Clinical Child and Family Psychology Review, Vol. 8, No. 2, June 2005 (© 2005)
DOI: 10.1007/s10567-005-4752-1

Translation and Cross-Cultural Adaptation of Assessment Instruments Used in Psychological Research With Children and Families

Brigit M. van Widenfelt,[1,3] Philip D. A. Treffers,[1] Edwin de Beurs,[2] Bart M. Siebelink,[1] and Els Koudijs[1]

With the increased globalization of psychology and related fields, having reliable and valid measures that can be used in a number of languages and cultures is critical. Few guidelines or standards have been established in psychology for the translation and cultural adaptation of instruments. Usually little is reported in research publications about the translation and adaptation process thus making it difficult for journal readers and reviewers to adequately evaluate the equivalency and quality of an instrument. In this study, issues related to the translation and adaptation of assessment instruments for use in other cultures and/or languages are addressed. Existing literature on translation is reviewed and examples from the clinical child and family psychology field are given to illustrate relevant issues. Suggestions are made for avoiding common translation errors.

KEY WORDS: translation; cross-cultural; assessment; child; family.

In clinical child and family psychology and related fields numerous assessment instruments are translated yearly into other languages and/or adapted for specific cultures. The majority of instruments are originally developed in English. For researchers working with populations in non-English speaking countries or cultural groups that differ greatly from the population used to develop the instrument, translating and adapting an established English language measure is an efficient solution for the lack of available instruments. Using established measures further allows for cross-cultural comparison of findings as well as use in international trials. For example, Bornstein et al. (1998) translated two American measures on parenting into six languages in order to compare mother's reports on parenting from seven different cultures. Though occurring less frequently, new assessment tools are also developed in non-English speaking countries and sometimes such instruments are then translated into other languages, including English. A Swedish instrument for assessing early parental rearing experiences, for example, has been translated into English and many other languages (EMBU; Perris, Jacobsson, Lindstrom, von Knorring, & Perris, 1980).

The least common, but most desirable scenario for cross-cultural use of assessment instruments is the development of a measure in several countries at the same time (Sprangers, Cull, Bjordal, Groenvold, & Aaronson, 1993). In such a case, a common set of items is generated that is relevant for a number of countries or cultures (parallel approach) or culture-specific items are developed at the same time in different cultures (simultaneous approach; Anderson, Aaronson, Bullinger, & McBee, 1996).

[1]Department of Child and Adolescent Psychiatry, Leiden University Medical Center, Leiden, The Netherlands.
[2]Department of Psychiatry, Leiden University Medical Center, Leiden, The Netherlands.
[3]Address all correspondence to Brigit M. van Widenfelt, Department of Child and Adolescent Psychiatry, Leiden University Medical Center, ACKJP Curium, Endegeesterstraatweg 27, 2342 AK Oegstgeest, The Netherlands; e-mail: b.m.van.widenfelt@umail.leidenuniv.nl.

Decentering is an important aspect of these approaches, referring to a process in which the source language measure is open to modification during the translation procedure, as all versions are viewed as equal (Brislin, Lonner, & Thorndike, 1973). Because the simultaneous or parallel approaches are uncommon, this study addresses the more common sequential approach (Anderson et al., 1996), in which a questionnaire is translated from an original source language into another language.

When a known assessment measure, such as the *Child Behavior Checklist* (CBCL; Achenbach, 1991) which has been translated into over 33 languages and more than 16 cultures, is used in a dataset collected in another culture and/or language, it is commonly assumed by journal readers, reviewers and editors that it is the same measure as the original version. That is, equivalency is assumed in terms of the translated or adapted version having items with the same meaning and general wording, corresponding response categories, identical instructions, similar psychometric properties, and appropriate norms. This assumption is quite often false. In most cases, the reader cannot even assess equivalency because of the current practice to report very little in research publications about the translation and adaptation process of the instructions, items and response categories, as well as the psychometric properties of the translated version. Often it is only reported that a measure was translated and administered to participants in another culture. In cases where a "translation procedure" is reported, it is usually stated that a (single) translation and then a back-translation was conducted. There is, however, a growing international literature indicating that a simple "single forward and back-translation procedure" is an insufficient method of making and checking the quality of a translation and, as a sole method, can result in a poor translation (Brislin et al., 1973; Hambleton, 2001; Perneger, Leplege, & Etter, 1999; Van de Vijver & Tanzer, 1997). The translation and adaptation task is complex and requires a combination of techniques (Hambleton, 2001). Only when the quality of a translation (and adaptation) has been established and the process is reported is it possible to make comparisons between studies and datasets, draw conclusions about the constructs assessed, or make statements about culture differences. To our knowledge, translation and cross-cultural adaptation of assessment instruments has not yet received adequate attention in child and family psychology journals.

The purpose of this study is to demonstrate that preparing a measure for a new language and culture is a major undertaking involving many steps and considerations. Foremost, our goal is to assist researchers in avoiding common errors in the translation process. Thus, it is not our intention per se to prescribe a set of guidelines aimed at making a perfect translation. There is insufficient empirical data on the use of guidelines to do that. Rather we would like to assist researchers in the child and family field in creating "good-enough" translations. We summarize important points in the translation specialty literature combined with our own experience. We use many of our own examples in translating English language instruments into Dutch throughout this paper. In addition, when possible we make reference to examples from the child and family literature. Secondly, this study aims to assist readers, reviewers, and editors of journals in evaluating the "equivalency" and quality of translated assessment tools.

