Guidelines for Establishing Cultural Equivalency of Instruments

Richard Ohrbach, DDS, PhD
University at Buffalo, NY, US

Jakob Bjorner, MD, PhD
Quality Metric, Inc, RI, US

MaryAnn Jezewski, RN, PhD
University at Buffalo, NY, US

Mike T John, DDS, PhD
University of Leipzig, Germany

Frank Lobbezoo, DDS, PhD
ACTA, The Netherlands

Developed for the
Committee for Translations and Protocols
International RDC/TMD Consortium Network
Updated May 11, 2013

Send any correspondence regarding this document to
Richard Ohrbach at ohrbach@buffalo.edu

Contents

Introduction

Research that crosses linguistic and cultural boundaries necessarily requires direct attention to the use of language and to cultural factors when verbal expression, verbal comprehension, or both are involved, at any level, in the systematic collection of data expected to exhibit comparable reliability and validity across the linguistic and cultural boundaries. Only if these criteria are met can the data serve as a foundation for generalization of findings. This set of Guidelines, representing a synthesis of literature and recommendations across multiple disciplines, was initially developed in response to identified research needs by the International RDC/TMD Consortium Network (Consortium), but it was intended at the outset to be broadly applicable to any instrument used across linguistic or cultural boundaries. The Research Diagnostic Criteria for Temporomandibular Disorders (RDC/TMD)[1] is a self-report and clinical examination instrument developed in English for use in the United States, and it was the core research instrument for Consortium activities when the Consortium was established; the RDC/TMD served as the source instrument for more than 20 approved translations. As a result of creating these translations, Consortium members acquired substantial experience in the methods necessary to create the best translations. That instrument served as an excellent example for the various challenges that emerge in the development of a translated instrument, as described by the guidelines in this document. The RDC/TMD has been replaced by the DC/TMD (Diagnostic Criteria for Temporomandibular Disorders), and it exhibits the same range of challenges in correctly translating an instrument for reliable and valid data collection.

The Consortium was established in 2000 with funding support from NIDCR/NIH in order to facilitate international research, and, through the efforts of Consortium members, the RDC/TMD has been translated and implemented into over 20 languages (as of November 2008) for use in other countries. These translations have proceeded according to accepted standards for valid document translation, but simultaneous developments in multi-national research have shown that semantic translation alone is not necessarily sufficient for a research instrument to produce valid data in another setting. Consequently, this set of guidelines was developed for the Committee for Translations and Protocols; this set of guidelines describes the procedures necessary for the development of a reliable and valid translated instrument required for Consortium-related pain research internationally. The backgrounds of the individuals who have contributed to this document include psychology, psychometrics, cultural anthropology, epidemiology, and clinical research; all have been involved in instrument development, instrument translation, and international clinical studies.

The intent of these guidelines is to foster valid instrument development, which will necessarily lead to better data that can be compared across multi-language and multi-cultural settings. The Committee recognizes that these standards will require substantial effort for the creation of international instruments, but, simultaneously, such efforts can be rewarded by productive collaborations among members and by publishable outcomes: the culturally equivalent instruments. Multi-national and cross-cultural instrument development itself is a demanding but highly rewarding scholarly activity. Recent efforts that nicely demonstrate the process of cultural adaptation have been performed by the Brazilian group[2], the Dutch group[3], and the German group[4], all members of the

Consortium. In addition, the SF-36, and its language adaptations, represent exceptional examples of this process, to which a special journal issue was dedicated[5].

This document is concerned with the development of instruments to be used for the collection of information from human subjects in multi-language or multi-cultural settings, based on the existence of a source instrument in another language. "Instruments", in this context, primarily refers to self-report questionnaires (which can include both checklists as well as measurement scales), but the term can also include any standardized verbal interaction between examiner and subject (patient) that has as its intent a response (verbal or otherwise) from the subject which will be systematically collected as data. An "item" is any single question or measurement procedure. The instrument to be developed via translation is termed the "target" instrument, and the instrument that it is developed from is termed the "source" instrument. The person who initiates and is responsible for the development of an instrument in another language or setting is termed "Team Leader"; the Team Leader is often also a forward translator, but the two roles are distinct in this document. The exact tasks that the Team Leader undertakes in the translation and equivalency process are Team Leader-dependent; a Team Leader could be primarily a manager, coordinating all aspects of instrument development, or the Team Leader could also be a content and methods expert, for example, and participate fully in multiple stages of instrument construction. There is no necessity or advantage for the Team Leader to be "blind" to any aspect of instrument development. However, translator, back-translator, and reviewer are three roles that are independent and must always be independent. These roles are described in greater detail subsequently.

There are three approaches to developing instruments in a second language: sequential, parallel, and simultaneous [6].

- The sequential approach starts with a completed instrument in a source language, and then translation, back-translation, and equivalency are performed on an adaptation in the target language; examples of instruments developed in this manner for second languages include the Sickness Impact Profile[7] and the SF-36[8]. This is the most commonly used approach.
- In the parallel approach (e.g., Euro-QOL[9]), culturally relevant input is present throughout the instrument development process; international discussions of each item are relied upon in order to help develop a single set of items appropriate to the measurement of the construct in each of the development settings.
- In the simultaneous approach, (e.g., see the WHOQOL[10] as a prime example), an assumption is made that both universal as well as culture specific assessments are required, and so a given version of the instrument in a particular language is comprised of both general items that exist in all language versions as well as items specific to that culture, in relation to the respective construct. This method is the least commonly used, in part due to the enormous resources and coordination required for simultaneous instrument development in more than one site at one time, utilizing both common and unique items.

Consequently, this document is primarily focused on the processes required for the sequential approach, in order to adapt an existing instrument from one language or cultural setting to another one.

The amount of attention any instrument requires, in terms of cross-cultural equivalence, will depend on the nature of the item(s) comprising an instrument or a section of an instrument; the nature of the items also includes consideration of the underlying type of validity (e.g., face, content, etc) as well as the underlying construct, latent variable, or trait. The latent variable (or latent construct) is the characteristic the instrument is intended to measure through the items. For example, the DC/TMD consists of self-report data and examination data and serves as a suitable example for the range of problems that might be encountered in translating an instrument for medical application.

The self-report DC/TMD Symptom Questionnaire includes such items as the following: (1) does your jaw click? and (2) do you have pain or stiffness in the morning? A distress questionnaire (PHQ-9) asks if you are feeling down? The Demographics questionnaire asks about marital status and income. The adaptation of each of these items has potential problems, making their translation less than straight-forward:

- "Click" (the word – in English – for a sound associated with the jaw joint) does not translate well across languages (or even within a language, with regional differences being noted in some languages). In passing, we note that most onomatopoetic terms including such non-medical words such as "meow" (what a cat says in English) and "woof "(what a dog says in English) also do not generally translate directly to other languages.
- "Pain" denotes a continuum of experience, with some languages containing more than one word for this continuum; separate words might identify specific aspects (usually, the lower intensity levels) of that continuum, such that "pain", semantically translated correctly, could inadvertently cleave off levels of aversive experience that might be intended in the original English word "pain" as used in a particular context and desired by the investigator. A term for "pain" should identify a sensory and emotional experience related to actual or potential tissue damage, or describe an experience in that context[11]) but in some languages other terms that relate to "hurt" or "ache" can be hard to distinguish from what "pain" (in English) is intended to represent. "Pain" may also have particular implications in some cultures, with different words needed, depending on intent of the item. And, pain at a particular time of day can have specific unique cultural meanings in some languages, again requiring close attention not to the semantic translation but to the context under which the item is intended to be administered.
- "Feeling down" refers to a spatial analogue to a feeling state. In some distress questionnaires (e.g., SCL-90), the word "blue" (as a characteristic of depression in English) is used for "feeling down"; both concepts are particularly interesting examples demonstrating the necessity for contextual review following semantic translation. "Blue", as a characteristic of depression, appears to only be relevant in the English-speaking world, so its literal translation would not be useful as an indicator of a depressed mood state in another linguistic setting that does not associate this color with feeling depressed. Similarly, feeling "down" may not translate correctly. Note, however, that "blue" itself, as a color, can be easily and correctly translated into most, if not all, languages, but the intent of the item would be lost; the challenge of translation would be to find the correct colloquial term that best identifies that phenomenological state of dysphoria.
- Seemingly, the last example about marital status and income level should be predictably translated into a culturally equivalent item based on semantic

equivalence; however, marital status has cultural standards, whereas the intent in a medical assessment tool is not the legal status but rather the relational status of the person. Income levels need to be normed to a given setting. Neither of these two questions are translation issues but rather reflect cultural adapatation.

Similarly, the DC/TMD clinical examination includes verbal commands to the subject, such as "open your mouth as wide as you can, even if painful", which may require special attention in translation and review in order for the command to the subject (patient) to accurately convey the intention of the assessment. In sum, an instrument in this document refers to any operation that includes language (verbal or written) used for the purpose of systematically collecting information from another person.

The above examples demonstrate why a translation must be performed by a number of individuals, acting independently, and why their produced translation needs to be reviewed by another set of individuals not involved in the translation process; the cultural and semantic challenges in creating a valid instrument are just as critical as any other aspect of clinical research.

Based on the International Quality of Life Assessment (IQOLA) Project [12] and the International Test Commission [13], as well as other sources, and summarized very nicely in the test development manual from Beaton [14], instrument development for multi-national, cross-cultural research consists of two major phases: (I) translation and cultural adaptation, and (II) translation validation and documentation [15 16 17 12 18 19].

- Phase I: Translation and Cultural Adaptation consists of 10 stages, with each stage resulting in written documentation via the translation Log:
  - (1) forward-translation;
  - (2) synthesis and resolution of discrepancies from 2 or more forward-translations;
  - (3) back-translation;
  - (4) independent review of back-translation vs source document;
  - (5) revision and iterative development related to discrepancies;
  - (6) consolidation of all translation and review activity into a single instrument appropriate for internal review;
  - (7) expert committee review and cultural validity revision;
  - (8) construction of a pre-final instrument;
  - (9) independent review of the translation process and documentation, and
  - (10) posting the translation so that others can begin to also contribute to Phase II.

- Phase II: Translation Validation and Documentation consists of 7 stages, with each stage resulting in written documentation via the translation Log:
  - (11) pre-testing and instrument review,
  - (12) field-testing,
  - (13) instrument revision,
  - (14) formal assessment,
  - (15) score standardization,

(16) validation research, and
(17) multi-national user manual.

For practical purposes, Phase I encompasses the creation of a culturally-adapted instrument that is ready to be administrated to the prospective population, while Phase II encompasses testing the instrument for its cultural validity, and then collating the empirical information into a user manual. At the end of Phase I, the instrument and the accompanying development documentation can be presented to an external reviewer (for example, the Translations Committee of the Consortium) and if accepted, the translation can then be posted (for example, on the Consortium web site) for distribution to other researchers who can now participate in the verification of cultural adaptation, scaling and validation process necessary for publication. Instrument translation experts regard pre-testing and field-testing as necessary stages before cultural adaptation is completed, and we agree. Only at that stage is a translated instrument suitable for submission to a journal for publication. However, in order to facilitate that process and in recognition that for many instruments translation to a language (e.g., Spanish, Arabic, English) does not necessarily delimit the cultural setting where such adaptation needs to be assessed, recommendation and approval by an external expert review committee would appear to be a sufficient minimal level of completion of a translated instrument that can now be implemented in other settings in order to more rapidly facilitate development for wide-spread application. Such release of an instrument is of course qualified by recognition that the instrument is still in development and that users at that early stage will need to participate in the completion of the Phase II studies before the instrument can be considered valid.

No one qualitative or quantitative step appears to be sufficient for ensuring valid translation and cultural adaptation of an instrument, and consequently multiple methods and checks are required [12]. Consequently, this document is entitled “Guidelines” in recognition that the creation of a culturally valid instrument cannot necessarily be performed by strictly following a pre-specified set of instructions (commonly regarded as a cookbook); each instrument, in each setting, as operated upon by a particular set of individuals comprising a unique combination of perspectives and skills, will present a potentially new situation requiring a new combination of approaches in order to create a new target instrument in an efficient manner consonant with the available resources and validity needs. For example, some Team Leaders might want to insert an additional pre-test qualitative evaluation of an instrument, prior to Expert committee review, in order to assess, as an example, the range of idiomatic expression anticipated in the multiple dialects likely to be encountered in the proposed research setting. This is a common need in areas where literacy is low. In other settings, where anticipated instrument respondents possess high socioeconomic status and who are medically literate, compressing the pre-test and the field-test together for a given instrument after the expert committee review might be most reasonable; alternatively, the developers might even side-step the pre-test and field-test in favor of moving directly to the formal validation research simply because the Team Leaders of the language translation might already have achieved a sufficient level of personal knowledge of the person characteristics in their population. In sum, the Team Leader needs to know the instrument being adapted, and the population for which it is being adapted. Because of the varying complexity of instrument translation and adaptation for another culture and language, instrument developers should consider, at the outset, inclusion of a consultant who will monitor the design of the entire process. This might avert particular forms of frustration later when

the Team Leader attempts to publish the target instrument, only to find that journal reviewers have a keen eye for potentially confounding elements in the new instrument.

The following series of recommended steps for instrument development are derived from methods that are established in the test-development and cross-cultural methods literatures. This set of steps should be regarded as *guidelines*; following each and every step exactly as described does not necessarily guarantee a valid culture-free instrument, while the development of a valid culturally equivalent instrument could nevertheless result despite (or, perhaps, because of) departures from these specifications. Moreover, there is a paradox involved in instrument development[20], which means that inappropriately strict adherence to these guidelines, in terms of having the goal of creating a perfectly equivalent target instrument, could actually result in an adapted target instrument that is *too much* like the source instrument, and consequently the adapted instrument might not be adequate for the intended cultural or language setting, simply because it might miss assessing the phenomenon of interest as it exists *in that setting*. Local resources, available subject populations, nature of the construct, prior instrument development results, and particular characteristics associated with a given language and culture are all relevant to how rigorously, or not, these guidelines should be followed for the development of a particular instrument in a particular setting. Consequently, multi-national collaboration during adaptation of a target instrument is highly recommended as a method for maintaining best practices in order to achieve the most culturally equivalent instrument in light of this paradox.

The development procedures leading to a final instrument must be documented so that other instrument users can evaluate the process of instrument development and determine whether the resultant target instrument is suitable for a particular purpose. The determination of validity of a target instrument remains an ongoing process. A companion document, Summary of Translation Procedures, contains specific instructions for each step, as well as logs for monitoring the translation and equivalency processes and summary forms for submitting a translated instrument for Committee review.

Methods for Establishing Cultural Equivalency

Preliminaries: Create a Translation Team

The individuals who will participate in the translation should be identified at the outset in order to clearly define the necessary roles such that sufficient blindness and objectivity are maintained from initial translation to final review. Based on translation experiences by many collaborators and across a wide range of settings, the following roles have been identified, and the activities of these roles are further described subsequently in the Guidelines.

- Team Leader: The Team Leader creates a translated instrument from the source into the target language. The Team Leader should have sufficient knowledge regarding the content of the items to be translated and any areas of insufficient knowledge should be anticipated at the beginning; specific selection of Independent Panel Members (for Stage 4: Expert committee review and revision) should be accordingly anticipated.

- Coordinator: The Team Leader may wish to appoint a second individual (Coordinator) to direct the activities of the translation team members and to monitor the construction and movement of the translation Logs. Based on experience, identifying a separate person for this role can greatly facilitate the rate of completion of a translation and, perhaps more importantly, permits the Team Leader to have additional roles (as described below).

- Forward Translator: The Forward Translator receives the applicable Log (B) from the Team Leader and renders a translation from the source into the target language. The Forward Translator must be fluent in the Target language as his/her first language, and the source language is a second language for the individual. Sometimes, translators have grown up in a dual-language household, particularly if the translator's parents were immigrants, and it may not be clear which of the two languages is the individual's mother tongue. In some situations, the mother tongue is considered the language the individual first spoke; in other situations, the mother tongue is considered the language the individual feels most comfortable with. When translation involves written documents, another criterion appears to be sufficient: which language was the first *written* form that the individual learned? This appears to be a pragmatic criterion for selection of an adequate forward translator. For translations that are complex or critical, utilizing two independent forward translators is highly advantageous in terms of both quality and efficiency when considering the overall process from start to finish.

- Back Translator: The Back Translator has characteristics that are the inverse of the Forward Translator: mother tongue is the source language, and the person's second language is the target language. See the Forwards Translator section for criterion regarding which language is the "first" language when the individual was raised in a dual-language home. Depending on how critical the instrument is, one Back Translator may be sufficient, with items marked as discrepancies relayed to the Expert Committee for independent review. If the instrument content is particularly complex or critical, it may be wise to use two Back Translators; this also facilitates the problems sometimes encountered with further Forward Translation of any items that require revision, as agreement between two

Backwards Translators that a forward translation resulted in a discrepancy between source and back-translation provides a benchmark for the next step.

