992;1:117–44.
14 Mullen MR. Diagnosing measurement equivalence in cross-national research. J Int Bus Stud 1995;26:573–96.
15 Davies S, Little IS, Ross R. Ensuring the measurement equivalence and appropriate use of personality assessments across cultures. Paper presented at the Annual Meeting of the Society for Industrial-Organizational Psychology, May 2006, Dallas.
16 Scott Lennox JA, Wu A, Boyer G, Ware JE. Reliability and validity of French, German, Italian, Dutch, and UK English translations of the Medical Outcomes Study HIV Health Survey. Med Care 1995.
17 Stahl E, Postma DS, Juniper EF, et al. Health-related quality of life in asthma studies. Can we combine data from different countries? Pulm Pharmaco Ther 2003;16:53–9.
18 Smith CS, Tisak J, Bauman T, Green E. Psychometric equivalence of a translated circadian rhythm questionnaire: implications for

between- and within- population assessments. J Applied Psycho 1991;76:628–36.
19 Wee HL, Ravens-Sieberer U, Erhart M, Li S. Factor Structure of the Singapore English version of the KINDL Children Quality of Life Questionnaire. Health Qual Life Outcomes 2007.
20 Timmerman EM, Hoogstraten J, Nauta M, Meijer K. Structural comparison of a translated dental attitude questionnaire: a factor analytic study. Community Dent Oral Epidemiol 1996;24:236–9.
21 Poirier S, Bureau V, Lehoux C, et al. A factor analysis of the Strauss and Carpenter revised outcome criteria scale: a validation of the French translation. J Nerv Ment Dis 2004;192:864–7.
22 Vandenberg RJ, Lance CE. A review and synthesis of the measurement invariance literature: suggestions, practices, and recommendations for organizational research. Organ Res Methods 2000;4–69.3
23 Scott NW, Fayers PM, Bottomley A, et al. Comparing translations of the EORTC QLQ-C30 using differential item functioning. Qual Life Res 2006;1103–15.
24 Pagano IS, Gotay CC. Ethnic differential item functioning in the assessment of quality of life in cancer patients. Health Qual Life Outcomes 2005;3:60.
25 Harvey PD, Fortuny LA, Vester-Blackland E, De Smedt G. Cross-national cognitive assessment in schizophrenia clinical trials: a feasibility study. Schizophr Res 2002;59:243–51.
26 Patrick DL, Burke LB, Powers JH, et al. Patient-reported outcomes to support medical product labelling claims: FDA perspective. Value Health 2007;10(Suppl.):S125–37.
27 Lenderking WR. Comments on the ISPOR Task Force Report on Translation and Adaptation of Outcomes Measures: guidelines and the need for more research. Value Health 2005;8:92–3.