RESEARCHERS' ORIENTATION TO CULTURE

Having different translations and/or adaptations of the same questionnaire for one language is not uncommon and can result from different methods used or choices made during the translation process. These differences may in large part be determined by the reseachers' orientation towards translating a measure for use in another language and culture.

To date, the most common assumption in health research appears to be that culture has only a minimal impact on the construct being measured, and therefore the way a construct is defined and operationalized in one culture can be applied directly in another culture (Herdman, Fox-Rushby, & Badia, 1997). This assumption can lead to standardized instruments being literally translated into other languages with the goal of having a back-translation that is linguistically the same as the original questionnaire. If this approach is used too strictly, it may result in inadequate consideration of the applicability of concepts in the new culture and may ignore culturally idiosyncratic ways of expression.

A contrasting orientation is when culture is viewed as having a potentially significant impact on how concepts are expressed in various cultures (Herdman et al., 1997). Psychopathology might be viewed as universal, but culture may play a role in variations of expression. In this view, it is assumed that instruments used in another country will likely need to go through some culture-specific adaptation. This assumption is consistent with recent pleas in the

literature for a more cultural sensitive approach to research (Garcia Coll, Akerman, & Cicchetti, 2000; Parke, 2000). The goal of a making a translation from a source language version into a new language would thus be to strive for equivalence in terms of keeping the new version as similar as possible to the source language version (i.e., wording of items). Yet at the same time strive for a conceptually equivalent version that measures the same constructs and in which constructs retain the same meaning, which may involve making adaptations to the original language version.

In the adaptation process, certain items may be more in need of culture-specific adaptation than others. Descriptive items assessing symptoms may in some cases be relatively easier to translate, than items intended to assess a more culturally determined construct such as self-esteem. For example, Wang and Ollendick (2001) discovered that in the Chinese culture there is not an equivalent term for "self-esteem" as defined in Western cultures. In the Chinese culture, the self is defined within relationships and is not valued as separate. Thus, for example, in contrast to Western culture, Chinese children learn to minimize their own role in their achievements and good behavior and attribute their strengths to the help of others. In such a scenario, simply translating items developed to capture a western-defined construct is of no avail. Instead new items need to be generated to capture how the Chinese make self-appraisals related to self-esteem.

Canino and Bravo (1999) point out that measuring impairment of a child's psychosocial functioning and adaptation in areas such as family, school, friends, and community may be particularly dependent on the social and cultural context. They describe differences on the CBCL (Achenbach, 1991) Social Competence Scores for Puerto Rican and Anglo children even after matching subjects for age, sex, and socioeconomic status. The Puerto Rican sample scored lower on social competence than the Anglo sample. They attribute these scores to the Puerto Rican sample reporting less involvement in sports, hobbies, organizations and jobs, apparently reflecting the children's limited access to such. Interestingly, the Puerto Rican sample reported to have more frequent contact with friends and better relations with family and siblings than the Anglo sample. Thus, how a construct is assessed in a new culture and language area may need to be adapted.

Though perhaps less complex than the above-mentioned examples, the translation of symptoms may also in certain cases require a culturally sensitive approach (see also Rogler, 1999). A problem or symptom may be more pathological or more familial in one culture versus another. Cultural norms and beliefs can influence the rater's view of the acceptability of individual behavior or characteristics as well as what types of interactions and relationships are acceptable (Rubin, 1998). Rubin (1998) gives examples from his work with his Chinese colleagues on behavioral inhibition. He states that extreme social wariness across cultures may be consistently viewed as an anxious reaction to novel situations and even manifest itself in the same way. However, the cultural "meaning" and the social responses may differ. Rubin (1998) further comments that the shy or inhibited behavior of children in more individualistically oriented countries such as in North America may be disadvantaged compared to more sociable and assertive peers. However, in his work with his Chinese colleagues, he did not find negative outcomes for such children, nor was such behavior viewed as maladaptive by either peers or adults. Another example is provided by Weisz et al. (1987). They report that the Thai translation of the *Child Behavior Checklist* (CBCL; Achenbach, 1991) item "swearing or obscene language" resulted in Thai parents reporting twice as much swearing among their children as did American parents. This likely resulted from the fact that the closest word to swearing in Thai also included language that would only be considered impolite in the United States, thereby reflecting cross-cultural differences in the breadth of the concept of swearing (Weisz et al., 1987). Thus even when a translation is done with great care, culturally idiosyncratic nuances may not be corrected for.

In the above we provide some examples of when instruments are used in another country. Geisinger (1994) points out, however, that some of the same issues in transporting assessment instruments across countries may also apply to subpopulations within a country if they differ enough from the original population for which the measure was developed for and sampled on (e.g., Asian Americans, African Americans, Hispanic Americans, and Native Americans).