- Reviewer: The Reviewer is the third critical member of the Translation Team and the role is often not recognized for its importance. Specifically, the Reviewer should be fluent with the source language as a mother tongue and should be an expert in the item content so that he/she can adequately compare the back-translation against the source document and decide if the translation was correct semantically. The Reviewer makes recommendations to the Team Leader (and hence, to the Forward Translator) regarding areas in the translation process that are questionable. Obviously, the Reviewer cannot be either a Forward Translator or a Back Translator. In order to expedite translations, the Team Leader might adopt the role of an unofficial reviewer and facilitate another cycle of forward translation and backwards translation, in addressing problematic items. This is acceptable practice as long as there is an external Reviewer who meets the criteria stated here and who will then review all documentation of the translation process. That external Reviewer should then be identified as the Reviewer for the translated instrument. This assures independence of the review process.
- Independent Panel Members: These individuals are selected in the target language geographical area, and their language skills (target vs source as first language) is a matter for the Team Leader to consider when comprising the Expert Committee. Instrument content, setting, and intended applications will determine the composition of the Expert Committee.

To summarize, translation teams are comprised, at a minimum, of a Team Leader, Forward Translator, Back Translator, and Reviewer. In this constellation, the roles of the Team Leader include creation of the Logs at each stage, synthesis of the two forward translations (if two translators were used), formation and chairing of the Expert Committee, interaction with the Reviewer regarding accuracy and any problems with the translation or back-translation, creating a final instrument for administration, and finalizing the documentation to be submitted to the Consortium. The Team Leader is often also a Forward Translator, and potential problems regarding blindness in instrument development should be considered at this stage of creating the Translation Team.

If there are problems that need to be addressed via the back-translation vs the source instrument, independence of evaluation can be compromised in two ways if the Team Leader is also a sole Forward Translator: (1) the Team Leader/Translator can be biased regarding the forward translation and hence not adequately consider the implications of what emerged in the back translation, and (2) the Team Leader/Translator will no longer be blind to any items that require additional translation (and another round of back-translation) since the Team Leader/Translator will have had perhaps extensive contact with the back translation. Solutions to this problem include the addition of a second Forward Translator, where agreement (and resolution of discrepancies) assures independence of translation, or including a Coordinator who will take over specific roles of the Team Leader.

Finally, the conduct of reliability and validity studies of the translated instrument may require a new set of individuals.

Phase I

Stage 1: Translation

Instruments not yet translated are forward translated from the source language (e.g., English) to the target language using standard methods [21 20 22 23]. Two bilingual translators (sometimes referred to as "forward-translators") whose native language is the target language, with the source language as their second language, will produce independent translations. Ideally, the forward translators reside in the country of the target language, in terms of being current with idioms and cultural usage. Two factors are important in selection of a forward translator: content knowledge, and bicultural status. One translator should be knowledgeable of the instrument content, while the other translator should be naïve to instrument intent and hence culturally representative of the subject population who would be using the instrument [20]. Parallel to the form of bias that content knowledge can lead to in a translation, possible bicultural status of a bilingual translator should also be considered for the tension that it creates as a potential source of undesired bias on one hand (e.g., the translator may unwittingly translate the term according to its usage in the source language) vs potentially greater facility at correctly rendering the intent in the source language to the correct usage in the target language. Bicultural status, for example, often emerges when the forward translator has lived for some time in the country of the source language. The selection of translators deserves attention, as translators with in-depth knowledge of the setting may provide more nuanced translations [19], and attention to these factors early in the development process minimizes problems subsequently in back-translation review and adaptation.

In addition to translation of items, translation is needed for other parts of the instrument: instructions, response categories, and scoring rules. Of particular note is response categories; they can be very sensitive to nuance in language yet have great impact on relative response frequency of endorsement [24], leading to differences in scaling and scoring. Further discussion is given to response categories in the section on Data Analysis.

Certainly, only one forward translator could be used; our experience, however, is that overall instrument development including semantic representation of concepts from the source language to the target language as well as cultural representation is facilitated by incorporating two forward translators at the outset, as described. This extra step at the outset eliminates a number of problems that otherwise have a predictable occurrence later on as terminology and phrasing come under evaluation. Moreover, the selection of two translators with differing backgrounds, as indicated, increases the probability that the target word selection and phrasing will contain both technical accuracy as well as common-sense usage. Some translation programs (such as the translation of the DC/TMD protocol) formally require two independent forward translators, for the reasons stated above.

Stage 2: Synthesis and resolution of discrepancies

If two forward translators are used, the two translators convene and produce a single translation. They will consider conceptual equivalence, colloquial language use, and clarity [6] as they combine the two translations into one. Depending on the choice of individuals for the two translator positions, the desired balance in negotiation and consensus might occur during the synthesis stage; or, if the two individuals differ in their style of dominance, the Team Leader might need to be present in order to facilitate the

synthesis from the two translation versions. The Team Leader should always conduct a final review of the synthesis translation before it is moved to the back-translation stage. As a minimum, the Team Leader will need to prepare the translation Log for the back-translator.

Stage 3: Back-translation

Quality control of instrument development is maintained via blinded independent back-translation, which is conducted by a translator whose native language is the source language and whose second language is the target language. Based on our experiences, the back-translator may reside in either source or target countries; suitable back-translators can generally be arranged from various sources such as university language departments, language and cultural institutes, and state departments. The back-translator should ideally be naïve to subject content in order to reduce information bias and to highlight unexpected meanings [15]. It is often recommended that two back-translators be used, followed by synthesis of the two back-translations, but the step of back-translation has some limitations such that the additional resources that could be allocated to a second back-translator are often better utilized elsewhere in the development of a translated instrument. Consequently, we believe that a single back-translation is generally adequate when supplemented with an expert panel (see Stage 4) [17]. Particularly when a source instrument will be translated to multiple other languages, back-translation helps to insure maximal similarity in item meaning in the maximal number of items across languages and cultures. Back-translation also serves to provide one index of clarity, common language use, and conceptual equivalence, by obtaining from the back-translator independent language rating of the translation [12]. Finally, back-translation provides concrete instance of instrument development, which is useful for basic documentation of progress and identifying the nature of any problems encountered in the development process; this documentation can be reviewed by other investigators who may have no skill in the target language.

Stage 4: Review

The back-translation is independently reviewed against the original source by someone not involved in either translation or back-translation. The reviewer is ideally the individual who developed the source language instrument, but this is seldom managed. The reviewer attempts to reconcile discrepancies between back-translation and source; these errors may have arisen in the back-translation rather than necessarily in the forward translation, but determining the source of the discrepancy is *not* the responsibility of the reviewer. The reviewer provides recommendations to the Team Leader for corrections, and the Team Leader must interpret the recommendations and determine if the error was in the forward-translation or in the back-translation; this is one reason why it is best if the Team Leader is not one of the translators (forward or back). This process of review, recommendations, and appraisal by the Team Leader is followed by another cycle of translation and review, as needed. Importantly, this process is documented so that any deviation, and its rationale, between source and target language versions will be available for other instrument users; it is known that some source language items do not translate directly to the target language, some source language items translate readily but the translation does not make sense in a second language, and some source language items may not be important in a second culture [22]. When these types of problems are found, the items should be marked for special attention by the expert panel, and the Team Leader should move forward with other parts of the instrument development.

Errors in translation are defined simply as a discrepancy between the back-translated version and the source version. A recommended standard for bench-marking the adequacy of the translation effort is a maximum of 8 errors per 300 words translated [21]. If more errors than this are encountered with a particular language version, the characteristics of the translators and/or back-translator should be reviewed, and another independent translation cycle should occur in order to resolve the unexpectedly higher error rate. Even more refined translation processes exist by using, for example, quality ratings of translation and sophisticated statistical analyses [25]. There is evidence that combining elements from "resource intensive" and "resource-saving" strategies in a *moderately resource-intensive translation* is able to produce adequate results [26,27] . The reader is encouraged to pursue these references if such strategies are important.

Stage 5: Revision and iterative development

Translation units (typically an item) identified during the reviewer process as exhibiting problems are appraised by the Team Leader. According to the source of the error, the translation unit is returned to the forward translator for revision or to the back-translator for revision. If the translation unit is returned to the forward translator, the revised translation will then be submitted to the back translator, and then on to the independent reviewer. This process continues until the review process identifies no problems.

Stage 6: Consolidation

When all translation units have been satisfactorily translated, the translation units are compiled so that all iterations of development associated with a single translation unit are collected together. This provides a complete record of the translation process and how the understanding in the target language of an individual translation unit may have been completely clear in the first attempt to translate it, or the understanding may have evolved over the course of perhaps many translations and reviews.

Stage 7: Expert committee review and revision

Once a valid translation has been consolidated, a Cultural Equivalency Panel should convene in the respective target language location. The Panel is ideally comprised of experts in language (e.g., linguistics, language instruction, professional translator), methods (e.g., epidemiology, anthropology), and content (e.g., disease-relevant specialist, clinical psychology); some experts should be bilingual (i.e., source and target language) and some should be monolingual (i.e., only target language). Some constructs or instruments may not require a Panel this extensive, and some research sites may not always have access to all of these different types of experts. The role of each type of member in such a Panel should be considered in all cases, with the instrument Team Leader selecting individuals who can best address the respective concerns about cultural equivalency of the items in a particular instrument in a particular setting.

The Panel ideally meets as a group, but this is clearly not always possible. Again, the Team Leader must make the decision regarding the objectives of Stage 7 and whether they can be met via (a) selective review of particular instrument items by each Panel member who then discusses it with just the Team Leader, (b) review of all content by each Panel member, who then discusses it with just the Team Leader, or (c) review of all content by all Panel members who meet together with the Team Leader. Option (c) is clearly preferred simply because group-based review has a much higher probability of identifying unique aspects of problematic items as well as creating better solutions to

those items. If the Team Leader chooses either option (a) or option (b), the full responsibility of integrating cultural concerns into item revision belongs to the Team Leader.

The Panel is asked to review the content of the instruments with respect to the four types of equivalences (semantic, idiomatic, experiential, and conceptual, as understood by a 12 year-old native speaker of the target language); these four types of equivalences must be established for an instrument to be used cross-culturally [15 28]. A simple method found to be useful in English for assessing level of readability of standard text is known as SMOG[29], but this method may not work well with single sentence questionnaire items; Team Leaders in other languages would have to assess the suitability of this or any other method for their language. The Panel may make recommendations for translation revision, which could result in the need for repeat forward translation and back translation.

Two particular concerns that may affect the cultural equivalency of instrument items – questionable content and questionable phrasing – need particular attention. Some items in translation will still retain clearly relevant cultural usage, and some items will clearly be irrelevant, regardless of how accurate the translation. For example, "blue" as an item in instruments assessing depression in English-speaking countries can be easily translated into any language yet the correctly translated term is clearly wrong as an indicator for depression-related mood in probably all languages other than English. For such items, a unique substitute that assesses the same (or closely similar) aspect of the targeted construct is desirable and readily performed in some cultural settings. In other cultural settings, however, the fundamental concept in English-speaking countries for which "blue" is an indicator, dysphoric mood, may simply not apply, or it may not be translatable into common language (cf. professional-level use of the language). Obviously, if an instrument contained this type of content, at least one Panel member should have a sufficient background in mental health in order to adequately explore the available and necessary options for these types of items.

In addition to classifying items as clearly relevant or clearly irrelevant, Panel members should also be instructed to include a middle category: instrument items with questionable content in terms of cultural relevancy. Including a third category for classifying item relevancy within the cultural setting leads to examining items that might otherwise be regarded as satisfactory, simply because they are not clearly irrelevant. However, to misclassify such items would be to sidestep potentially important cultural differences between source and target languages in a particular item. That is, semantic or idiomatic accuracy in translation may not adequately map onto the desired experiential or conceptual equivalence that is needed for a measurement scale, but to assess these latter forms of equivalences sometimes requires more attention and focused consideration. Once items have been identified as having questionable content (see next para for method), the content validity of the item should first be assessed for relevancy, in order to determine the priority for item review or revision.

The identification of content relevancy is based on use of a quantitative approach for relevancy ratings from the panel members. One method for assessing relevancy is through the use of the Content Validity Index, CVI[30 31], which starts as a set of ordinal ratings (1 = not relevant, 2 = somewhat relevant, 3 = quite relevant, 4 = very relevant) regarding relevancy, from which the proportion agreement statistic is computed (after

recoding the responses, with 1 and 2 considered together, and 3 and 4 considered together, creating a binary outcome measure of low vs high relevancy). The ratings by the Panel members can be compared for agreement; items with a high rater agreement would presumably be retained, while items with a low rater agreement would be questioned and potentially dropped. Consequently, the CVI provides a method for standardizing the process of agreement/approval of specific items for a given construct across multiple settings, and for systematizing the process of item review for inclusion in the translated instrument. One problem with this approach of basing relevancy assessment on agreement of the raters is that an item might exhibit an overall high average relevancy to the intended construct, despite medium to low agreement, and another item might exhibit high agreement regarding its low relevancy. Since the outcome from the prior example would most likely be to trust the item with the high average relevancy rating (though with poor rater agreement), and not the item with low relevancy about which one is certain, using the average relevancy rating would appear to be the essence of the method. Consequently, one alternative is to use the CVI-based ordinal rating scale, and then simply take the mean of the relevancy ratings from the Panel members. The Team Leader would inspect the relevancy scores and determine a meaningful set of cutoffs for making decisions regarding dropping the item concept, finding a substitute item, or revising the present item.

The Team Leader then makes decisions regarding whether to retain items, revise items, or drop items. The following rating combination of adequacy and relevancy are associated with the suggested next step by the Team Leader:

Relevancy	Translation Quality		
	Inadequate	**Questionable**	**Adequate**
Low	*Drop item*	*Drop item*	*Drop item*
Moderate	*Retranslate*	*Review translation or item content, or both*	*Find better item*
High	*Retranslate*	*Revise translation*	*Retain item*

The final product from the Panel review should be a culturally acceptable translation; items that require input from the target population should be highlighted for the pre-test.

Stage 8: Preparation of pre-final instrument

Based on the prior reviews and recommendations by the Expert Panel members, the Team Leader makes final decisions regarding how a translation unit will be rendered in the pre-final instrument. The pre-final instrument is not considered definitive, as it has not yet been subjected to pre-testing and field-testing, but the pre-final instrument should be considered sufficient for dissemination to a closed group of collaborators so that other investigators can collaborate in this initial testing phase.

Stage 9: External review
For the Consortium to post an instrument, an independent review of the translation process and documentation is performed in order to identify possible problems in the translation process. The purpose at this stage is to confirm that discrepancies were identified, that alternative forms were appropriately considered, that final decisions incorporated sufficient perspectives, and that the pre-final instrument correctly reflects the translation process.

Stage 10: Making the translation available to others
The Consortium is ideally suited for collaborative testing of a newly translated instrument in order for an instrument to be evaluated in a range of settings. Typically, a new translation has not yet been subjected to pre-testing and field-testing (the next two stages), and therefore having no empirical support the new translated instrument should not be considered for publication. The new translated instrument can, however, be considered for limited availability to other collaborators who can facilitate the next stages of development.

PHASE II

Stage 11: Pre-test and review
The revised instrument should be subjected to a series of evaluations, each with a particular purpose relevant to the sample size and effort required. The perspective taken here is that final instrument development can be facilitated via use of carefully planned studies that incrementally increase in complexity so that each level of assessment need only be performed once. That series of evaluations is comprised of subject selection with respect to language, pre-test, instrument review, field-test, and instrument development documentation. Formal assessment of test reliability and validity occurs in Stage 16.

Recognizing that for any particular instrument, given resources, prior instrument development, nature of the construct, and intended use of the final translated instrument, the following set of steps might be modified (e.g., further separated, or combined) with respect to perhaps sample size and /or omitting either the pre-test or the field-test (with the omitted test being rolled into another step). Other considerations for how this next series of evaluations might proceed center around the nature of the instrument content. Psychological instruments always have a focus on reliability and validity with respect to conventional psychometric theory. Medical instruments, in contrast, may need to have as rigorous a focus as a psychological instrument, but again they may not. For example, medical information commonly collected is some type of medical history. Such items (e.g., history of chest pain) are critical for medical assessment, and yet because such items might be solitary there is little that can be typically said about it in psychometric terms other than simple reliability. Yet, from a cross-cultural perspective, reliability and validity are still important. Other kinds of items relevant to a medical instrument might include a clinical examination, or there may be a group of symptom items that coalesce into a quasi-construct. The choice of study design for the pre-test and the field-test, as steps toward the evaluation of the translated instrument, will depend on the item-type of the instrument. An instrument that is, essentially, just single items will benefit moreso from a bilingual assessment since conventional within-sample statistics are not useful in for this type of items.