PRACTICAL ASPECTS OF MAKING A TRANSLATION

Through conducting a literature search in the databases Pub Med and PsychINFO and reviewing reference lists from relevant articles, we discovered several sources of guidelines for translating

assessment instruments in specialty books and journals that can be useful in psychological research with children and families as well. Two books from the early 70's on cross-cultural methodology provide guidelines for translating (Brislin et al., 1973; Werner & Campbell, 1970). A second source of guidelines specifically for educational tests are those described in articles by authors of the *International Test Commission* (e.g., Hambleton, 2001). Further, Guillemin, Bombardier, and Beaton (1993) have developed translation and adaptation guidelines for evaluating cross-cultural adaptations of quality of life questionnaires. We did identify one relevant article in a psychology journal written by Geisinger (1994). And finally, we identified one chapter that was included in a book on child and adolescent diagnostic assessment which reviews types of equivalency and addresses some issues related to translating diagnostic interviews (see Canino & Bravo, 1999).

The steps in preparing a translation of an assessment instrument described in the above literature overlap to some extent with steps described in the literature on questionnaire development in general. Procedures such as testing initial versions on peers, piloting with representative populations, and conducting data analyses need to be carried out again with a translated version of a measure. Because such procedures are described elsewhere (see Sudman & Bradburn, 1982), we will not discuss them in detail. Rather, we focus in the following section on the practical aspects and issues specifically relevant to translating and adapting instruments.

Contacting the Original Author

Prior to embarking on the demanding task of making a translation and cross cultural adaptation, it is obviously important to first find out if a version of the questionnaire already exists in the language and culture of interest. We recommend contacting the original author, reviewing journals, dissertations, and contacting same-country colleagues. Multiple versions of translated questionnaire often exist and original authors are usually not aware of this. For example, during the initial phase of investigating a Dutch version of the *Children's Depression Inventory* (CDI; Kovacs, 1992), we found three unofficial Dutch versions in circulation in the Netherlands and an additional version in Belgium, each version differing greatly.

Note, even if the translation of a questionnaire exists in another same language region, it still must be reviewed for the need to make adaptations related to cultural differences (such as differences between English spoken in Great Britain and North America or Dutch spoken in Belgium and the Netherlands). The same words in the same language may not have semantic equivalence across cultures or countries. For example, in the translation of the *Minnesota Multiphasic Personality Inventory* (MMPI-2; Derksen, de Mey, Sloore, & Hellenbosch, 1993) into Dutch, the word "cry" was translated as "huilen" for the Netherlands version. However, "huilen" in Belgium refers to howling (as a wolf). Therefore, the Belgian version needed to use another word "wenen" to properly capture the type of crying intended in the item.

A second reason to contact the original author is that the quality of the translation is likely to be better if the original author is available for discussion about translating difficulties and/or interpretation of data. In our experience, original authors tend to be the most available and involved if they choose to keep the author rights of the instrument (versus using a publisher). Dr Robert Goodman in the United Kingdom has set an excellent example for the field by setting up a website in which a brief measure for parents, teachers and children to rate emotional and behavioral problems, the *Strengths and Difficulties Questionnaire*, is available for use in over 40 languages (see www.sdqinfo.com). The website provides scoring information and an overview of research reports. Goodman's model is not only a user-friendly way to make questionnaires available; but it also assures that there is only one version of the questionnaire per language in use. In contrast, if the original questionnaire has been published by or is in the process of being published by a publishing company, the original author will likely refer the "translating" author to communicate further with the publisher due to the transfer of legal ownership of the instrument. Unfortunately in this type of construction the original author sometimes takes distance from the questionnaire and the new language versions. The translating author instead deals with a publisher. In such a set up there is sometimes less attention given to and allowed for the quality and psychometric properties of translated versions.

Creating a Translation Team

The first step in making a translation is deciding *how many* persons and *who* to involve in the

translation and adaptation task. Using only a single translator is far from ideal (Hambleton, 2001). Guillemin et al. (1993) recommend having at least two independent translators. Our experience is that it is fruitful to work in a team of at least three to four persons. More important is *who* is involved in the process. It is our observation that differences in how a questionnaire is translated can in part be accounted for by whether or not a *native speaker* is involved in the translation process. Ideally we recommend that a native speaker who is *bilingual* and bicultural should be involved early on in the translation process, or at least involved in a later phase in which he or she can review the translation and assist in choosing among competing alternatives. Some studies also use bilinguals to fill in the questionnaire in both the original and translated version (Brislin et al., 1973).

The type of professionals who are involved in the translation process may also influence the outcome of a translation. For example, one of the Dutch versions of the CDI (Kovacs, 1992) that we reviewed was translated by (nonclinical) researchers and adapted for use in a nonclinical population. In this version items are "softened" in general and the item on suicide is removed in order to make the CDI more acceptable to children and adolescents who would be assessed. Involving a clinician in the adaptation process could be helpful in assessing the clinical importance of the phrasing of certain items that might be crucial for eventually distinguishing cases in the clinical range. If the questionnaire is about a specific disorder, such as obsessive compulsive disorder (OCD), then a person with *expertise* in assessing and/or treating OCD could be useful in reviewing the phrasing of items. Finally, it is useful if team members have experience or understanding of the cultural context and/or developmental phase of the population that will be studied. Thus, it is logical, for example, to include a person who studies or works with young children in the case of translating items for younger children as they may have experience with how to choose words and phrases at the right developmental level. In sum, the translators need to have qualities beyond knowledge of the two languages, such as knowledge of the subject matter of the instrument and the two cultures (Geisinger, 1994; Hambleton, 2001).