A suggested set of steps (in principle based on a reasonable pathway balancing efficiency with quality, with the least amount of redundancy when considering a population of tests being adapted) for this series of instrument evaluations is as follows, from which modifications or consolidations may be made:

1. Language base of subject samples. The potential to use bilingual samples at this stage in instrument development raises a number of issues that are generic to the entire instrument development process. Assumptions that underlie the use of a bilingual sample include: (1) that the subjects are equally proficient in both languages, which may not be true; (2) that the subjects are homogenous with respect to education and background context for learning the second language, which may not be true; (3) that no carry-over effects occur from one administration to the next (if using a single-sample design), and (4) that bilinguals and monolinguals are directly comparable, which is questionable because of differing backgrounds that can give rise to being bilingual [32]. Departure from these assumptions may render a bilingual sample less optimal for assessing items that comprise measurement instruments, particularly when educational status might affect scale scores. Designs other than a single group can be employed in order to control for these assumptions; see Sireci [32] for further information. The assumptions of equal proficiency in both languages and of homogeneity of the bilingual subjects with respect to educational status can both be addressed via careful attention to the reading level of the items relative to the general education of the intended target population. Typically, setting a maximal grade level reading ability of the average 12 year old can largely resolve these problems. However, some instruments might be intended for very high functioning individuals, and hence the reading level may necessarily have to be higher; correspondingly, if an instrument is to be used in a setting where subjects have minimal or no literacy, then attention to even easier reading levels might be important or to even consider that a structured oral version may be necessary. Carry-over effects can be controlled, to some extent, by either counter-balancing sequence of administration of the two language versions via randomizing assignment of subjects to each of the two sequences if using a single-sample design, or, alternatively, bilingual subjects can be randomly assigned to either the target language or the back-translated version in order to compare responses and observe difficulties [33]. In some settings, all potential subjects of future applied research using the target instrument might be bilingual, and selecting *only* bilingual subjects for the field-test may make the most sense with respect to all aspects of target instrument testing, at least for that planned use of the instrument. The documentation of the instrument would, in this last example, clearly indicate that the validity of the instrument was established for bilingual individuals.

2. Pre-test. The instrument is administered to a small number of representative subjects, in order to *qualitatively* assess test performance by assessing comprehensibility, feasibility, and acceptance of the instrument at the item level. There are two approaches that can be considered: focus groups and clinical sample testing. The purpose of the pre-test is to highlight ambiguities in item construction and differences in understanding, between respondents vs test developers and consultants, in the intended meaning of the items. The expected outcome of this stage is to revise the wording of problematic items, in order to hopefully improve assessment of the trait underlying the particular item response. While all types of instrument items can be usefully assessed in the pre-test, items that can only be assessed qualitatively (e.g.,

checklist items, examination instruments) are of particular importance during the pre-test stage.

(a) Focus Groups. Some instrument development groups have relied extensively upon focus groups (such as for the pre-test) as opposed to the back-translation in order to better capture the intent of an item and emphasize less a literal translation [34]. We agree with that perspective in terms of cultural adaptation, but we also believe that a place for back-translation remains; in the sequence of steps recommended here, back-translation provides a quality control check, particularly useful for the original instrument developer if monolingual, and archival documentation of the translation that is otherwise non-existent. Nevertheless, focus groups provide a unique perspective on content development of an instrument that is not readily obtainable via any other method, and it is strongly recommended as the better choice at this stage in our overall sequence of instrument development, assuming that the Team Leader (or other individual) has the skills necessary for managing a focus group.

Focus group sizes range between 6-8 subjects, with 8 perhaps optimal in terms of maximizing diversity of opinion and maintaining an active dialogue. Two hours is the maximal period that such groups can meet due to fatigue; consequently, additional focus group meetings or additional new focus groups might be needed if there are many instruments to be evaluated. An assistant moderator is very useful, and the discussion by such groups should be audio or video tape-recorded rather than rely upon hand-written notes. One question that commonly arises in this stage of instrument development is how does the Team Leader know when he/she is "done" with respect to item review. With focus groups, one is "done" when there is agreement within the group about a given item; if disagreement is the outcome of the discussion regarding a particular item, then the item needs further attention with respect to conceptualization and content validity. Further information about conducting focus groups is readily available and should be consulted in order to maximize the outcome [35,36].

(b) Clinical sample. Subject samples should include about 10 subjects [17], and the sample can contain both monolingual and bilingual subjects. For monolingual subjects, the primary objective is a subject interview following instrument administration. For bilingual subjects, the primary objective is to understand discrepant items following administration of both source and target versions of the instrument. Following administration of the instrument, the subjects should be interviewed regarding intended meaning of items vs chosen responses, and possible translation alternatives should be explored [22]. Items targeted by the expert panel should be particularly emphasized by the interviewer; questions to be asked of subjects include, for example, requests for the subject to (1) describe in his/her own terms their understanding of the concept, (2) provide an alternative phrasing for the item, or (3) explain their experience of the concept. With respect to the question of recognizing when the review is "done", inquiry about an item can be considered complete when 8-10 subjects (of this recommended sample size) provide the same feedback, leading to saturation of information.

Following pretest, inappropriate items and items with errors should be clearly identified, hopefully accompanied by recommendations for how the problematic items can be improved. Problematic items are reviewed by the Team Leader and, as needed, further revised by the translator so that the item intention is best reflected by

item content. As a result of pre-testing, some items may be removed from the instrument due to irrelevance in that culture, or items may be altered in such a way that the fundamental item intent is met but in a manner very different from how it was operationalized in the source version within the context of that particular environment. While this type of instrument modification – altering an item to meet the intent – creates significant analysis problems under conventional measurement models, this type of problem is easily handled by Item Response Theory measurement models (see below), and consequently, altering the item to meet the intent should be regarded as completely appropriate in order to develop a set of translated items with content most relevant to the particular culture.

3. Instrument review. At this point, several possible levels of instrument review are available, all of which serve to increase overall instrument validity [12]. These levels include: (a) national, by the Team Leader and his/her review panel, (b) bi-national, by members of the Translations Committee (or other international colleagues) as well as the Team Leader, and (c) cross-national, by having additional site investigators review all translations via the back-translations and support documents.

Stage 12: Field test

For a field-test, the instrument is administered using a sample size of 30-40 [15 17], in order to *quantitatively* assess the more basic aspects of instrument performance with respect to measurement scales. Consequently, this stage is the assessment of adequacy of cultural adaptation of the items in context of the entire instrument. Subjects recruited for this evaluation should be selected on the basis of the following: (1) representative of the usual clinical population in terms of symptom characteristics; (2) no limitations in understanding verbal and written language as used during the administration of the instruments; and (3) clearly members of the cultural setting (that is, non-assimilated immigrants should not be included in this sample). Subjects can be recruited as either monolingual or as bilingual, and the choice should be based less on availability of mono- vs bi-lingual subjects and more on the type of instrument content to be assessed. Checklist items are much better assessed via a bilingual sample, in that within-subject pairwise (source vs target language) comparisons such as percent agreement (or Kappa, if a more formal statistic is desired) can be readily made. Monolingual samples present a much greater challenge for assessing checklist items, in that there are no psychometric statistics useful for such types of items. Rather, more formal statistics of test integrity assume that the items form a viable construct. For such an instrument, testing it in a monolingual sample may be desirable particularly if item content is specifically linked to the cultural identity.

Initial analyses should highlight items that clearly do not work in terms of actual responses. As described under “Stages 5-6: Data analysis”, the analyses (1) evaluate basic item characteristics in the instrument, and (2) assess performance of the target instrument as a whole compared to the source instrument. The sample size of the field test is not sufficient to demonstrate test validity. The field-test provides a final check on what has been developed, before expending the resources necessary for formal evaluation of instrument validity and the documentation of scaling.

(a) Monolingual sample. A monolingual sample will, in practical terms, limit test administration to only those instruments which are measurement scales.

(b) Bilingual sample. One option at this stage is to use only bilingual subjects, in that it permits a more ready analysis of equivalency. There are several conceptual and methodological assumptions underlying the validity of this approach which require attention. The bilingual approach is particularly useful for assessing items that are part of a checklist (i.e., they do not comprise a measurement scale), as an extension from the qualitative assessment in the pre-test.

In using a bilingual field-test sample which is administered the instrument in both target and source language versions, item reliability between language versions can be assessed, in addition to some of the other analyses described under Data Analysis (below). One simple analysis method for cultural equivalency based on the use of bilingual subjects is to assess percent agreement of raw responses to each item from the respective administrations of the target and source language version; 80% is typically taken as a minimal level of agreement that suggests consistency in understanding the item in both languages. Kappa and weighted Kappa statistics can also be used for assessing item agreement at this stage of development. While more complex statistical analyses can be performed such as IRT-based differential item function (DIF) analysis (assuming a larger sample size than that recommended here) as one approach, the percent agreement method provides a ready index of comparability of the item across the two languages. The goal at this stage is to evaluate discrepant items, not necessarily assess measurement properties of validity. However, as stated at the beginning of this section, depending on how the next stage of instrument development might be construed, more formal assessment using a larger sample size could occur at this point in instrument development; this would lead more readily to more sophisticated statistical analyses regarding item validity in the target language.

The interpretation of any analysis from data collected at this stage, whether the psychometric statistics described under Stages 5-6 from a monolingual sample or percent agreement or a Kappa-type statistic from a bilingual sample must take into account the case mix of subjects selected for this stage of instrument development. The criteria for selecting subjects at this stage must be carefully considered, given the relatively small sample size. While some asymptomatic subjects (i.e., individuals with an absence of the trait or characteristic being assessed) might be desirable, depending on the nature of the target instrument to be tested, the symptomatic subjects are typically more useful in terms of retaining adequate item variability in order to have unbiased estimates.

Interpretation of item analysis must also take into account the character of the item. All of the translation steps in Phase I are intended to create an item that is as equivalent as possible to the source item, taking into account any specific characteristic of the target culture relevant to the intended content validity of the item. However, the translation process, as good as it might be, is nevertheless a judgment, and the validation of that judgment occurs in this step. A formal review of the item content has been recommended at this stage in order provide qualitative information regarding each item as a supplement to the quantitative information that emerges from the target population interacting with the instrument. When Phase II is administered on an instrument that was translated by others, and where the deliberations underlying the construction of each item may not be known, the qualitative review of the content validity is all the more important. Specific procedures for the completion of this part of Stage 12 are described in Appendix 3 and documented on Log QTR.

Finally, scaling of threshold values for rating scale response formats should be done in order to insure that the choices of words for such scales in the target language vs the source language (e.g., fair, poor, good, excellent) produces consistent responses across both source and target languages; IRT models provide a ready method to make this assessment quantitatively.

Stage 13: Instrument revision

At this stage, the instrument is formally finished; a complete working draft of the instrument is compiled. Characteristics of the translators, back-translators, review panel, subjects in the pre-test, and subjects in the field-test should be carefully described. This information will be included in the translation documentation for the instrument. This information can be submitted to a review group at this stage in order for further research to be facilitated by not only the development group but also by other interested groups. The translation Logs contain forms for documentation of the cultural adaptation process.

Stage 14: Formal instrument assessment

The instrument is administered in a much larger sample (see Sample Size Determination, below) in order to formally evaluate test validity properties, including scaling, final item retention, and differential item functioning. Sufficient statistical power to evaluate actual instrument performance is of concern at this stage. The critical component of this step lies within the statistical analysis of the instrument that has resulted from all prior steps. Items may again be dropped, but equally likely is the emergence of specialized scaling and scoring rules that permit the instrument to be uniquely suited for the particular culture/language for which it was adapted.

Subjects should again be selected according to the same considerations as listed for the field-test, i.e., representative case mix, adequate language skills in the target language, and unambiguous member of the cultural setting. Of note at this stage is whether subjects who are bilingual with respect to the source language should be considered for inclusion into the formal instrument assessment. Since many cultural settings contain individuals who are bilingual if not polylingual, and where a bilingual subject may be bilingual in the target language *and some other language but not the source language*, the most sensible principle to maintain at this stage is to insure that the subjects are representative of the individuals planned for further study with respect to the clinical condition and other relevant psychosocial factors and that the subjects meet the appropriate definition of cultural identity. Cultural identity has no single definition; a *minimal* standard would include all of the following: (1) at least one parent and the subject were born in that culture, (2) the subject spoke the host (target) language at home while growing up, and (3) self-identity by the individual as a member of that culture (as distinct from ethnic identification or citizenship). Team Leaders may want to make these criteria more rigid, depending on the instrument and the intention with respect to the particular research question as it pertains to cultural concerns.

The goal of the preceding steps is to achieve equivalence between two instruments (source and target). But there is a paradox in performing translation in order to achieve equivalence, which requires some consideration in making final decisions about item retention, modification, or elimination [20]. If the requirement for equivalence is that comparable results are obtained from both target and source versions of an instrument, then the more equivalent the forms become the less likely one will find differences

between the two cultures. The methods described above appear to provide the best basis for evaluating and revising an instrument such that item content is best suited to the target population and thus reveals any real differences. The final aspect of this paradox is bridged through data analysis methods of item response theory and the associated assessment of differential item functioning, described in the next section, which permit a particular language version of an instrument to retain unique characteristics (through the use of subsets of items for comparison between source and target versions) and yet still provide commensurate data.

Stage 15: Standardize scores across cultures

The normalization process for commensurate measurement is based on pair-wise evaluation of instruments (any two languages), using common items across the two languages. A minimum of 4, but preferably at least 20% of the items in each scale, should be common across the languages under consideration. One fundamental decision is whether raw scores or IRT-derived logit scores are used as the fundamental metric. The advantage of raw scores is that they are easy to compute and can be instantly compared to reference tables; the advantages of logit scores are that (1) missing values require no special treatment with respect to creating a summary score, and (2) scores across diverse instruments are, in principle, comparable – if the assumption of local independence is true for the respective instruments. Raw scores can be, however, perhaps equally robust to missing data if the mean score (across items), rather than raw score, is used.

If all steps outlined above have been performed carefully and no items with DIF have been found, then scores across different language versions of the identified instrument will generally be directly comparable. If the items fit a Rasch model, raw scores are sufficient as the standard scoring method. If the items does not fit a Rasch model but fits a general IRT model, the raw score is still a simple and robust scoring method – provided that the difference in slope parameter is not too large, as evaluated from the IRT model.

If DIF is found between language versions, IRT-based scores will still be comparable if there are sufficient common items to establish a robust anchor and if language version-specific item parameters have been established for the items with DIF. In this case the raw score will not be directly comparable between countries, and equivalence (cross-walk) tables will need to be constructed that for each country link the raw score to the IRT score. Such tables can be developed from the IRT model through the computation of predicted values for each item.

In addition to IRT approaches that may be more robust with respect to instrument assessment across different groups, resampling approaches developed for cross-validation may also be applicable for these problems in that the results will be more robust [37]. A variety of methods can be used for the bootstrap approaches which are particularly useful when instrument cut-points (or threshold values) are being estimated. One method, for illustration, is termed 10-fold cross validation. In this method, 100% of the data are used to estimate an initial threshold; then, the sample is randomly divided into 10 groups, the analysis is performed on 9 of the ten groups, the new estimated threshold is compared on the remaining group with respect to who is above vs below the threshold, and the process then repeats nine more times. The initial estimate of threshold, using all of the data, is then compared to the aggregate outcome, using the 10-fold cross-validation procedure, for determining how robust the initial threshold estimate was. A typical sample may be as small as 100 subjects for this method to be useful.

Stage 16: Validation research

Validity of an instrument has traditionally been regarded as emerging from convergent and discriminant validity evaluations based on correlational structures of an instrument of interest compared to other instruments measuring similar and/or dissimilar constructs. The pattern of results in one language can then be compared to the pattern as obtained in another language [19]. Newer conceptions of methods appropriate for the development of a valid instrument for cross-cultural research indicate that IRT models provide an additional assessment of an instrument's validity, with other methods (such as confirmatory factor analysis, as described under Data Analysis, Method 2) contributing to this overall assessment [38 39,40].

Stage 17: Instrument documentation

Based on the empirical documentation developed in Phase II, as well as the results from the subsequent data analyses, the instrument characteristics should be carefully summarized into a user manual which includes the analysis results from each target language version.

Methods for Modifying an Existing Translation

1. Any modifications, for whatever purpose, in the current existing translation of an instrument need to be accompanied by the same translation steps, as described above: independent translators, independent back-translators, and independent reviewers of the back-translation against the original. The independence of reviewers is stressed, as challenges to an existing translation should be accompanied by a stringent process leading to high confidence level that the revision represents an improvement.

2. When the translation being challenged is one previously conducted via the Consortium, collaboration among the relevant individuals (e.g., the original translator, and the new translator who identifies problems) is the only practical method for addressing identified problems and improving an instrument via consolidating the two translations.

3. When the translation being challenged is one previously created by individuals not affiliated with the Consortium, the best method is to contact that translator and determine if collaboration towards a revised translation is possible. If the translator of the existing translation is unknown or if collaboration is not possible, then the translator of a potential revised version of the instrument should proceed forward, with the goal that the subsequent steps outlined here will provide the necessary support and documentation for the improved translation.

4. Challenges to a translation will typically call for a Cultural Equivalency Panel, even if the challenge to the translation did not intend to include that step. This specific step is recommended in order to address original intent in the revised translation when trying to decide upon the existing translation vs the suggested new translation. If disagreement about the correct translation (whether semantic or cultural) persists between the original translator and a new translator of that instrument, then a third translator may need to be included in order to consolidate the versions. When the original translator is not available, then the third translator becomes equally important for peer review and consultation. The emphasis is that the translated versions need to be consolidated rather than considering one version vs another. Collaboration is key for the Consortium at all levels of activity.