Team Procedures and Guidelines

It is our practice that each team member first makes an independent translation of the measure. After individual translations are made, the team members meet together, and through dialogue decide on the best version for each item. If team members cannot agree or are unsure about the best translation of an item, it is often helpful to consult a dictionary (and a thesaurus), the original authors, and/or discuss difficult items with family, friends and colleagues to get "outside" input. Some researchers use more than one team so that the teams can check, review and discuss the translations made. For example, Jeanrie and Bertand (1999) used a team of four translators to make independent translations of an American personality measure for English and French Canadian adults and then used a second team to judge the translations. Guillemin et al. (1993) recommend a multidisciplinary review committee should be organized to evaluate the translation, along with the back-translations and the original. Such a committee may be especially beneficial if the initial translation team is small and not representative enough to evaluate the measure (i.e., including a bilingual member and professionals in the area of study). Geisinger (1994) also suggests using a reviewer team to react to the translation made by the original team both in writing and verbally. Translator teams cannot necessarily know which translation will be the best in this phase of the translation process, however. Thus sometimes it is useful to temporarily make one item into two items and wait with making a decision for the best item until field-testing and/or comparison through statistical analyses on a larger dataset has been done.

We think it is important for team members to be familiar with literature on questionnaire development (i.e., Sudman & Bradburn, 1982) and keep the principles in mind when making translation decisions. The following recommendations for questionnaire development have also been cited in previous translation guidelines (Brislin et al., 1973; Guillemin et al., 1993): use short and simple sentences; use active rather than the passive voice; repeat nouns rather than use pronouns; use specific rather than general terms; avoid using metaphors and colloquialisms; avoid using the subjective mode such as verb forms with could or would; avoid adverbs and prepositions telling "where" or "when"; avoid possessive forms where possible; avoid words indicating vagueness such as "probably"; avoid having two different verbs in one sentence if the verbs suggest different actions. Finally, the level of comprehension should be attended to as well. Even adult questionnaires should be at the reading level of elementary school

children. For example, Brislin et al. (1973) report that instruments that were of a third-grade reading level of difficulty (8 years of age) produced better translations than those of a seventh-grade level of difficulty (12 years of age). Similarly, Guillemin et al. (1993) recommend using language that can be understood by 10–12-year olds.

Sources of Error in Translating

Literal Translations

Problems with translated questionnaires are most often of the nature of items being translated too literally. This may result in phrases that do not make sense, sentences that are poorly constructed, or even the loss of the original meaning of the item. It is understandable that the translator would like to "stay close" to the original questionnaire, and that is in part the object of translating. However, this intention should not be at the cost of creating a good questionnaire! Further, translators should critically evaluate and note which items may not translate literally because of, for example, use of colloquialisms, a description of activities that are not common in the new culture or wording that results in awkward phrasing.

Sometimes an item may appear to translate easily and only through piloting experience and further data collection and statistical analyses will the researchers discover that a literal translation of the item is not applicable. For example, one of the items from the Negative Affect Self-Statement Questionnaire (NASSQ; Ronan, Kendall, & Rowe, 1994) that we translated, "I am a winner," was literally translated in a Belgian version of the questionnaire as "ik ben een winnaar." For the same item, it was clear from our pilot interviews with Dutch children and adolescents that such a literal translation was not culturally appropriate for use in the Netherlands, and thus risked insufficient endorsement and a skewed response pattern. Some children relayed to us that they understood the item but said that they would never say or think that, nor did they expect other Dutch children would. We discussed with the children similar translations such as "I am the best," but the children had the same reactions as they had to "I am a winner." Hence together with the children that we interviewed and later with our translation team we decided on "alles lukt me" (I will succeed in everything I do/I can do anything). This is an example of when a literal translation appears to risk little endorsement by respondents or is likely to be unclear to the respondents, and thus the emphasis should rather be placed on the translation of the intent of the item Our choice for an item that captured the intent of the original item, was supported by our data analyses that revealed the item had sufficient variance and loaded on the same factor as in the American version.

Thus in some cases, items may appear to be easily understood in a new culture, yet they are less common or relevant or less socially acceptable and thereby may warrant adaptation or what Guillemin et al. (1993) refer to as trying to achieve "experiential equivalence." It is often a difficult decision to make in the translation phase whether the item should be changed or kept as is until further data analysis. For example, Dutch translators of the *Multidimensional Anxiety Scale for Children* (MASC; March, Parker, Sullivan, Stallings, & Conners, 1997) were faced with the problem of translating "summer camp." It was possible to literally translate it (zomerkamp) but the relevance in Dutch culture was quite minimal as Dutch children spend their summer holidays with their family. Because some camps do exist (e.g., 1-week sailing camp) the authors chose to leave the item as is in the translation phase. In such a case, the item needs to be re-examined once data is collected to evaluate the endorsement pattern. Sometimes examples are broadened to increase relevancy. In an item of the Dutch translation of the *Anxiety Disorders Interview Schedule* for children (parent version) (ADIS- P; Siebelink & Treffers, 2001; Silverman & Albano, 1996) on school refusal, parents are asked if their child often goes to see the school nurse. Since nurses are rarely present in Dutch schools, the item was broadened to "school mentor, school doctor or school nurse."

Mistranslations

Another source of error in translations is mistranslations due to foreign author(s) not understanding an original item well enough. In the Dutch version of the SCL-90 (Derogatis, 1977; Arrindell & Ettema, 1986), Item 51, "Your mind going blank" has been translated as "Een gevoel van leegte" which back translates as "An empty feeling" or "Feeling empty." This is an obviously problematic translation as the meaning of the item is changed. Results of factor analyses confirm this fact: In the English version, item 51 belongs to the Obsessive-Compulsive subscale. The translated item of the Dutch version of the SCL-90 not surprisingly loads on the Depression subscale. Though factor

structures in new cultures may differ from that of the original culture, such as with the Dutch SCL-90, it is important to be able to interpret the new factor structure based on possible cultural differences and not on problematic item translations.

Altering, Adding and Deleting Items

Altering Items

Problematic items in a translation are, however, not always because of the translation being too literal or as a result of an insufficient understanding of an item. Often, several translations appear feasible and the translation team is faced with the choice of selecting the best one. Once alternatives are tried out with research subjects and data analyses are conducted, the best translation can be chosen. Slight alterations to an item may be necessary to achieve the most acceptable wording for the respondents. For example, we tried several Dutch translations for the word "friends" in the questionnaire item "some teenagers find it hard to make friends they can really trust but other teenagers are able to make close friends they can really trust" (*Self-Perception Profile for Adolescents*; SPPA; Harter, 1988). Our experience revealed subtle differences in the translation choice to be very influential for the way in which the item was understood and endorsed (Treffers et al., 2002). In this case, difficulty in "making" friends versus finding it difficult to "get" friends made a significant difference on which factor the item loaded on. In the Dutch culture, "making" friends was likely to be more associated with a social skill and loaded on the factor "social acceptance," whereas "getting" friends loaded on the same factor as the original questionnaire: close friendship.

A more obvious example of needing to alter the wording of an item is when idioms or colloquialisms are used that do not translate directly into a new culture. For example, in adaptations of the Fear Survey Schedule for Children—Revised (Ollendick, 1983) for British children, it was necessary to change items such as "Getting a shot from a nurse or doctor" to "Getting an injection from a nurse or doctor," "Being hit by a car or truck" to "Being hit by a car or lorry," and "Strange or mean looking dogs" to "Nasty-looking dogs" (Ollendick, Yule, & Ollier, 1991). Similar adaptations were necessary for use of the instrument in Australia (Ollendick, King, & Frary, 1989). Rogler (1999) provides additional examples. In such cases equivalents need to be found that capture the meaning of the item or expression (Guillemin et al., 1993).

On occasion an English word cannot be directly translated into another language, as there is no equivalent word and instead needs to be circumscribed. For example, an item in the Dutch ADIS-P (Siebelink & Treffers, 2001; Silverman & Albano, 1996), to assess conduct disorder reads, "Is your child known as a bully in your school?" Since there is no Dutch equivalent for "bully," yet the item is relevant as there are certainly Dutch bullies, it was decided to describe a bully as someone who "plays the boss and teases." Here it should be noted that teasing ("pesten" in Dutch) includes a broad range of behavior from light teasing to harassing and intimidation.

Generating New Items

Though generating new items may be an important step in the translation/adaptation process, it is most desirable if items for a final version remain similar in concept to the original version (and if possible it is practical for future comparison if the number of items are the same). If new items are generated, this might best be done in collaboration with the original author of the questionnaire. If the final translated version results in the addition of a number of new items for the purposes of the researcher (e.g., adding a new scale), it is advisable to review whether the questionnaire is still comparable to the original version or should be named differently. For example, when Weisz et al. (1987) translated the CBCL (Achenbach, 1991) for use in Thailand they wanted to make sure that the questions captured the patterns of child behavior problems in Thailand. Referral problems of admissions to Thai child guidance clinics were analyzed and on that basis at least 20 new items were added. To avoid confusion, the authors chose to adapt the name to "Thai Youth Checklist." One scenario in which adding new items may be necessary is when a replication of a factor structure is sought. In a case where a factor structure is not replicated yet item translations seem adequate, perhaps different items need to be added to capture the construct in a new culture.

Deleting Words, Items, or Scales

In some cases researchers may find that part of an item, the entire item or even an entire scale is not applicable or appropriate for use in a new culture. In such cases, if no alternatives are available,

the translators may decide to remove the item(s). For example, Weisz et al. (1987) dropped the item "Is your child in a special class?" from the CBCL (Achenbach, 1991), because most Thai schools have no special classes. Sometimes, only part of an item is not relevant and can be deleted. For example, one of the questions on the ADIS-P (Silverman & Albano, 1996) intended to assess school refusal instructs the interviewer to ask about anxiety related to "walking in the hallways or standing at his or her locker." because there are no school lockers and children carry their books with them in a school bag in the Netherlands, reference to lockers was removed from the Dutch version of the item and walking in the hallways was kept as is.

When an entire item is deleted and no alternative is created, then in some cases a problem for scoring is created. One possible solution if one wants to use the scoring system of the original questionnaire, is to use the original author's guidelines for dealing with missing data. For example, there may be a recommendation to prorate the scale score, such as with the SDQ, where at least 12 of the 20 items must be completed in order to prorate a total score.