5. A Translation Summary document should be constructed, as described above, documenting the changes to the existing translation. The Face pages should provide users with a clear indication of a new version, with guidance to the older version. Finally, the user can benefit from an appendix to the translation, containing a summary of the changes that were made.

Data Analysis

Data analysis is primarily focused on item responses, as a characteristic of an underlying trait, using two separate but intersecting methods, conventional test analyses coupled with Item Response Theory analyses. Preliminary item analyses for cultural equivalence can be conducted using the field test data (n=30-40); further, and more definitive, analyses should be performed using larger sample sizes, in order to evaluate the data from operational, scalar, functional, and metric perspectives [41]. Analyses from the field test data are described first, followed by a description of the additional analyses using a larger sample size.

Summary of Methods for Analysis of Translation, Cultural Equivalency, and Validity		
Stage	**Statistic**	**Purpose**
Field-test (n = 30 – 40) *Overall Objective: Item review, in order to eliminate, revise, or retain an item*		
Method 1	Internal consistency	Flag target language items with markedly different reliabilities compared to source language for review
Method 2	IRT: person-fit statistics	Assess translation adequacy
Clinical Evaluation (n of approximately 200 per setting; see Sample Size section) *Overall Objective: Assess dimensionality of construct and discard or dummy-code biased items; develop scoring algorithm across language versions*		
Method 1	Internal consistency	Assess internal fit of items of target language version compared to source language
Method 2	Confirmatory factor analysis	Compare factor structure in target vs source language versions
Method 3	Pearson correlation	Convergent and divergent validity assessment
Method 4	IRT: item-fit statistics	Verify and refine measurement scale
Method 5	Differential item functioning (DIF)	Determine whether items functions in the same way (i.e. same difficulty and discrimination) in target vs source language versions

Overview of Measurement Models Based on Frequency Responses. Measurement models that rely upon frequency of endorsement are typically referred to as item response models or as Rasch models. While the two approaches, Item Response Theory (IRT) and Rasch modeling, share a number of similarities, there are also some foundational differences. From the IRT perspective, IRT approaches can include 1-, 2-, and 3-parameter models; for these applications, the 1-parameter (also often termed the Rasch model) and 2-parameter models or their extension of polytomous data [42] [43] [44] [45] are most likely sufficient. The IRT model predicts the probability of selecting each response choice of an item as a function of person characteristics (e.g. ability) and item parameters (e.g. item difficulty and item discrimination). The easier an item for endorsement, the more likely it

is endorsed; the more able a person, the more likely he/she will endorse the item compared to a less able person. The probability of item endorsement then is a logistic function of the difference between the person ability and the item difficulty. The model handles dichotomous as well as rating scale data, and it also handles missing data [46] .

Aside from the difference in the number of parameters which distinguishes the Rasch model (which is a 1 parameter model) from the 2- and 3-parameter IRT models, there is a fundamental conceptual difference: advocates of the Rasch model view it as a strong modeling approach to measurement: either the data (as generated by the items appropriate to the underlying latent trait) fit the measurement model, or they do not. If they fit the model, then unidimensional measurement is considered a direct outcome of the selected items, and if they do not fit the model, then some other model needs to be considered for the development of a measurement scale. From the Rasch perspective, the inclusion of a second (or third) parameter only creates noise that detracts from unidimensional measurement possible when relying on only one parameter. From the Rasch perspective, the 1-parameter model necessarily results in interval-level measurement; the only question is whether the data fit the model or not. In the IRT approach, an IRT model (either 1, 2, or 3 parameters) is selected in which to model the data. There are many published arguments pertaining to which approach, IRT vs Rasch, is acceptable; as our aim is to discuss analysis as the means to the end-point of developing a measurement scale, other published literature must be consulted with respect to which approach one might want to adopt. For our purposes, we will subsume the Rasch approach within IRT, broadly, in order solely to simplify discussion and explanation. This does not imply a preference for one approach over the other.

A number of fit statistics have been developed in the IRT framework in order to evaluate the fit of each item [47-50] as well as the assumption of unidimensionality and local independence [51,52]. Most fit statistics use estimates of person ability and item parameters to construct an expected response pattern that is compared with the observed pattern. Further, fit statistics have been developed that evaluate the fit of the response pattern of each person in order to calculate an index of response consistency for each person (person fit index) [53 54].

Popular fit indices for the 1-paremeter model are the information-weighted fit statistic (infit) and outlier sensitive fit statistic (outfit); infit refers to how individuals with threshold values close to that of an item respond to that item, and outfit refers to how individuals with dissimilar threshold values respond to that item. For infit and outfit, 95% of the fit statistics have expected values between -2.0 and +2.0, and statistics outside these limits indicate poor fitting items. The 2-parameter model adds a discrimination function which allows items to differ in how acutely they differentiate respondents. Together, these five statistics of item value, error, infit, outfit , and discrimination lead to assessment of how well (and which) items fit a unidimensional construct. The model also produces threshold scaling values, which permits an evaluation of the adequacy of translation of response choice labels [24].

Interval level scoring derived from the logit transformation of the raw item response is an often stated advantage of IRT models, particularly in the Rasch model. Note that the possibility of interval level scoring is based on the presence of a continuous latent variable; there is, however, no direct way to establish that the underlying latent variable is continuous. One method, though cumbersome and also possibly unavailable, for assessing whether the underlying latent trait is continuous is to increase the number of items that are included in the instrument, rescale with an IRT model, and assess whether

the data still fit the model and the number of latent classes. This process is continued until the number of items is increased such that the number of latent classes now exceeds the original number of interval steps. If the data continue to fit the model at that point, then it is probable that the underlying latent variable is continuous. More simply (though not directly empirical), if it makes conceptual sense to regard the underlying latent variable as continuous, and if in any of the analyses the hypothesis of a continuous latent trait cannot be disproved, then it seems safe to conclude that the underlying latent trait is continuous, and therefore further conclude that the instrument scale is of interval quality.

IRT models provide a sensitive and flexible approach to investigation of cross-cultural equivalence through the testing of differential item functioning (DIF): (1) IRT methods can evaluate cross-cultural bias both in form of differences in item difficulty (also called uniform DIF [55]) and in item discrimination (discrimination is equivalent to loading in factor analysis and is also called non-uniform DIF) [55,56]; (2) the number of items can be reduced in order to improve instrument efficiency; (3) items unique to a particular language or culture can be retained via dummy coding if item content is essential to the measurement of the construct in that setting; and (4) the IRT model provides a metric for cross-calibrating of scale scores across cultures, even if some items are unique to a specific culture [57].

Standard IRT approaches to DIF analyses fit an IRT model for each subgroup and then test the assumption that the item parameters are equal across subgroup through a likelihood ratio or a chi-square test [56-58]. Significant differences in item parameters provide evidence that respondents from different Groups (e.g., gender, cultures, age) endorse an item differently, controlling for overall severity in the measured construct. A minimum sample size of 100 persons in the smallest group is recommended in order to achieve sufficient statistical power with this approach [59,60]. With smaller sample sizes, cross-cultural measurement equivalence can be evaluated through person-based fit statistics if an IRT model has been estimated with adequate sample size and found to fit in one culture [61].

In sum, for purposes of developing instruments for cross-cultural use, instrument development that includes IRT analytic methods may be more robust [62 63 57]: (1) cross-cultural bias can still exist despite very similar factor structure (and hence, adequate internal consistency); (2) the number of items in an instrument can be reduced in order to improve instrument efficiency; (3) items unique to a particular language or culture can be retained via dummy coding if item content is essential to the measurement of the construct in that setting; and (4) the IRT model may lead to interval level scoring thus providing a more ready basis for commensurate scales across cultures [57 64].

Analyses using field test data (n=30-40)

Method 1: The first method is based on classical test theory [5]. Conventional item statistics of internal consistency using Cronbach's alpha are performed and the secondary item vs whole (minus that item) correlation is calculated. Items with markedly different correlations, in comparison to statistics derived from the source language version, should be flagged for review. This evaluation of item correlations, however, is highly contextual, depending on the pattern and magnitude of internal reliability coefficients in the source language. We recommend as a threshold for item review a change of 0.3 in the item-whole coefficient of the target item, compared to the statistics associated with the source version; simultaneously, it is expected that well-fitting items will have an an

item to scale correlation of at least 0.6 [28] though some sources recommend a threshold of 0.40 as acceptable [6], but ideally will have a pattern and magnitude overall similar to that of the source language. This value is chosen based on correlation of 0.3 accounting for approximately 10% change in covariance between the targeted item and the remaining items. The overall internal consistency coefficient (α) should exceed 0.70 [6]

Method 2: Using the field test data, the IRT models can only provide an initial assessment of item fit because the suggested sample size results in too little power to reliably detect misfitting items. Consequently, only a preliminary analysis of model fit can be determined. For instruments with an established IRT data in the source language, an IRT analysis can be conducted with focus on the person-fit statistics [61] Using item parameters from the source language person-fit statistics can be computed from data in the target language. If person fit in the target language compares well with that in the source language, it is probable that the items have been translated well. For misfitting persons, problematic items can be detected by analyses of the response pattern.

Items exhibiting poor internal consistency or poor fit in terms of misfitting persons should be considered for reevaluation by the expert panel for elimination, revision, or retaining the item as is. Strict quantitative assessment of item fit, however, cannot be reliably performed at this stage. Thus, the instrument review and revision for cross-cultural equivalency will contain a qualitative component (expert panel; and pre-testing by target population), a quantitative component for instrument reliability, and a more in-depth probe into possible translation problems.

Analyses using larger subject samples (n of approximately 200 per setting)
Data analyses for more definitive assessment of cultural equivalency of an instrument include models that focus on covariance relationships as well as on item parameters. Analyses at this stage should consider using gender, age, educational level, and culture as between-subjects variables for assessing differential item functioning (DIF) and overall validity of the instruments [17]. At this stage, the paradox of equivalency becomes more important, in that misfitting items must be evaluated qualitatively as well as quantitatively.

Method 1: The analyses of internal consistency, as described previously, should be repeated in order to confirm, using the larger sample size, prior findings.

Method 2: Regular factor analysis provides a simple approach to the structure of the data via Eigenvalue decomposition. Confirmatory factor analysis (CFA) is used to test factor structure in unidimensional instruments, in terms of whether the same single structure exists in a target language. For multi-construct instruments, CFA should be performed in order to confirm existing findings of the multiple factors [8]. Confirmatory factor analysis for cross-cultural validation, using models that allow for categorical responses [65], allows comparison of factor structure that is more in keeping with actual types of raw data that are collected. CFA also allows an evaluation of the assumption of local independence via the residual correlation matrix. The resultant factor structure should be clear and replicate the original scales based on the same items and comparable variances [6].

Method 3: Convergent and discriminant validity testing across, respectively, similar and dissimilar instruments can also be used, in order to confirm that the respective constructs remain unique within each culture [19 39]. Correlation coefficients of at least $r = 0.50$ are recommended [6] for the convergent validity. Discriminant validity can also be based on known groups validity. A target language instrument that performs similar to the source

language instrument and which has excellent psychometric properties can be assumed to be culturally equivalent [6].

Method 4: The IRT models should be used for verification and refinement of the measurement scale within each culture, that is, for scaling and not for factor delineation. Within the IRT model, item fit statistics [47-50] can evaluate the fit of each item, while the assumption of local independence among the items in the model can be tested through other statistics [51,52]. The absence of such associations, in conjunction with good item fit, support a unidimensional construct and allow for commensurate measurement instrument across cultures, even if optimal measurement of the construct across cultures requires that some items be allowed to float independently within the measurement scale due to differing interpretations of the item across any pair of cultures.

Method 5. With the larger sample size, it is possible to perform DIF analyses [66 67]. The goal is to evaluate each instrument from each culture in reference to the source language version through pair-wise comparisons of target version with source version. This permits development of unique instruments for each culture, allowing items that do not fit the measurement model within a particular culture in reference to the target culture to either not be included in the instrument-as-used, or to allow items that ostensibly measure the same construct to nevertheless be included in the instrument but unlinked in terms of scaling across two cultures; that is, the item can remain within the instrument (measurement model) but float independently across the two cultures. To require each cultural version of an instrument to be identical across all cultures, with respect to all items, runs the risk of there not being any items in common for the more complex constructs such as QOL [10]; in contrast, allowing items to uniquely remain in the measurement model leads to more generalizable instruments [57].

In order to assess DIF, an iterative "top-down purification" approach can be used and is recommended as being perhaps the most robust to potential problems in item content; this approach is based on selecting a set of items that fit the model well across any two languages (source, target) which then serve as a reference while the non-fitting items are reintroduced back into the model in order to test cross-cultural equivalence [66 68]. If an item displays DIF (but fits the IRT model in each culture), its scaling is considered unique to that group. For example, two cultures could respond differently to the same item, which indicates that the item does not assess the respective domain in the same manner, yet the item could be equally important within both cultures; the item would remain within each instrument but would be scored independently (i.e., dummy coded for an IRT analysis) since the item contributes to the overall score differently in each culture. The resultant instrument would contain both common (i.e., unsplit items, which serve as links across the respective versions) and unique items (which vary between the two assessed versions), thus maximally addressing the construct within each culture. Fit to the measurement model is again reassessed, and if the item still displays misfit to the model, then it is removed from the scale entirely and does not contribute to the person estimate. Previous evidence indicates that when all items fit a construct within a culture but some items display DIF in cross-cultural comparison, those items displaying DIF may affect the full scale minimally [67].

DIF can be assessed using one or more of several established approaches, of which two are briefly described here:

(1) One standard approach is based on IRT models. An IRT model is established for each subgroup and the assumption that the item parameters are equal across subgroups is

tested through an likelihood ratio or a chi-square test [56-58]. Significant differences in item parameters provide evidence that respondents from different groups (e.g., gender, cultures, age) endorse an item differently, controlling for overall severity in the measured construct – measured through the IRT score. Most IRT software packages now implement some form of DIF testing. For two-parameter models, a distinction is made between uniform DIF (difference in threshold parameters only) and non-uniform DIF (difference in slope parameters). The primary limitation of such DIF implementations is that predictors can generally only be categorical, and more complex relationships among predictors cannot be assessed, if hypothesized to be important. There are also stand-alone software programs for assessing DIF using these statistics.

(2) Another approach utilizes ordinal logistic regression and model comparison procedures. In this approach the sum score of all items is used to control for the overall severity in the measured construct. Uniform DIF is tested as a main effect of group membership on the item score (when controlling for the sum score of all purified items and the item in question) while non-uniform is tested as a group times sumscore interaction [55]. With this approach the "grouping" variable can also be continuous. One adaptation of this approach is DIFdetect [69], an ado program for Stata 8.0 (and which will also be adapted for use in SPSS and SAS). DIFdetect uses an ordinal logistic regression model with the response options from the selected item as the dependent variable, and the predictors are the sum score, categorical group identifiers and/or continuous identifiers (e.g., age), and the interaction term(s) between group/continuous identifiers and trait (i.e., summary) score. DIFdetect implements the possibility of using IRT score estimates as the scale score instead of the total sum score, which is desirable as a method for compensating for missing data that can occur in clinical and epidemiological studies. Also, an importance criterion (in terms of magnitude of change of the beta coefficient [70]) can be implemented. Multiple pair-wise test comparisons occur, and alpha protection is maintained through Bonferroni adjustment. For example, a test with 41 items would result in 210 different DIF hypotheses being assessed. DIFdetect has been updated and revised into a program that runs under R, with improved reporting of statistics and graphical depiction of results; readers should search the web for these continually updated programs.

In such a situation with a very large number of multiple comparisons, however, the Bonferroni correction will often be too conservative. An alternative approach to this problem resides in resampling related to multiple comparison procedures [71], which is available in SAS, Stata, and other software packages.

Sample Size Estimates

For the field test evaluation of a translated instrument, estimates for a sufficient sample size depend on three intersecting variables: degree of confidence interval desired, homogeneity of the study sample with respect to the measured construct, and how well the items fit the construct as determined by the measurement model. In our work to date, we have found that for items that are extremely well-fitting with respect to the model, 30 subjects can be sufficient, while for items that exhibit a more average model fit, 50 subjects might be the minimum. These sample sizes are based on good selection of subjects who exhibit qualities relevant to the construct being assessed. In terms of 95% confidence intervals, these rough sample size estimates are consistent [72] and will serve as a first quantitative view of the data, utilizing an iterative process towards commensurate culturally sensitive measurement scales.

For the more definitive analyses regarding cultural equivalency, the minimum number of subjects required to produce stable IRT parameter estimates will depend on the model and the centering of item to the population; that is, there is no general right or left skew in the data due to the population being either more capable compared to the trait being assessed, or less capable compared to the assessed trait. At least 250-500 subjects are typically regarded as necessary in a single sample in order to have stable estimates in general IRT models [38]. In contrast, the Rasch (1-parameter IRT) model does not contain a discrimination parameter and consequently can be estimated with smaller samples. Note that the centering of items to the population is always important in terms of assessing the adequacy of a particular sample size for a particular set of IRT analyses.