Translation of Instructions, Response Categories, and Scoring Materials

Carefully translating the instructions of a questionnaire is also important. Not doing this in a precise way is a common oversight. Sometimes when items are lifted from a journal article without consultation with the original authors the instructions are not available to the translation team. Instructions can influence how the questionnaire is filled in (for example perhaps a time period is indicated, "the last 2 weeks," or an example is given that helps clarify how to fill in the questionnaire). Some questionnaires include a practice item, which may be particularly helpful for using questionnaires with children. Guillemin et al. (1993) recommend to carefully evaluate the translation of the instructions.

The response categories of the items should also be carefully translated so that the translated version is scaled the same way as the original version and/or adapted to the culture to avoid a skewed response style. This can prove to be especially complicated as subtle changes in wording of the response categories can have an important influence on the response characteristics of the item. Changes are sometimes made in the number of response categories and type of responses in translated versions of questionnaires. For example, in one of the experimental Dutch translations of the *Childhood Anxiety Sensitivity Index* (CASI; Silverman, Fleisig, Rabian, & Peterson, 2001), the three point scale of the original CASI (1 = *none*, 2 = *some*, and 3 = *a lot*) is adapted without explanation to a differently anchored 4-point scale: 1 = *never*, 2 = *sometimes*, 3 = *often,* and 4 = *always* (Muris, Schmidt, Merckelbach, & Schouten, 2001). Unfortunately when such a change is made without documentation, the reader does not know if this is due to simply not having translated the original scale correctly or if it is the result of careful statistical analyses. Van de Vijver and Tanzer (1997) give an example in which the rating scale needed to be adapted in a study with Hispanics in which respondents tended to choose extremes on a 5-point rating scale more frequently than White Americans. When a 10-point rating scale was used, this tendency was no longer found.

Another potential error found when changes are made in response categories is that there is no longer a fit between the type of response and the phrasing of the item. In any case, if changes are made to the response categories, such changes need to be reported, including the rationale for doing so, as they may be influential as to how the participants respond; if they are not done carefully, they could result in a method bias, which often leads to a shift in the average scores (Van de Vijver & Poortinga, 1997).

Back-Translation

Common Usage

The back-translation procedure has become the most commonly reported method used in evaluating the appropriateness of a translation. It is usually seen as the "Final step" of the translation procedure as a way to report back to the original author the content of the foreign version. We advocate to primarily use the back-translation procedure in an earlier phase of the translation process as an important feedback method for the translation team. It can be an extremely useful tool to identify problematic translations and in our experience often leads to additional discussion and adaptation of the items.

Back-Translator

A back-translation of the questionnaire is to be made by a person whose mother tongue is the

original language of the questionnaire. It is helpful if the back-translator is not only familiar with the culture of the original language of the questionnaire, but also has other knowledge relevant to the questionnaire subject (see section on qualities of the translation team). In contrast to this view, Guillemin et al. (1993) advocate that back-translators should preferably not be knowledgeable to the questionnaire subject, and thereby free of biases. If available, perhaps both types of back-translators can be useful in different ways for detecting problems in the translation.

Controversy Over Back-Translations

It should be pointed out that even though the back-translation procedure is often recommended, findings are mixed on the value of the procedure (Brislin et al., 1973; Perneger et al., 1999). Guillemin et al. (1993) highly advocate incorporating the making of back-translations and even recommend that all the initial translations made are back translated independently from each other. In contrast to Guillemin et al.'s enthusiasm for back-translations, other authors express criticism of the way back-translations are used. That is back-translations are mostly used with the goal of demonstrating to the original author that a literal translation was made. However the goal of absolute measurement equivalence is questionable and viewed as a source of cultural insensitivity by some (Erkut et al., 1999; Rogler, 1999). Hambleton (2001) further states that a back-translation actually provides little evidence of equivalence. In any case, if back-translations are rigidly used to assure that the translated instruments are kept exactly the same as the original version, the translated questionnaire is likely to suffer. Back-translations should reflect however that the content and meaning of the original items, instructions and response categories are the same as the original. Heavy reliance on back-translation as a sole method is highly discouraged (Brislin et al., 1973; Hambleton, 2001).

PRETEST OR PILOT DATA

After a first version of a translation is agreed upon, it is critical to administer it to a small group of persons that represent the target population. Hambleton and Patsula (1999) further recommend to pilot test an instrument and then compare the results to those obtained from the original language sample. Piloting is a very valuable step, but according to a review conducted by Guillemin et al. (1993) few studies on translated measures conduct a pilot or pretest phase. This is also true for the child and family literature. It is productive to include an interview or discussion with the pilot participants to find out their reactions to the instrument instructions, items and response categories. When selecting persons to pilot the measure with, we suggest varying age, gender and background/region of participants.

Through discussing the items of the questionnaire with pilot participants, there is much to be learned for example from children on how certain constructs are spoken about in their world and whether or not terms that are chosen are acceptable to them. In our research with Dutch children, we ask children if they understand the item and instructions; what they think of the question or item; what they think the question is asking. Sometimes the item is explained and then the child is asked to assist in explaining how he or she would express the idea in Dutch. It is important to consider regional differences in expression and to try to avoid slang when using the feedback of pilot participants. Brislin et al. (1973) describe a probe for the meaning of the participant's response: After a participant answers an item, he or she is asked, "What do you mean?" If the respondent's justification of their response is bizarre, then one should question if the intent of the question has been clearly conveyed. Finally, once initial pilot data have been collected, it is important to meet again as a team to consider any necessary changes. Adaptations can then be made on the basis of the pilot experience (Geisinger, 1994). Again, if the team is unsure about an item, temporarily two alternatives can be included in the version.