According to the Rasch literature, stable IRT parameter estimates can be obtained from a sample size of 150 subjects, which will estimate the item difficulty with an α of 0.01, to within $\pm$ 0.5 logits [73 74] where one logit is equal to an increase in the odds ratio of 2.716. This sample size is also sufficient to test for DIF where (a) an α of 0.01 and a difference of 0.05 SD within the residuals can be detected for any 2 groups with β of 0.20; or (b) with identical α and β, a difference of SDs within the residuals can be detected where this difference varies by 0.1 logits across each of 6 groups [66]. Bonferroni corrections are applied to both fit and DIF statistics due to the number of tests undertaken in the pair-wise testing process. Note, however, that Bonferroni can be too conservative when large numbers of comparisons are made, and resampling procedures are a recommended alternative.

For pair-wise comparison procedures for DIF among multiple cultures, choosing a single reference language is recommended, with English most likely often being that choice for a number of practical (but not scientific) reasons. For such multiple pair-wise comparisons, while 150 subjects might be sufficient for the comparison countries using standard IRT analyses (particularly, with Rasch models), an enhanced sample size for the reference language group is recommended. The reference sample should include approximately 300 subjects as a minimum (see [60] for discussion). When performing DIF analyses at the Educational Testing Service, Princeton NJ, at least 100 subjects are required for the smaller sample, and the smaller and larger group combined must sum to at least 500 subjects. The increase from 400 (100 minimum in smaller group + 300 in the larger group) to a total of 500 probably represents a negligible effect on power, but simulation studies would be needed to assess this.

Software

Some recommendations for software are made, based on experience in performing the analyses described in this manual. However, there are many other software packages available for any of the described analyses.

Winsteps software, for the 1-parameter model estimates [75], uses an unconditional maximum likelihood estimation procedure to estimate item calibrations and person measures. The rating scale model assumes a uniform underlying structure, but if the model fit is not adequate, that assumption can be relaxed via use of the partial credit model which makes no *a priori* assumption regarding the internal structure of the measurement scale [76 77]. The developers of Winsteps hold training courses in the US and Australia throughout the year, and possibly in other countries; see www.rasch.org. Another recommended program is RUMM2010, which uses a pairwise conditional maximum likelihood estimation procedure.

Parscale or Multilog software programs are often recommended for the 2-parameter model. Both these programs use Marginal Maximum Likelihood estimation. Choice of whether to use 1- vs 2- parameter estimates is based on evaluation of discrimination and overall model fit.

For assessment of DIF, many approaches can be used. Winsteps and RUMM2010 implement chi-square procedures for assessing DIF between two groups. There is an add-on (ado file: DIFdetect [69]) for Stata 8.0 software that provides excellent DIF analysis for multiple group comparisons based on the use of ordinal logistic regression, permitting the modeling of uniform and non-uniform DIF with either categorical or continuous exogenous variables. Under R, the program lordif replaces DIFdetect.

For factor analyses that assess assessing covariance relationships in data that are not continuous, several software packages that include the appropriate models are available: M-plus software [65], Mx, and Lisrel, among others.

Sources

Winsteps software (www.winsteps.com)

RUMM2010 (www.rummlab.com)

Parscale (www.assess.com/Software/parscale.htm)

Multilog (www.assess.com/Software/MULTILOG.htm)

DIFdetect (www.alz.washington.edu/DIFDETECT/welcome.html)

M-plus software (www.statmodel.com)

Mx (www.vcu.edu/mx/)

lordif (http://cran.r-project.org/web/packages/lordif/index.html)

R (www.**r**-project.org)

Acknowledgments

The authors extend their appreciation to the following individuals who have provided helpful comments and assisted with the preparation of this document: Dr. Yoly Gonzalez, University at Buffalo; Dr. Stuart Gansky, UCSF; Dr. José Velasco-Neri, University of Guadalajara; Ms Anup Kang, University at Buffalo; Dr. Mohammed Al-Harthy and Dr. Thomas List of Malmö University, and to the members of the International RDC/TMD Consortium for their encouragement and commitment to the standards embodied in this set of Guidelines.

Version Notes

Nov 16, 2008

- Roles for individuals involved in a translation team identified, and end point of the first phase of translated instrument development redefined to be the end of the expert committee which then permits posting of an instrument to the user community for their now participating in the next stage of adaptation and validation.

December 1, 2009

- Logs and Appendix 2 substantially revised in order to improve identification of the different stages of translation, better reflect the translation process, and include new Appendix 3 as a hand-out for panel review members.

May 11, 2013

- Entire text edited for errors.
- “Developer” replaced with term “Team Leader”, as more self-explanatory.
- Software tools updated.
- Procedures updated to reflect DC/TMD replacing the RDC/TMD.

References

1. Dworkin SF, LeResche L. Research Diagnostic Criteria for Temporomandibular Disorders: Review, Criteria, Examinations and Specifications, Critique. J Craniomandib Disord Facial Oral Pain 1992;6:301-355.

2. Kosminsky M, Lucena LBS, de Siqueira JTT et al. Adaptação cultural do questionário "Research Diagnostic Criteria for Temporomandibular Disorders: Axis II" para o português (Cultural Adapation of the "Research Diagnostic Criteria for Temporomandibular Disorders: Axis II). Jornal Brasileiro de Clinica Odontologica 2004;43.

3. Lobbezoo F, van Selms MKA, John MT et al. Use of the Research Diagnostic Criteria for Temporomandibular Disorders for multinational research. Translation efforts and reliability assessments in The Netherlands. J Orofacial Pain 2004;19:301-308.

4. John MT, Hirsch C, Reiber T et al. Translating the Research Diagnostic Criteria for Temporomandibular Disorders into German: Evaluation of content and process. J Orofacial Pain 2006;20:43-52.

5. Ware JE, Jr., Gandek B. Methods for testing data quality, scaling assumptions, and reliability: the IQOLA project approach. J Clin Epi 1998;51:945-952.

6. Bullinger M, Anderson R, Cella D et al. Developing and evaluating cross-cultural instruments from minimum requirements to optimal models. Qual Life Res 1993;2:451-459.

7. Bergner M, Bobbitt RA, Carter WB et al. The Sickness Impact Profile: development and final revision of a health status measure. Medical Care 1981;19:787-805.

8. Ware JE, Jr., Kosinski M, Gandek B et al. The factor structure of the SF-36 Health Survey in 10 countries: Results from the IQOLA project. J Clin Epi 1998;51:1159-1165.

9. Anonymous. EuroQol--a new facility for the measurement of health-related quality of life. The EuroQol Group. Health Policy 1990;16:199-208.

10. Leplege A, Ecosse E, WHOQOL Rasch Project Scientific Committee. Methodological issues in using the Rasch model in select cross-culturally equivalent items in order to develop a quality of life index: The analysis of four WHOQOL-100 data sets (Argentina, France, Hong Kong, United Kingdom). Journal of Applied Measurement 2000;1:372-392.

11. Merskey H, International Association for the Study of Pain--Subcommittee on Taxonomy. Classification of chronic pain: descriptions of chronic pain syndromes and definitions of pain terms. Pain 1986;Suppl. 3:S1-S225.

12. Bullinger M, Alonso J, Apolone G et al. Translating health status questionnaires and evaluating their quality: the IQOLA project approach. J Clin Epi 1998;51:913-923.

13. Coyne I. ITC Test Adaptation Guidelines (April 21, 2001 Version). 2001. International Test Commission.

14. Beaton D, Bombardier C, Guillemin F et al. Recommendations for the cross-cultural adaptation of health status measures. 2002. Toronto, Institute of Work and Health, and American Academy of Orthopaedic Surgeons.

15. Beaton DE, Bombardier C, Guillemin F et al. Guidelines for the process of cross-cultural adaptation of self-report measures. Spine 2000;25:3186-3191.

16. Ware JE, Jr., Keller SD, Gandek B et al. Evaluating translations of health status instruments. International Journal of Technology Assessment in Health Care 1995;11:525-551.

17. Geisinger KF. Cross-cultural normative assessment: translation and adaptation issues influencing the normative interpretation of assessment instruments. Psychological Assessment 1994;6:304-312.

18. Guillemin F, Bombardier C, Beaton D. Cross-cultural adaptation of health-related quality of life measures: literature review and proposed guidelines. J Clin Epi 1993;46:1417-1432.

19. Behling O, Law KS. Translating Questionnaires and Other Research Instruments. Thousand Oaks, CA: Sage, 2000.

20. Sechrest L, Fay TL, Zaidi SMH. Problems of translation in cross-cultural research. Journal of Cross-Cultural Psychology 1972;3:41-56.

21. Brislin RW. Back-translation for cross-cultural research. Journal of Cross-Cultural Psychology 1970;1:185-216.

22. Trust introduces new translation criteria Medical Outcomes Trust Bulletin 1997;5.

23. Massoubre C, Lang F, Jaeger B et al. Translating questionnaires and tests: techniques and problems. Canadian Journal of Psychiatry 2002;47:63-69.

24. Keller SD, Ware JE, Jr., Gandek B et al. Testing the equivalence of translations of widely used response choice labels: results from the IQOLA project. J Clin Epi 1998;51:933-944.

25. Bullinger M. German translation and psychometric testing of the SF-36 Health Survey: preliminary results from the IQOLA Project. International Quality of Life Assessment. Soc Sci Med 1995;41:1359-1366.

26. Perneger TV, Leplege A, Etter JF. Cross-cultural adaptation of a psychometric instrument: two methods compared. [Clinical Trial. Journal Article. Randomized

Controlled Trial] Journal of Clinical Epidemiology. 52(11):1037-46, 1999. J Clin Epi 2005;52:1037-1046.

27. Guyatt GH. The philosophy of health-related quality of life translation. Qual Life Res 2005;2:461-465.

28. Flaherty JA, Gaviria FM, Pathak D et al. Developing instruments for cross-cultural psychiatric research. J Nerv Ment Dis 1988;176:257-263.

29. McLaughlin G. SMOG grading: A new readability formula. Journal of Reading 1969;12:639-646.

30. Wynd CA, Schmidt B, Schaefer MA. Two quantitative approaches for estimating content validity. Western Journal of Nursing Research 2003;25:508-518.

31. Lynn MR. Determination and quantification of content validity. Nurs Res 1986;35:382-385.

32. Sireci SG. Using bilinguals to evvaluate the comparability of different language versions of a test. In: Hambleton RK, Merenda PF, Spielberger CD, eds. Adapting Educational and Psychological Tests for Cross-Cultural Assessment. Mahwah, NJ: Lawrence Erlbaum Associates,2005:117-138.

33. Brislin RW, Lonner W, Thorndike RM. Cross Cultural Research Methods. New York: John Wiley and Sons, 1973.

34. Hunt SM. Cross-cultural comparability of quality of life measures. Drug Information Journal 1993;27:395-400.

35. Morgan DL. Focus Groups as Qualitative Research. Thousand Oaks: Sage Publications, Inc., 1996.

36. Morgan DL, Krueger RA. The Focus Group Kit: Volumes 1-6. Thousand Oaks: Sage Publications, Inc, 1997.

37. Hastie T, Tibshirani R, Friedman J. The elements of statistical learning : data mining, inference, and prediction. New York: Springer, 2001.

38. Embretson SE, Reise SP. Item Response Theory for Psychologists. Mahwah,NJ: Lawrence Erlbaum, 2000.

39. Borsboom D, Mellenbergh GJ, van Heerden J. The concept of validity. Psychol Rev 2004;111:1061-1071.

40. -----The theoretical status of latent variables. Psychol Rev 2003;110:203-219.

41. Hui CH, Triandis HC. Measurement in cross-cultrual psychology: a review and comparison of strategies. Journal of Cross-Cultural Psychology 1985;16:131-152.

42. Andrich D. A Rating Formulation for Ordered Response Categories. Psychometrika 1978;43:561-573.

43. Masters GN. A Rasch model for partial credit scoring. Psychometrika 1982;47:149-173.

44. Muraki E. A Generalized Partial Credit Model. In: van der Linden WJ, Hambleton RK, eds. Handbook of Modern Item Response Theory. Berlin: Springer,1997:153-164.

45. Samejima F. Estimation of latent ability using a response pattern of graded scores. Psychometrika Monograph Supplement 1969;17.

46. Andrich D, Lyne A, Sheridan B et al. RUMM 2010 (version 3.3). Perth: RUMM Laboratory, 2000.

47. Smith RM. Fit analysis in latent trait measurment scales. Journal of Applied Measurement 2000;2:199-218.

48. Glas CAW, Verhelst ND. Testing the Rasch Model. In: Fischer GH, Molenaar IW, eds. Rasch Models - Foundations, Recent Developments, and Applications. Berlin: Springer,1995.

49. Orlando M, Thissen D. Likelihood-Based Item-Fit Indices for Dichotomous Item Response Theory Models. App Psych Measurement 2000;24:50-64.

50. Stone CA. IRTFIT_RESAMPLE: A Computer Program for Assessing Goodness of Fit of Item Response Theory Models Based on Posterior Expectations. App Psych Measurement 2004;28:143-144.

51. Chen W-H, Thissen D. Local Dependence Indexes for Item Pairs Using Item Response Theory. Educational and Behavioral Statistics 1997;22:265-289.

52. Wright BD. Local Dependency, Correlations and Principal Components. Rasch Measurement Transactions 1996;10:509-511.

53. Drasgow F, Levine MV, Williams EA. Appropriateness measurement with polychotomous item response models and standardized indices. Br J Math Stat Psychol 1985;38:67-86.

54. Molenaar IW, Hoijink H. The Many Null Distributions of Person Fit Indices. Psychometrika 1990;55:75-106.

55. Zumbo BD. 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. Ottawa,ON: Directorate of Human Resources Research and Evaluation, Department of National Defense, 1999.

56. Reise SP, Widaman KF, Pugh RH. Confirmatory factor analysis and item response theory: two approaches for exploring measurement invariance. Psychol Bull 1993;114:552-566.

57. Raczek AE, Ware JE, Bjorner JB et al. Comparison of Rasch and summated rating scales constructed from SF-36 physical functioning items in seven countries: results from the IQOLA project. J Clin Epi 1998;51:1203-1214.

58. Thissen D, Steinberg L, Wainer H. Detection of Differential Item Functioning Using the Parameters of Item Response Models. In: Holland PW, Wainer H, eds. Differential Item Functioning. Hillsdale NJ: Lawrence Erlbaum Associates,1993.

59. Camilli G, Shepard LA. Methods for Identifying Biased Test Items. 1994; London: Sage publications. London: Sage Publications, 1994.

60. Zieky M. Practical Questions in the Use of DIF Statistics in Test Development. In: Holland PW, Wainer H, eds. Differential Item Functioning. Hillsdale, NJ: Lawrence Erlbaum Associates,1993:337-347.

61. Custers JW, Hoijtink H, van der NJ et al. Cultural differences in functional status measurement: analyses of person fit according to the Rasch model. Qual Life Res 2000;9:571-578.

62. Prieto L, Alonso J, Lamarca R. Classical test theory versus Rasch analysis for quality of life questionnaire reduction. Health and Quality of Life Outcomes 2003;1:27-40.

63. Tanzer NK. Cross-cultural bias in Likert-type inventories: perfect matching factor structures and still biased? European Journal of Psychological Assessment 1995;11:194-201.

64. Prieto L, Alonso J, Lamarca R et al. Rasch measurement for reducing the items of the Nottingham Health Profile. Journal of Outcomes Measurement 1998;2:285-301.

65. Muthen BO, Muthen L. Mplus User's Guide. 3[3]. 2004. Los Angeles, Muthén & Muthén.

66. Tennant A, Penta M, Tesio L et al. Assessing and adjusting for cross-cultural validity of impairment and activity limitation scales through differential item functioning within the framework of the Rasch model. Medical Care 2004;42:I-37-I-48.

67. Bjorner JB, Kreiner S, Ware JE, Jr. et al. Differential item functioning in the Danish translation of the SF-36. J Clin Epi 1998;51:1189-1202.

68. Lange R, Irwin HJ, Houran J. Top-down purification of Tobacyk's revised paranormal belief scale. Person individ Diff 2000;29:131-156.

69. Crane PK, Jolley L, van Belle G. DIFdetect (for use with Stata 8.0). 2004.

70. Maldonado G, Greenland S. Simulation study of confounder-selection strategies. American Journal of Epidemiology 1993;138:923-936.

71. Westfall PH, Young SS. Resampling-Based Multiple Testing: Examples and Methods for p-Value Adjustment. New York: John Wiley and Sons, 1992.

72. Linacre JM. Sample Size and Item Calibration Stability. Rasch Measurement Transactions 1994;7:328.

73. -----Sample Size and Item Calibration Stability. Rasch Measurement Transactions 1994;7:328.

74. Tabachnick BG, Fidell LS. Using Multivariate Statistics. New York: HarperCollins, 1989.

75. Linacre JM. A User's Guide to Winsteps Rasch-model Computer Programs. Chicago: www.winsteps.com, 2004.

76. Wright BD, Masters GN. Rating Scale Analysis. Chicago: Mesa Press, 1982.

77. Andrich D. Rasch Models for Measurement. Newbury Park, CA: Sage Publications, 1988.

Appendix 1. Expert Panel Review of a Translation

Panel Member Instructions

Version: December 1, 2009

The instructions in this Appendix are derived from material developed in Stage 7 of the Guidelines document. These instructions are intended to be used as a set of stand-alone instructions to the members of the Expert Panel.

1. Introduction. While the goal in translation is to create a target document that is an exact copy of the source document, specific challenges determine how accurately that goal is met for the construction of instruments used in clinical practice and research. The theory of translation contains an extensive range of issues associated with text translation, and there are many ways to assess translation sufficiency. The guidelines adopted by the Consortium describe one accepted set of procedures for the construction of measurement instruments. See the source document for more detail and references.

2. Panel Structure. Panel members are chosen by the Team Leader of the target translation. With regards to source and target languages, the Panel should be comprised of individuals who are bilingual as well as those who are monolingual (i.e., speak target language but not source language). Bilingual members will review the source and target language versions of the instrument while the monolingual language members will review only the target language instrument; obviously, the review goals differ depending on whether the reviewer is bilingual vs monolingual. Ideally, the members meet jointly for group discussion with the translators, but the members can also evaluate the translation independently of one another. In either case, they convey their findings to the translation Team Leader who coordinates recommendations and next actions for any revision.

3. Focus of Review. The Team Leader will inform the committee members regarding the scope of the review: selected parts of the translation or the entire translation. The level of understanding language for the instrument text should generally be that of a 12 year old. Panel members should pay particular attention to questionable content and questionable phrasing; that is, literally correct translation from the source may not be appropriate to the intended construct in the target language. The Panel members are provided with the source language and target language content as well as the Logs used during development of the translation so that discussion by the translation team regarding any particularly difficult items for translation can be referenced in the review by the Panel members.

4. Sufficiency of Translation. For the purposes of determining the sufficiency of translation of a clinical or research instrument, each translation unit (instructions, item content, and response options) should be evaluated with regard to each of the following 4 types of equivalences. Not all types of equivalence need be met for a given item to be considered acceptable. Some aspects of an equivalence can be more readily reviewed either by bilingual or by monolingual reviewers.

a. **Semantic equivalence**: Does the word or expression have the same literal meaning between source and target language? Are there multiple meanings of the word in the target language that might create different response sets by respondents? Are there grammatical problems in the translation?

b. **Idiomatic equivalence**: Idiomatic or colloquial expressions in the source language are often difficult to translate and will usually not be literally translated to the target language; an idiom in the target language may even be needed. Or a non-idiomatic item in the source language may sometimes be best translated into an idiomatic or colloquial expression in the target language in order to best capture the intended meaning.

c. **Experiential equivalence**: An item attempts to capture the respondent's everyday life experience. Such experiences are often culturally linked; the words used in a target translation need to capture that experience.

d. **Conceptual equivalence**: Words between two cultures might map semantically but have different connotations with regards to the intent of the instrument. Does the target word or expression describe the concept in the source language and adequately map to the intended purpose of the item?

5. Item Ratings by Panel Members. As shown on the review form (Log G1), two types of ratings and a comment should be provided for each translation unit:

a. *Sufficiency*. The sufficiency of the translation is rated as adequate, questionable, or inadequate, based on the 4 types of equivalences.

b. *Content Validity Rating*. Content validity refers to whether the item belongs, on the basis of its surface meaning, in the domain as established by the instrument. A rating should be applied to each item: 1 = not relevant, 2 = somewhat relevant, 3 = quite relevant, 4 = very relevant. Item content in translations generally follows that of the source instrument, as determined by the construct, but content relevant in the source language for that culture may not necessarily be equally relevant for the target language or culture.

c. *Comments*. If the sufficiency is questionable or inadequate or if the content validity rating is 1 or 2, insert brief description of the problem for the given item as related to the 4 types of equivalences or cultural relevancy, as appropriate. If an item is questionable in some way but the reviewer is not sure that the item should be considered for revision or being dropped at this stage, the reviewer can also indicate that the item should receive special review at the next stage of translation validation – the pre-test, where qualitative interviews are conducted with the intended respondents. If the translation is adequate and the content validity rating is a 3 or 4, then write "None" in this field.

Appendix 2. Phase II Translation Review

Reviewer Instructions

Version: Jan 2, 2011

This Appendix provides additional information regarding translation review as part of Phase II assessment of reliability and validity of a translation. When the translated instrument being compared for cultural equivalency was not created by the Phase II evaluation research team, qualitative review is particularly critical and these guidelines are also applicable.

1. Introduction. After a translated instrument has been developed through the end of Phase I, which included panel review regarding the correctness of the translation, Phase II consists of administering the instrument to the target population in order to test the reliability and validity of the translation. The Guidelines text describes the necessary methods for quantitative assessment of item and test characteristics as part of Phase II in order to better interpret any discrepancies in item performance as revealed by quantitative analysis.

2. Review Member Structure. At this stage of review of a completed instrument, the goal is to generate impressions regarding the nature of the translated items. In general, two Reviewers are probably sufficient, and both Reviewers should be bilingual. One Reviewer should be a content expert, and the other Reviewer should represent the target population. Reviewers will meet initially in order to review all items, and then they may meet an additional time in order to focus on selected items as determined by the quantitative analysis. Ideally, the Reviewers meet together in order to construct a consensus evaluation; however, if conjoint meeting is not possible, they may perform their reviewers separately, and then the main reviewer will consolidate the two separate reviews into one (with additional consultation of the other Reviewer, as needed).

3. Focus of Review. The intent of this Review is to assess the translation that has been performed in order to identify any residual problems with the translation and usage of each item. The Reviewers will focus on the four types of equivalency as well as sufficiency.

   The four types of equivalency – semantic, idiomatic, experiential, and conceptual – are considered during the translation and review stages of Phase I, with the goal of producing an item that assesses the same phenomenon as intended by the source instrument. However, cultures and languages differ, and a variety of considerations enter into the final determination of a translation. Whether that final translation is consistent, in a measurement sense, with the source instrument must be empirically evaluated. In addition to the quantitative assessment of item consistency, qualitative review of a translation has been recently included as a necessary step during this Phase. Qualitative review consists of formal comparison of the translation to the source in order to detect possible sources of discrepancy. That is, during a translation, the translation team endeavors to recreate, under best practices,

a target instrument comprised of items that are as relevant as possible to the latent construct underlying the source instrument. And so, again, the four types of equivalency must be considered for whether a given type of equivalency is sufficient for item agreement with the source language instrument so that the qualitative review can be considered conjointly with the quantitative review.

An item that has been translated with full semantic equivalence may reflect a perfectly correct translation or it may not, depending on the context of the item within a given culture. Not all types of equivalence need be met for a given item to be considered acceptable. Consequently, qualitative review is intended to provide yet another perspective regarding the characteristics of the items that comprise a translated instrument.

4. Equivalency of Translation. Each translation unit (instructions, item content, and response options) should be evaluated with regard to each of the following 4 types of equivalences.

    a. **Semantic equivalence**: Does the word or expression have the same literal meaning between source and target language? Are there multiple meanings of the word in the target language that might create different response sets by respondents? Are there grammatical problems in the translation?

    b. **Idiomatic equivalence**: Idiomatic or colloquial expressions in the source language are often difficult to translate and will usually not be literally translated to the target language; an idiom in the target language may even be needed. Or a non-idiomatic item in the source language may sometimes be best translated into an idiomatic or colloquial expression in the target language in order to best capture the intended meaning.

    c. **Experiential equivalence**: An item attempts to capture the respondent's everyday life experience. Such experiences are often culturally linked; the words used in a target translation need to capture that experience.

    d. **Conceptual equivalence**: Words between two cultures might map semantically but have different connotations with regards to the intent of the instrument. Does the target word or expression describe the concept in the source language and adequately map to the intended purpose of the item?

5. Sufficiency of Translation. The sufficiency of the translation is rated as adequate, questionable, or inadequate, based on the 4 types of equivalences.

6. Method. As discussed in the Guidelines, some aspects of an equivalence can be more readily reviewed either by bilingual or by monolingual reviewers; for the present purpose, however, the more critical distinction between two reviewers is that of content expert vs an individual with more general knowledge in that the latter more closely represents the primary characteristic of the target population.

    When the translator or developer of a translated instrument is also involved with conducting Phase II, it would seem that this type of review would not be necessary; contrast this situation with that of a researcher evaluating the cultural characteristics

of an instrument that he did not create: he would not be informed as the process underlying a given translation, and thus would not know whether the translation is to be trusted or not. In either situation, however, qualitative review is necessary. In the former situation a separate review of the instrument allows a more objective perspective on what was previously created via the translation and review process.

The following steps are involved with qualitative review:

1. Identify individuals to act as the Reviewers.
2. Transfer all source and target language items from Log H to Log QTR which needs to be created for this purpose by creating an additional column for review comments for each item.
3. Arrange for conjoint review, if possible, of the items.
4. Document the consensus outcome for each item, indicating “type of equivalence” that applies as well as the “sufficiency” of the translation for that equivalence. Finally, describe any concerns regarding the item translation.

.

Appendix 3: Handbook of Translation and Validation Procedures

Prepared by Richard Ohrbach, Anup Kang & Mohammed Al-Harthy
for the Committee for Translations and Protocols, International RDC/TMD Consortium Network
Revised December 1, 2009

The following guidelines, organized by Phase and Stage, summarize the step-by-step procedures to be followed in instrument translation and establishment of cultural equivalency. This material is identical to that presented in the body of the Guidelines; the intent of this section is to provide the user with a readily accessible handbook for implementing each of the procedures and to perhaps serve as a checklist during the translation and validation steps. The rationale for each of the procedures listed below is found in the respective section of the Guidelines. Note that while the Guidelines for Equivalency document explicitly describes the recommended procedures as *guidelines*, the listed steps are the essential operations to be carried out for developing an instrument. The exact method, and the adjunctive steps to be taken at any given stage, however, must be determined for a given instrument and hence are under the control of the Team Leader and, as stated in the Guidelines for Equivalency, depend on the nature of the instrument, the resources available, and the intended usage of the translated instrument.

Procedures applicable to all Logs:

- **Item content for source and target languages will be in those languages.**
- **Use English for all other documentation unless indicated otherwise.**

Stage	Task	General Procedures	Log	Translation Log Procedure
Phase 1: Translation and Cultural Adaptation				
—	Setup by Team Leader	• Review Ohrbach et al, Guidelines for Establishing Cultural Equivalency of Instruments. Principles are important and must inform decisions in the translation process. • Define the intended population (country, group, culture) of the translated instrument; these criteria will be used as reference for translation and for pretesting & field testing (Phase 2). • These criteria should address: o gender o age range o psychosocial & educational status, ethnicity, language fluency, cultural identification	A	***Team Leader:*** Complete Tables 1, 2, and 3 on **Log A**.

Stage	Task	General Procedures	Log	Translation Log Procedure
1	Forward Translation (F-T)	• Select Forward (F) Translator per these criteria: o Native language is target language, and second language is source language. o Using 2 F Translators: F Translator 1 is informed (aware of health concept) & F Translator 2 is uninformed (naïve to concept), with respect to instrument content. o Using 1 F Translator: preference is given to being uninformed (naïve to concept). o If Team Leader is a F Translator, and only 1 F Translator is used, a Coordinator will also be needed for the project. • Each F Translator produces an independent translation (F-T1/F-T2) of the source language into the target language.	B	***Team Leader:*** • Complete Table 1 on **Log B**. • Phase is "1" for the first time **Log B** is used, and for each iteration used for translation corrections, the Phase number is increased. • In Tables 2-5 on **Log B**, enter each translation unit (instructions, each item, each recurring response option, scoring rules, respectively), from source instrument into respective sets of cells. ***Forward Translator:*** • Target translation entered, unit by unit, into cells labeled "Forward Translated Version (F-T1/F-T2)."
2	Synthesis	• If 1 Forward Translator: Log C not needed. • If 2 Forward Translators, and Team Leader is not a translator: o Team Leader resolves any differences in translation style in order to facilitate the synthesis from the two translation versions. o Team Leader can consult with translators as needed. • If 2 F Translators, and Team Leader *is* one of the Forward Translators: o The Team Leader and the other F Translator must sit together in order to do the synthesis.	C	***Team Leader:*** • Complete Table 1 on **Log C**. Indicate correct Phase number. • Combine translations F-T1 & F-T2 into one common translation F-T1&2. • Enter combined translation into cells labeled "Synthesis Translation (F-T1&2)." • Submit notes on discrepancies and their resolution in Table 6 "Synthesis Process Report of Discrepancies"

Stage	Task	General Procedures	Log	Translation Log Procedure
3	Back Translation (B-T)	• Select Back (B) Translator per these Criteria: o Native language is source language, and second language is target language. o Totally blinded from original source. o Choice of 1 or 2 B Translators. o At least one B Translator should not be medically oriented. • Each B Translator produces an independent B-T (B-T1, B-T2) from F-T or F-T1&2 into source language.	D	***Team Leader:*** • Complete Table 1 on **Log D**. Indicate correct Phase number. • Copy item # and target translation from Translation **Log B** or **Log C** to "Synthesis Translation (F-T or F-T1&2)" column on **Log D**. ***Backwards Translator:*** • Create B-T and enter into "B-T" column on **Log D**. • Rate overall quality of the Translation and enter in Table 1 on **Log D** ***Team Leader:*** • Copy original source language items from Translation **Log B** to "Source" column on **Log D**.

Stage	Task	General Procedures	Log	Translation Log Procedure
4	Indepen-dent Review	• Select Reviewer per these Criteria: o Should be an expert in the item content. o Should be an expert regarding the source instrument. o Should be fluent with the source language as a mother tongue. o Does not need to have skill in target language. o Not involved in F-T or B-T process. o Makes recommendations to the Team Leader (and hence, to the Forward Translator(s) regarding areas in the translation process that are questionable.	E	***Team Leader:*** • Complete first line of Table 1 on **Log E**, and highlight phase #. ***Reviewer:*** • Complete Name, Date, and Profile in Table 1 on **Log D**. • Compare B-T against source items on **Log D** from each B-T. • For problematic items, enter Item # and Source Item (copied from **Log D**), and describe nature of inconsistency in "Reviewer: Describe Problem" column. • Enter Total Discrepancy Count (the total # of discrepancies between B-T and source document) into Table 1 on **Log E**. Consider one error in a multi-word phrase as one discrepancy, and a single item may have more than 1 discrepancy. Do not count errors in preposition as discrepancy unless the relational properties are critical, and do not include spelling errors in the B-T. ***Team Leader:*** • Review **Log E** and consult with F-T, B-T, or both in relation to the identified problems, and cultural equivalence in the translation should be incorporated. • Use "Team Leader Response" column, and preface F-T responses with "F-T:" and B-T responses with "B-T:". • Additional forward translation may be entered on **Log E** as a discussion item for response by Team Leader or Reviewer. • If Back Translator revises the B-T to current version of the F-T, the revised back translation should be entered onto Log. • If item requires forward translation revision, then it has to be retranslated (Stage V: Iterative Development) and new **Logs B, C, and D** are created; indicate on **Log E** that new translation will be performed. ***Reviewer:*** • Enter recommendation in response to Team Leader Response, indicating whether current iteration of translated item is acceptable or whether revised translation should be pursued. Log is returned to Team Leader. ***Team Leader:*** • Provide short summary of new step to be taken, using "Reviewer Recommendation and Team Leader Action" column; Team Leader response should be highlighted in yellow.

Stage	Task	General Procedures	Log	Translation Log Procedure
5	Revision and Iterative Develop-ment	• Stage V entails return to Stages I-IV for only the discrepant items. • Team Leader coordinates and repeats forward translation (and synthesis), back translation, and review until Reviewer approves the forward translation.	B C D E	**Logs B, C, D, and E** are incremented using Iteration number in Table 1. File names are incremented with numbers (e.g., original **Log B** is renamed to "**Log B1**" and the next **Log B** for F-T is named "**Log B2**"). ***Team Leader:*** For any items that require revision in the translation, copy source from items to be retranslated onto new **Log B**. ***Forward Translator, Team Leader, Back-Translator*:** • Stages I-III are repeated. ***Team Leader:*** • Repeat Stage IV. • Continue with iterations, incrementing Phase # each cycle until Log E indicates that there are no discrepancies.
6	Consoli-dation	• For each Source item, the F-T and B-T from each iteration are compiled onto the Log. • Comments are added as necessary for each item. • Goal is to produce in a single document a complete record of the translation process so that Panel members can readily review the translation.	F	***Team Leader:*** • For each Source item on **Log B**, enter F-T (or F-T1&2) and B-T, including the Phase # associated with the F-T and B-T, on **Log F**. This is done for each area: instruction, item, response options, and scoring instructions. • For units that required revised translations, the source item need only be entered once. • Increment iteration number sequentially for any revised forward translation content on **Log B** and **Log C** and back translation content on **Log D**; each new phase in the iteration process is added on rows beneath the Source item. • For each item, process comments from all **Log E**s can be added on an inserted line immediately following the translation content from the respective Phase.