DATA COLLECTION AND ANALYSES

Empirical analyses of a dataset are an essential part of the translation and adaptation process (Hambleton, 2001). Even when a translating team takes all of the above steps and considerations, many problems can still remain unidentified until the measure is field-tested (Hambleton & Patsula, 1999). Many statistical procedures need to be applied because there is not one method without shortcomings (Hambleton, 2001). For data collection and analyses we recommend that methods advised in statistics books for questionnaire research are used (e.g., Nunnally & Bernstein, 1994). Geisinger (1994) recommends performing traditional or

item-response theory item analyses. We start with conducting a number of item analyses to evaluate response biases, such as frequency distributions and evaluation of skewness and kurtosis of items. A procedure called Differential Item Functioning (DIF) analyses, in which items that are more or less difficult than another item in the new version are compared to the source language version, can also be used (Geisinger, 1994). Using the chi-square test to evaluate if an item is biased is also recommended by Geisinger (1994).

Internal consistency of the (sub)scales can be analyzed next to examine if the individual items form a homogeneous set. Comparison of psychometric data on the translated questionnaire with data reported on the original questionnaire is important for examining differences in response styles that may be related to the translation of the instrument.

Factor analysis is the most frequently used technique, and more recently LISREL (Jöreskog & Sörbom, 1993) and structural equation programs (Bentler, 1992) are being used as well (Van de Vijver & Poortinga, 1997). The structure of the scale(s) should be examined to investigate transcultural factorial invariance. If a factor structure does not replicate, this may be attributed to the way items are translated or perhaps to cultural differences such as familiarity with the items or the influence of social desirability. If certain items load on a different factor, these too need to be re-evaluated. Incompatibility of the samples compared can result in a method bias as well (Van de Vijver & Tanzer, 1997). If samples are not similar it is important to consider that a sample bias may account for differences in outcome from the original questionnaire. Thus for comparison purposes, it is best to have similar samples in terms of, for example, age and psychopathology level of subjects. For example, sometimes factor structures differ for age groups or clinical versus normal groups.

In sum, there are a number of data analyses that can provide feedback on the choice of translation of the items. Hambleton (2001) notes however that methodological skills alone are insufficient to interpret statistical findings on a new translation and persons with knowledge of the cultures, languages and subject material (see translating team section) need to assist in interpretation.

CLINICAL UTILITY

Further psychometric evaluation that is recommended includes gathering new data to establish new norms, as well as assessing the stability of the instrument with test-retest reliability. Both are important for the clinical utility of a measure. Original language version norms are unlikely to be able to be used with the new language version of the instrument (Geisinger, 1994). The most obvious case is when an instrument is being used in a new culture and language that greatly differs from the population in which the source instrument was established. In a comparison of American and Dutch scores on the MMPI-2, a number of differences were found, such as the Dutch scoring higher on the neurotic and social introversion scales than Americans. Such differences justified creating new norms for the Dutch population (Derksen et al., 1993). Sometimes new norms also need to be established for use in a different country despite the fact that the measure is in the same language. For example research in Belgium and Holland on the Dutch translation and adaptation of the *Perceived Competence Scale for Children* (PCSC; Harter, 1982) revealed large differences between the two countries: Dutch children reported perceiving themselves as more competent in almost all areas measured by the PCSC (e.g., scholastic achievement, global self-worth) making it necessary to develop different norms for the two countries (Veerman, Straathof, Treffers, Van den Bergh, & ten Brink, 1997).

Finally, new norms may need to be established within a same language area if a specific group has not been included in the original norming of the instrument. For example, though the norming of the American CDI is extensive, only a small proportion of the samples are based on ethnic-minority children (Barreto & McManus, 1997). It is also current practice to sometimes use adult assessment tools with adolescents (e.g., SCL-90; Derogatis, 1977). In this scenario instruments may need to be adapted given the different life experiences of youth, but even if not, they still may need to be renormed for the younger population. Geisinger (1994) cites the MMPI as an example in which a separate adolescent version has been made. Collecting clinical data is also important for establishing clinical cut off scores in a new culture, as the cut off scores may differ as well. Thus, it is important to note that even if a measure proves to be reliable and valid, scores cannot necessarily be interpreted in the same way as in the culture of the original language.

SUMMARY AND CONCLUSION

Research across cultures is flourishing because of increased global communication and contact

among researchers. Cross-cultural data is also being regarded as increasingly important for our understanding of human behavior and psychological processes. Several persons in the child and adolescent field have commented on the importance of cross-cultural research for the field. Schneider (1998) makes the plea that cross-cultural research is necessary to test the limits or the generalizability of American based research on child and adolescent adjustment. Ideally it would be preferable for paradigms and theories to be constructed and tested in collaboration with researchers from different countries to get an intercultural viewpoint (Saarni, 1998; Schneider, 1998). Along these lines, Rubin (1998) further advises child development researchers not to generalize their (culture specific) theories of normal and abnormal development to other countries.

Before we can attribute differences in psychological processes and behavior to culture, however, we first need to examine the quality of a translated questionnaire and the norms established in a new culture. It is often the tendency of researchers to prematurely attribute differences in outcome from an original measure to cultural differences, but there are many sources of "error" along the translation path that could explain differences in outcome from the original measure. Within-culture differences also need to receive greater research attention (Geisinger, 1994; Rubin, 1998).