Stage	Task	General Procedures	Log	Translation Log Procedure
7.1	Expert Panel Review	Expert Committee has no required membership but should always be comprised of at least 3 individuals not involved in any of the prior steps in instrument development and, depending on instrument content, is generally comprised of individuals with the following types of expertise: • clinician • psychologist • methodologist • language expert • patient • other specialists as needed • .A panel coordinator may be needed at this stage in case the Team Leader is the only Forward Translator. Logs. • It is expected that, in general, panel members will not meet in person but rather work independently and provide their review to the Team Leader. • If they are able to meet in person, it is preferable that they still complete the Log GR prior to the meeting. Else, Log GR can be completed at the Panel meeting. Recommendations are made with respect to 4 areas: • *Semantic Equivalence* for multiple meanings to a given item • *Idiomatic Equivalence* for colloquialisms or idioms • *Experiential Equivalence* for a given task that may not be experienced in the target language even if it is translatable. • *Conceptual Equivalence* for words with different conceptual meanings between cultures.	A GR F	***Team Leader:*** • Create the Expert Panel. • Enter names and credentials in Table 4 on **Log A**. • Appoint Coordinator if Team Leader was the only Forward Translator. • Complete Table 1 on **Log GR**. • Distribute **Log F** and **Log GR** to each Panel member. ***Panel Member:*** • Review **Log F**. • Enter Sufficiency rating (Adequate, Questionable, and Inadequate) of the Target translation. • Provide Content Validity Rating for item relevance as described on **Log GR**. • In Comments field, enter comments as follows: • If Sufficiency is “adequate” and Content Validity Rating is 3 or 4, enter “None”. • If the item exhibits mild and perhaps hard-to-define problems, enter “Needs special review at next stage”. • Else, describe nature of problem for either sufficiency of questionable or inadequate or content validity rating of 1 or 2.

Stage	Task	General Procedures	Log	Translation Log Procedure
7.2	Expert Panel Review - Summary	• The goal of the Panel is to obtain a wide range of perspectives regarding the translation. The goal of the summary stage is to integrate the description of the problems and recommendations. • Panel Members and Reviewer are advisory to the Team Leader who makes all final decisions.	GS	***Team Leader:*** • For each item noted by any Panel Member to have a discrepancy, enter in Table 2 Source and Target Translation, note the Content Validity Index for the item, and describe the problem. • Using the Item # as the key between Table 2 and Table 3, summarize in Table 3 the recommendations from the Panel member(s). • Table 3: Enter Team Leader Response and explanation for the indicated response. • Send **Log GS** to Reviewer. ***Reviewer:*** • Identified problems may be same or different from those that emerged during the translation process. • Reviewer task at this stage is to consider recommendations made at the levels of relevance and cultural equivalency. • In Table 3, provide response to Team Leader's plans. ***Team Leader:*** • Review recommendations from Reviewer. • Make final decisions for each problematic item. • With assistance from Forward and Back Translators, as needed, modify instrument content to reflect recommendation from Expert Panel.
8	Pre-Final instrument draft	• Team Leader synthesizes content from Log F with recommendations from Committee, as documented on Log GS, and enters that on Log H as a final record linking each source item to its determined translation at this phase in the translation process. • Create final draft form of instrument for initial administration, and include it with Log H as a separate document.	A H I	***Team Leader:*** • Complete Tables 4 and 5 on **Log A**. • From **Log F** and **Log GS**, consolidate and enter final instrument translation on **Log H** for each of instructions, items, item responses, and scoring. • Attach to the end of this document the pre-final form of the completed instrument which can be administered. • Complete Instrument Information and Table 1 on **Log I**; leave Table 2 blank. • Submit **Logs A – H** to the Consortium for review.

Stage	Task	General Procedures	Log	Translation Log Procedure
9	Consortium Review	Team Leader compiles logs and sends to Chair of Committee on Translations who will review all documentation and determine whether the documentation for the translation process is acceptable or whether revisions are required.	I	***Team Leader:*** • Enter name of instrument, source language and target language, source culture and target culture, source country and target country. • All Logs (A-H) are compiled into single pdf file (in order, alphabetically, by Log letter designation) and submitted, along with Log I in Word format, to Consortium Review Committee.

Stage	Task	General Procedures	Log	Translation Log Procedure
10	Posting Translation on Consortium web site.	After Consortium review is completed and the translation is approved, the instrument and Logs are posted on the Consortium web site. • Instrument for administration may be posted in public access area of web site or may only be accessible to Consortium members, depending on state of development of the instrument (i.e., validity). • Generally, instruments developed only to the stage of cultural equivalency but without pretested are not released to the public but are available through the Consortium web site to all registered members of the Consortium. • Archived Logs are posted on the Consortium web site, for permanent access only by Consortium members.	—	Consortium Translation Committee will post the pre-final completed instrument and the development logs compiled as a single document.

Stage	Task	General Procedures	Log	Translation Log Procedure
Phase 2: Translation Validation and Documentation				
11.1	Initiate Pretesting	Team Leader (who may be same or different from Team Leader for Phase 1) designs validation studies which includes determination of subject groups. **Language Methods** <u>Bilingual Method</u> ● Primary objective is to understand items with discrepant meaning between source & target versions. ● Both target & source language versions are used. <u>Monolingual Method</u> ● Primary objective is to understand items with discrepant meaning between subjects' interpretation and developer intentions. ● Subjects interviewed regarding intended meaning of items vs. chosen responses and possible translation alternatives. **General Procedures** ● Goal is qualitative review. ● Interview subjects in order to assess their understanding of item. ● Highlight differences between respondents and test developers, with respect to the understanding of the intended meaning of the items. ● Obtain content validity ratings, as needed, from subjects when items are problematic. ● Desired Outcome: Revise wording of problematic items.	J	***<u>Team Leader</u>:*** Complete Tables 1 and 2 on **Log J**.

Stage	Task	General Procedures	Log	Translation Log Procedure
11.2	Pre-Testing	**Two Approaches:** • Focus Groups o Assistant moderator is very helpful. o Tape record discussion for later/review. o 6-8 subjects meet for a period of 2 hrs (max). o Disagreement among group members is signal for further evaluation of the item. o When everyone in group comes to agreement then review of that item completed. o If group does not finish task within the 2 hrs, a new focus group is needed for additional review. • Clinical Sample Testing o Use approximately 10 subjects. o Sample can contain monolingual, bilingual subjects, or both. Review questions for subjects: • Describe in own terms their understanding of the concept.	K	***Team Leader:*** • Enter sample description, inclusion criteria and demographics on **Log K**. • Manage focus group. ***Team Leader, Translator:*** • As a result of pre-testing, some items may be removed from the instrument due to irrelevance in that culture, or • Items may be altered in such a way that the fundamental item intent is met but in a manner very different from how it was operationalized in the source version within the context of that particular environment. ***Focus Group*** • Facilitators • Describe each problematic item, including subject's CVR o assessment and probe interview notes as appropriate. • Summarize recommendation for each iteM listed in (2). • Enter item #, source item, pre-final translation, and revised translation based on protesting. ***Clinical Sample Testing*** • Include language status (monolingual, bilingual) in description. • Describe each problematic item & number of subjects who reported that item; include probe interview notes as appropriate. • Summarize recommendation for item identified in (2). • Enter item #, source item, pre-final translation, and revised translation based on pre-testing.
11.3	Instrument Review	Team Leader can facilitate multiple levels of review of the documentation collected at this stage: • national, by the Team Leader and his/her review panel; • bi-national, by members of the Translations Committee (or other international colleagues) as well as the site director; • cross-national, by having other (i.e., more than one) site investigators review all translations via the back-translations and support documents.	J	***Team Leader:*** • Begin Table 4 on **Log J**.

Stage	Task	General Procedures	Log	Translation Log Procedure
12	Field Test	**Subject Criteria** • Symptoms representative of clinical population. • Understand verbal and written source language. • Clear members of cultural setting are selected, and non-assimilated immigrants are not included (NB:cultural status needs to be qualified by the Team Leader). **General Procedures** • 30-40 subjects from target setting • Goal is quantitative review • Complete questionnaires suitable for internal reliability assessment **Analysis Stage** • Evaluate basic item characteristics in instrument. • Assess performance of target instrument compared to source instrument. • Use bilingual field-test/psychometric statistics sample to ensure test validity **Qualitative Review** • Two reviewers evaluate the 4 types of equivalence of each item in the translation. • Interpretation of DIF analyses in the Analysis Stage use this qualitative information to evaluate sources of disagreement in item performance across settings.	L QTR	***Team Leader:*** Enter sample description, study description, outcomes (Tables 1, 2, 3, on **Log L**). Reviewers complete the Log.

Stage	Task	General Procedures	Log	Translation Log Procedure
13	Instrument Revision	• Review outcomes from pretesting and field testing and determine if any aspect of instrument content requires revision. • Convene Expert Panel (Log J) as needed. • Team Leader creates revised content of instrument on Log M and creates <u>final</u> form for administration. • Final-revised instrument is submitted, along with Log M, at this time to the Consortium; this version replaces the previously posted version and this version can be made public, per the Team Leader's request, so that other investigators outside the Consortium can also assess validity.	M J	***<u>Team Leader:</u>*** • Compile content into **Log M** from **Log H** and add modifications from **Logs K and L**. • If there is no change in the translation for a given translation block, indicate "No change" in the cell under "Revised Translation" for each Table 2-5. • Complete Table 6; Explanation and Impact information is for legacy documentation for any data collected using prior version. • Complete Pretest and Field Test sections of Table 3 and complete Table 4 on **Log J**. • Include with **Log M** the final-revised form of completed instrument for administration, and submit **Logs J-M**, with the final-revised instrument, to the Consortium for updating the website.
14	Formal Assess-ment	<u>Subject Criteria</u> • Subjects should be representative of clinical population <u>Formal Assessment</u> • Administer instrument in larger sample. <u>Statistical Analysis</u> • Cronbach alpha and factor analysis (as needed). • IRT for item characteristics and differences.	N J	***<u>Team Leader</u>*:** • Enter sample description, study description, outcomes (Tables 1, 2 and 3, on **Log N**). • Complete Table 3 on **Log J**.

Stage	Task	General Procedures	Log	Translation Log Procedure
15	Standardize Scores across cultures	• Pain-wise evaluation of instruments. • Requires minimum of 4 items in common between versions. • Decide upon raw scores or IRT-derived scores: o If items fit Rasch model, then raw scores are sufficient. o If items do not fit Rasch model, can use raw score if slope parameters are relatively close. o If items are assessed with 2PL IRT model, then IRT scores are necessary. o If DIF, can compare IRT scores if (1) have enough common items, (2) country-specific parameters are established for items w/DIF, and (3) equivalence tables are constructed based on predicted values. • Consider re-sampling methods for estimates.	J	***Team Leader*:** • Complete Table 5 on **Log J**. • Team Leader should write a narrative summary of these procedures which will serve as the Methods and Results sections of a manuscript. This may be attached to the end of **Log J**.
16	Validation Research	• Conventional validity statistics. o Convergent & discriminant evaluations. o Compare pattern of correlation in one language vs. other. • IRT models also provide validity statistics.	O J	***Team Leader*:** • Enter sample description, study description, and outcomes (Tables 1, 2, and 3, on **Log O**). • Complete Table 6 on **Log J**.
17	Instrument Docu-mentation	Multi-national user manual • Instrument characteristics, reliability & validity data, administrator, and scoring are addressed. • Include comparative statistics between source and target version. • Submit outcomes of Formal Assessment (Log N), Score Standardization (narrative summary), Validation Research (Log O), and User Manual to Consortium for Archiving.	J	***Team Leader*:** Complete Table 7 on **Log J**. Compile **Logs J-O** into pdf, complete Tables 1 and 2 on **Log P**, and send both documents (**Logs J-O** pdf, **Log P** doc) to Consortium Translations Committee.

Appendix 4: Translation Logs

A separate ZIP file contains the accompanying Translation Logs which collectively comprise the "Summary Report on the Cross-Cultural Adaptation of an International RDC/TMD Consortium Instrument" (Summary Report). The Logs are intended to be generally applicable to the essential steps in the process of achieving a sound and defensible translation. The Logs will provide the information needed to review the adequacy of a translation and to then serve as an archival document attesting to that process for that particular instrument translation and serving as a reference for any subsequent modifications to the translated instrument as empirical evidence accumulates suggestive of such changes. A guiding principle in designing these Logs, as adaptations of the procedures first suggested by Dorcas Beaton, was that the translation process should be sufficiently self-documenting so that completion of the Logs requires very little additional effort beyond the actual translation process itself.

The level of documentation suggested in the Translation Logs is based on the complexities and difficulties associated with instrument translation, given its inherently iterative structure and consequent understandable reluctance to adequately document that complex process. Failure to provide that information, however, leads to greater difficulties later when other researchers inquire about why an item was translated in a particular manner or want to suggest that the translation itself is faulty. When those moments arrive – and they will for most medical research instrument translations if others are actively interested in the use of that particular instrument – the availability of documentation that adequately demonstrates accountability by the Team Leader of the translated document is generally regarded as a relief. At the same time, the structure of the current version of the Translation Logs (revision December 1, 2009) is based on extensive use of several prior versions and feedback with users.

Consequently, the International Consortium proposes that the level of documentation constructed via the Summary Report is a realistic step that will further the process of scientific instrument development. Inquiries about this document and/or suggestions for improvement should be sent to Richard Ohrbach, DDS, PhD, at ohrbach@buffalo.edu.

The Consortium website contains the starting Log (Log B) for many of the instruments translated by Consortium members. Team Leaders and translators are encouraged to consult the website for Log B document pertaining to an instrument to be translated; if that Log B does not yet exist, users are encouraged to contribute to the library of resource materials.

Please consult separate MS Word-based files, as contained in a ZIP file, for completion of Logs during Translation.

The blank Logs are attached in the following pages.

Summary Report on the Cross-Cultural Adaptation of an International RDC/TMD Consortium Instrument Phase I: Translation and Cultural Adaptation

Review Handbook before completing Tables.
Use TAB key within Tables in order to create more rows as needed for Instrument related tables.

Table 1. Team Leader Information

Name–Family	Name–Given	Name–MI	Date (dd/month/yyyy)
ENTER	ENTER	ENTER	ENTER

Institution	Street1
ENTER	ENTER

Street2	City	Postal Code	Country
ENTER	ENTER	ENTER	ENTER

Phone	Fax	Email
ENTER	ENTER	ENTER

NB: If more than one Team Leader, i.e., Co-Team Leaders, is involved, create a blank copy of Table 1 for the other Team Leader; name the Tables as 1a& 1b.

Table 2. Instrument Information

Source Instrument Name	Version date/number	Source Language
ENTER	ENTER	ENTER

Target Country	Groups or Cultures for this instrument (include all)	Target Language
ENTER	ENTER	ENTER

Target Instrument Name (in target language)	Target Instrument Name (back-translated to English)
ENTER	ENTER

Table 3. Translation Team Participants		
Stage	**Name**	**Qualifications/Title**
Translation Coordinator		
Forward Translator 1		
Forward Translator 2		
F-T1&2 prepared by		
Back Translator 1		
Back Translator 2		
B-T1&2 prepared by		
Reviewer		
Table 4. Expert Review Panel (list everyone, even if already listed above)		
Panel Coordinator		
Clinician		
Psychologist		
Methodologist		
Language Expert		
Other 1: <specify role>		
Other 2: <specify role>		
Other 3: <specify role>		

Table 5. Translation Summary Checklist of Translated Document Sections		
Stage	**Documentation**	
	Date	**Comments**
Forward Translation (F-T)		
Forward -Translation 1		
Forward - Translation 2		
F-T1&2 Synthesis		
Back Translation (B-T)		
Back - Translation 1		
Back - Translation 2		
B-T1&2 Synthesis		
Independent Review		
Independent Review		
Expert Panel Review		
Pre-Final Translation of Instrument		
Report: Discrepancies & Recommendations		
Final Translation		

Review Handbook before completing Tables.
Use TAB key within Tables in order to create more rows as needed Tables 2-5.

Table 1. Instrument and Translator Information			**Phase** (*highlight one*) **-1-** **-2-** **-3-** **-4-**
Source Instrument ENTER	Source Language ENTER	Target Language ENTER	
Select Translator (*highlight one*) F Translator 1 F Translator 2	Forward Translator Name ENTER	Date ENTER	
First language ENTER	Forward Translator Profile (*highlight one*): Aware of health status concept Naive to concept		

Table 2. Instrument Instructions	
Source	**Forwarded Translated Version (F-T1 / F-T2)**

Table 3. Item Content		
Item #	**Source**	**Forwarded Translated Version (F-T1 / F-T2)**

Table 4. Recurring Response Options		
Item #	**Source**	**Forwarded Translated Version (F-T1 / F-T2)**

Table 5. Scoring Instructions	
Source	**Forwarded Translated Version (F-T1 / F-T2)**

Review Handbook before completing Tables.
Use TAB key within Tables in order to create more rows as needed for Tables 2-6.

Table 1. Instrument Information			**Phase** (highlight one) **-1- -2- -3- -4-**
Source Instrument ENTER	Source Language ENTER	Target Language ENTER	Date ENTER

Table 2. Instrument Instructions		
Forward Translation: F-T1	**Forward Translation: F-T2**	**Synthesis Translation (F-T1&2)**

Table 3. Item Content			
Item #	**Forward Translation: F-T1**	**Forward Translation: F-T2**	**Synthesis Translation (F-T1&2)**

Table 4. Recurring Response Options			
Item #	**Forward Translation: F-T1**	**Forward Translation: F-T2**	**Synthesis Translation (F-T1&2)**

Table 5. Scoring Instructions		
Forward Translation: F-T1	**Forward Translation: F-T2**	**Synthesis Translation (F-T1&2)**

Table 6. Synthesis Process Report of Discrepancies		
Item #	**Describe Issue**	**Resolution**

Review Handbook before completing Tables.
Use TAB key within Tables in order to create more rows as needed for Tables 2-5.

Table 1. Instrument and Translator Information			**Phase** *(highlight one)* **-1-** **-2-** **-3-** **-4-**
Source Instrument ENTER	Source Language ENTER	Target Language ENTER	
Select translator *(highlight one)* B Translator 1 B Translator 2	Back Translator name ENTER	Date ENTER	
First language ENTER	Back Translator Profile *(highlight one)*: Aware of health status concept Naive to concept	**Back Translator:** Quality Rating of Forward Translation *(highlight one)*: Fair Good Excellent	

Table 2. Instrument Instructions		
Synthesis Translation (F-T or F-T1&2)	**Back Translation (B-T)**	**Source**

Table 3. Item Content			
Item #	**Synthesis Translation (F-T or F-T1&2)**	**Back Translation (B-T)**	**Source**

Table 4. Recurring Response Options			
Item #	**Synthesis Translation (F-T or F-T1&2)**	**Back Translation (B-T)**	**Source**

Table 5. Scoring Instructions		
Synthesis Translation (F-T or F-T1&2)	**Back Translation (B-T)**	**Source**

Review Handbook before completing Tables.
Use TAB key within Tables in order to create more rows as needed for Table 2.

Table 1. Instrument and Reviewer Information		**Phase** *(highlight one)* **-1- -2- -3- -4-**
Source Instrument ENTER	Source Language ENTER	Target Language ENTER
Reviewer Name ENTER		Date ENTER
Profile of reviewer (*highlight one*): Very familiar with health status concept Familiar with health status concept Aware of health status concept Naive to concept		Total Discrepancy Count: ENTER

Table 2. Independent Reviewer				
Item #	Source Item	Reviewer: Describe Problem	Developer Response	Reviewer Recommendation and Team Leader Action

Review Handbook before completing Tables.
Use TAB key within Tables in order to create more rows as needed for Tables 2-5.

Table 1. Instrument information			
Source Instrument ENTER	Source Language ENTER	Target Language ENTER	Date ENTER

Table 2. Instrument Instructions			
Source	Iteration	Forward Consolidated Translation (F-T or F-T1&2)	Back Translation (B-T)

Table 3. Item Content				
Item	Source	Iteration	Forward Consolidated Translation (F-T or F-T1&2)	Back Translation (B-T)

Table 4. Recurring Response Options				
Item	Source	Iteration	Forward ConsolidatedTranslation (F-T or F-T1&2)	Back Translation (B-T)

Table 5. Scoring Instructions			
Source	Iteration	Forward Consolidated Translation (F-T or F-T1&2)	Back Translation (B-T)

Table 1. Instrument and Reviewer Information		
Source Instrument ENTER	Source Language ENTER	Target Language ENTER
Panel Member Name ENTER	Language Base (highlight one) Bilingual Monolingual	Date ENTER
Target Condition, Culture, Group ENTER		

Instructions:

Use TAB key within Tables in order to create more rows as needed for Tables 2-5.

Developer: Complete first line of Table 1 and name of Panel Member, and transfer source and target translation units from Log F to this Log.

Expert Panel Member:

1. Select Language base in Table 1, and enter date.
2. Enter Sufficiency rating of the Target translation:
   Adequate, Questionable, Inadequate.
3. Provide Content Validity Rating for item-relevance:
   1 = not relevant
   2 = somewhat relevant
   3 = quite relevant
   4 = very relevant
4. In Comments field, enter explanations as follows:

Content validity	***Sufficiency***		
	Adequate	***Questionable***	***Inadequate***
1: not relevant ***2: Somewhat relevant***	Describe nature of the problem.	Indicate "Needs special review at next stage". Comments may also be added.	
3: quite relevant ***4: very relevant***	No comment required.	Describe nature of the problem.	

:

Table 2: Instrument Instructions					
Item #	**Source**	**Target**	**Sufficiency**	**Content Validity Rating**	**Comments**

Table 3: Item Content					
Item #	**Source**	**Target**	**Sufficiency**	**Content Validity Rating**	**Comments**

Table 4: Recurring Response Options					
Item #	**Source**	**Target**	**Sufficiency**	**Content Validity Rating**	**Comments**

Table 5: Scoring Instructions				
Source	**Target**	**Sufficiency**	**Content Validity Rating**	**Comments**

Table 6: Overall Comments

Review Handbook before completing Tables.
Use TAB key within Tables in order to create more rows as needed for Tables 2-3.

Table 1. Instrument Information		
Source Instrument ENTER	Source Language ENTER	Target Language ENTER

Content Validity Rating for item relevance based on median score of the individual ratings from committee members. Include range of scores as appropriate.
Guideline for interpretation is: 1 = not relevant
2 = somewhat relevant
3 = quite relevant
4 = very relevant

Table 2: Report of discrepancies and their recommendation:				
Item #	**Source**	**Target**	**Content Validity Rating**	**Describe problem**

Table 3. Panel Recommendation, Developer Action, and Reviewer Response			
Item #	**Expert Panel Recommendation**	**Developer Response & Explanation**	**Reviewer Response**

Review Handbook before completing Tables.
Use TAB key within Tables in order to create more rows as needed for Tables 2-5.
INCLUDE WITH THIS LOG A FULLY-FORMATTED INSTRUMENT SUITABLE FOR INITIAL ADMINISTRATION

Table 1. Instrument Information			
Source Instrument ENTER	Source Language ENTER	Target Language ENTER	Date ENTER

Table 2. Instrument Instructions	
Source	**Pre-final Translation (F-T or F-T1&2)**

Table 3. Item Content		
Item #	**Source**	**Pre-final Translation (F-T or F-T1&2)**

Table 4. Recurring Response Options		
Item #	**Source**	**Pre-final Translation (F-T or F-T1&2)**

Table 5. Scoring Instructions	
Source	**Pre-final Translation (F-T or F-T1&2)**

Review Handbook before completing Tables.

** This evaluation is based on the guidelines given in Guillemin (1993).

Table 1. Instrument and Reviewer Information		
Source Instrument Name ENTER	Version date/number ENTER	Source Language ENTER
Target Instrument Name (in target language) ENTER	Version date/number ENTER	Target Language ENTER
Reviewer for Consortium ENTER		Date ENTER

Table 2. Cultural Characteristics of Developed Instrument		
	Source	Target
Language		
Culture		
Country		

Table 3. Summary Evaluation (For use by the Consortium Review Committee)		
	Evaluation (highlight one)	**Score (number of "yes" checks)**
1. Translation technique Used two or more translations? Translating into their mother tongue? Was one translator aware of concept & condition of clients? Was one translator naive to concept?	 **Yes No** **Yes No** **Yes No** **Yes No**	 _____**/4**
2. Synthesis of translated versions Synthesis of translations done?	 **Yes No**	 _____**/1**
3. Back translation Translating into their mother tongue? Not aware of concepts/condition?	 **Yes No** **Yes No**	 _____**/2**
4. Expert committee Committee review done? Membership of committee appropriate? Details of decisions & issues provided?	 **Yes No** **Yes No** **Yes No**	 _____**/3**
TOTAL SCORE (range: 0-1)		**ENTER**

Summary Report on the Cross-Cultural Adaptation of an International RDC/TMD Consortium Instrument Phase II: Cultural Equivalency and Validation

Review Handbook before completing Tables.

Table 1. Team Leader Information					
Name–Family ENTER	Name–Given ENTER		Name–MI ENTER		Date (dd/month/yyyy) ENTER
Institution ENTER	Street1 ENTER				
Street2 ENTER	City ENTER		Postal Code ENTER		Country ENTER
Phone ENTER	Fax ENTER	Email ENTER			

NB: If more than one Team Leader, i.e., Co-Team Leaders, is involved, repeat Table 1 for the second individual. Rename tables as Table 1a & 1b.

Table 2. Instrument Information					
Source Instrument ENTER			Version date/number ENTER		Source Language ENTER
Target Country ENTER	Target Culture ENTER				Target Language ENTER
Target Instrument Name (in target language) ENTER			Target Instrument Name (back-translated to English) ENTER		

Table 3. Summary Checklist of Qualitative and Quantitative Item Testing		
Phase	**Documentation Includes**	
	Date	**Comments**
Pretesting, Field Testing, and Formal Assessment		
Pretesting 1		
Pretesting 2		
Pretesting 3		
Field Testing		
Formal Assessment		

Table 4. Expert Panel for Revision (if use bi-national or cross-national review, modify content of table accordingly)		
Panel Coordinator		
Clinician		
Psychologist		
Methodologist		
Language Expert		
Other 1: <specify role>		
Other 2: <specify role>		

Table 5. Score Standardization (please follow this outline for a narrative summary, and enter below or on separate sheet attached at end of this Log).	
Describe reference population	
Describe method for standardization	
Describe scoring rules	
Explain norms	

Table 6. Validation Research (please follow this outline for a narrative summary, and enter below or on separate sheet attached at end of this Log).	
Describe population(s)	
Describe method(s) for validation	
Describe outcomes	

Table 7. User Manual		
Title		
Author		
Publication availability (highlight one)	Copy-righted	Public Domain
Source to contact for obtaining manual		

Review Handbook before completing Tables.
Use TAB key within Tables in order to create more rows as needed for Table 5.

Table 1. Instrument Information		
Source Instrument Name ENTER	Version date/number ENTER	Source Language ENTER
Target Instrument Name (in target language) ENTER	Version date/number ENTER	Target Language ENTER

Table 2. Pretesting - 1			
Mode of Administration (highlight one):	Single Interview	Focus Group	
Sample Description			
Inclusion Criteria			
Demographics	Total	Males	Females
Age (mean, SD)			
Sample Size (n)			

Table 3. Pretesting - 2			
Brief explanation of why Pretesting-2 was performed:			
Mode of Administration (highlight one):	Single Interview	Focus Group	
Sample Description			
Inclusion Criteria			
Demographics	Total	Males	Females
Age (mean, SD)			
Sample Size (n)			

Table 4. Pretesting - 3			
Brief explanation of why Pretesting-3 was performed:			
Mode of Administration (highlight one):	Single Interview	Focus Group	
Sample Description			
Inclusion Criteria			
Demographics	Total	Males	Females
Age (mean, SD)			
Sample Size (n)			

Table 5. Pretest Based Revised Translation			
Item #	Source Item	Previous Translation	Revised Translation

Review Handbook before completing Tables.

Table 1. Instrument Information		
Source Instrument Name ENTER	Version date/number ENTER	Source Language ENTER
Target Instrument Name (in target language) ENTER	Version date/number ENTER	Target Language ENTER

Table 2. Sample description			
Mode of Administration (highlight one):	Single Interview	Focus Group	
Description			
Inclusion Criteria			
Demographics		**Males**	**Females**
Age(mean, SD			
Sample Size(n)			

Table 3. Study description (please follow this outline for a narrative summary, and enter below or on separate sheet attached at end of Log I).	
Reliability: (internal consistency, test-retest reliability)	
Describe methods	
Summarize results	
Validity	
Describe methods used (list constructs, how they were measured)	
Summarize results for each construct:	
Responsiveness	
Describe methods	
Summarize results	
Other psychometric testing (e.g. Rasch modeling, CFA)	
Describe methods	
Summarize results	

LOG M – Final-revised Translation (form version 2013 May 11)

Review Handbook before completing Tables.
Use TAB key within Tables in order to create more rows as needed for Tables 2-5.

INCLUDE WITH THIS LOG A FINAL FULLY-FORMATTED INSTRUMENT FOR ADMINISTRATION

Table 1. Instrument Information		
Source Instrument Name ENTER	Version date/number ENTER	Source Language ENTER
Target Instrument Name (in target language) ENTER	Version date/number ENTER	Target Language ENTER

Table 2. Instrument Instructions		
Source	**Translation from Log H**	**Final-revised Translation**

Table 3. Item Content			
Item #	**Source**	**Translation from Log H**	**Final-revised Translation**

Table 4. Recurring Response Options			
Item #	**Source**	**Translation from Log H**	**Final-revised Translation**

Table 5. Scoring Instructions		
Source	**Translation from Log H**	**Final-revised Translation**

Table 6. Summary Description of Changes		
Change	**Explanation**	**Impact on Scoring or Interpretation**

Review Handbook before completing Tables.

Table 1. Instrument Information		
Source Instrument Name ENTER	Version date/number ENTER	Source Language ENTER
Target Instrument Name (in target language) ENTER	Version date/number ENTER	Target Language ENTER

Table 2. Sample description			
Mode of Administration (highlight one): Single Interview Focus Group			
Description			
Inclusion Criteria			
Demographics	Total	Males	Females
Age(mean, SD**)**			
Sample Size (n)			

Table 3. Study description (please follow this outline for a narrative summary, and enter below or on separate sheet attached at end of this Log).	
Reliability: (internal consistency, test-retest reliability)	
Describe methods	
Summarize results	
Validity	
Describe methods used (list constructs, how they were measured)	
Summarize results for each construct:	
Responsiveness	
Describe methods	
Summarize results	
Other psychometric testing (e.g. Rasch modeling, CFA)	
Describe methods	
Summarize results	

Review Handbook before completing Tables.

Table 1. Instrument Information		
Source Instrument Name ENTER	Version date/number ENTER	Source Language ENTER
Target Instrument Name (in target language) ENTER	Version date/number ENTER	Target Language ENTER

Table 2. Sample description			
Description			
Inclusion Criteria			
Demographics	Total	Males	Females
Age(mean, SD**)**			
Sample Size(n)			

Repeat Table 2 for any additional validation study groups used. Name them Table 2a, 2b, etc.

Table 3. Study description (please follow this outline for a narrative summary, and enter below or on separate sheet attached at end of this Log).	
Reliability: (internal consistency, test-retest reliability)	
Describe methods	
Summarize results	
Validity	
Describe methods used (list constructs, how they were measured)	
Summarize results for each construct:	
Responsiveness	
Describe methods	
Summarize results	
Other psychometric testing (e.g. Rasch modeling, CFA)	
Describe methods	
Summarize results	

Review Handbook before completing Tables.

** This evaluation is based on the guidelines given in Guillemin (1993).

Table 1. Instrument and Reviewer Information		
Source Instrument Name ENTER	Version date/number ENTER	Source Language ENTER
Target Instrument Name (in target language) ENTER	Version date/number ENTER	Target Language ENTER
Reviewer for Consortium ENTER		Date ENTER

Table 2. Cultural Characteristics of Developed Instrument		
Name of translated instrument:		Source version used
	Source	Target
Language		
Culture		
Country		

Table 3. Summary Evaluation (For use by the Consortium Review Committee)			
	Evaluation (highlight one)		**Score (number of "yes" checks)**
1. Pretesting			
Used focus group?	**Yes**	**No**	
Used clinical sample?	**Yes**	**No**	
Used bilingual sample and reported % agreement?	**Yes**	**No**	
Probe technique used and reported on?	**Yes**	**No**	_____**/4**
2. Field Testing			
Sufficient sample size?	**Yes**	**No**	
Bilingual sample used?	**Yes**	**No**	
Item reliability used?	**Yes**	**No**	
Sample response characteristics reported?	**Yes**	**No**	_____**/4**
3. Instrument Revision			
Translation units adequately considered for revision?	**Yes**	**No**	
Expert Panel involvement?	**Yes**	**No**	
Revised form of instrument prepared?	**Yes**	**No**	_____**/4**
Summary description of changes provided?	**Yes**	**No**	
4. Formal Assessment			
Defined subject population?	**Yes**	**No**	_____**/2**
Sufficient statistics?	**Yes**	**No**	
5. Score Standardization			
Defined reference population?	**Yes**	**No**	
Justified choice of raw vs IRT-derived scores?	**Yes**	**No**	_____**/3**
Scoring rules and norms established?	**Yes**	**No**	
6. Validation Research			
Defined population?	**Yes**	**No**	
Validation methods appropriate?	**Yes**	**No**	
Outcomes support validity?	**Yes**	**No**	_____**/3**
7. User Manual			
Manual constructed?	**Yes**	**No**	
Manual available?	**Yes**	**No**	_____**/3**
Manual sufficient?	**Yes**	**No**	
TOTAL SCORE (range: 0-1)			**ENTER**