In addition, in order for journal readers and reviewers to interpret findings from other countries and/or cultures using the "same" instrument, it is essential that the translation and adaptation procedures, as well as the related analyses, are adequately carried out and reported. Any changes or adaptations made should ideally be documented (Hambleton, 2001). Furthermore, documentation of specific translation problems from one language and culture to another, such as problems with verb tense, certain words or concepts could be very helpful to other researchers (Hambleton, 2001). In any case, the current common practice to give a minimal report of procedures (i. e., "a translated version was used"; "a back-translation was conducted") needs to be more rigorous. Though in psychology journals there is generally a uniform way to present a method section, when presenting information on the translation and adaptation of an instrument there is no uniformity. It appears from reviewing articles on translated instruments in the child and family literature that there is a tendency to say a little bit more about the translation process if the instrument has been translated for countries in Asia or Africa (e.g., Weisz et al., 1987) compared to instruments adapted or translated for non-English European populations. When questionnaires are used in the same language but used with a population not included in the norm group, we found there is usually no reference to possible cultural adaptations made, especially when American instruments are used in non American English speaking countries (e.g., Heubeck, 2000 when referring to use of the CBCL in Australia).

No definite statements can be made based on the existing literature as to which are essential or optional steps for adapting assessment instruments into another language and culture given the lack of research on the guidelines suggested here and the limited research in previous reports in the literature. Thus, we do not have specific recommendations on what exactly should be conducted and documented in research reports. Research set up to evaluate guidelines with child and family assessment instruments is certainly needed and would be helpful in this regard. We do conclude, based on the guideline literature and our own experience, however, that when (a) care is taken in the translation/ adaptation process to include a team composed of different experts; (b) team members attempt to find a balance between a literal translation and a culturally specific translation; (c) back-translations are used in the process as a feedback method for the translation of content and intent (and not overly relied on as a sole method to establish equivalency); (d) field tests with small groups of persons are conducted prior to collecting larger datasets; (e) statistical analyses on a larger dataset are used as a method of identifying flawed items; (f) as well as for establishing reliability, validity and new norms; and (g) regular contact is kept with original source language authors—then a good translation is likely to result, and any large errors are likely to be eliminated.

REFERENCES

Achenbach, T. M. (1991). *Manual for the Child Behavior Checklist. 4-18 and 1991 profile*. Burlington: University of Vermont, Department of Psychiatry.

Anderson, R. T., Aaronson, N. K., Bullinger, M., & McBee, W. L. (1996). A review of the progress towards developing health-related quality-of-life instruments for international clinical studies and outcome research. *PharmacoEconomics*, *10,* 336–355.

Arrindell, W. A., & Ettema, J. H. M. (1986). *SCL 90: Handleiding bij een multidimensionaele psychopathologie-indicator.* [SCL-90: Manual for a multidimensional indicator of psychopathology]. Lisse, The Netherlands: Swets & Zeitlinger.

Barreto, S., & McManus, M. (1997). Casting the net for "depression" among ethnic minority children from the high-risk urban communities. *Clinical Psychology Review*, *17,* 823–845.

Bentler, P. M. (1992). *EQS Structural equation program manual*. Los Angeles: BMPD Statistical Software.

Bornstein, M. H., Haynes, O. M., Azuma, H., Galperín, C., Maital, S., Ogino, M., et al. (1998). A cross-national study of self-evaluations and attributions in parenting: Argentina, Belgium, France, Israel, Italy, Japan, and the United States. *Developmental Psychology*, *34,* 662–676.

Brislin, R. W., Lonner, W. J., & Thorndike, R. M. (1973). *Cross-cultural research methods*. New York: Wiley.

Canino, G., & Bravo, M. (1999). The translation and adaptation of diagnostic instruments for cross-cultural use. In D. Shaffer, C. P. Lucas, & J. E. Richters (Eds.), *Diagnostic assessment in child and adolescent psychopathology* (pp. 285–298). New York: Guilford Press.

Derksen, J. J. L., de Mey, H. R. A., Sloore, H., & Hellenbosch, G. (1993). *MMPI-2 Handleiding bij afname, scoring, en interpretatie*. [MMPI-2: Manual for administration, scoring and interpretation]. Nijmegen, the Netherlands: Pen Tests Publisher.

Derogatis, L. R. (1977). SCL-90. *Administration, scoring and procedures manual-I for the R(evised) version and other instruments of the psychopathology rating scales series.* Baltimore: Clinical Psychometrics Research Unit, John Hopkins University School of Medicine.

Erkut, S., Alarcon, O., Garcia Coll, C., Tropp, L. R., & Garcia, H. A. V. (1999). The dual-focus approach to creating bilingual measures. *Journal of Cross-Cultural Psychology*, *30,* 206–218.

Garcia Coll, C., Akerman, A., & Cicchetti, D. (2000). Cultural influences on developmental processes and outcomes: Implications for the study of development and psychopathology. *Developmental Psychopathology*, *12,* 333–356.

Geisinger, K. F. (1994). Cross-cultural normative assessment: Translation and adaptation issues influencing the normative interpretation of assessment instruments. *Psychological Assessment*, *6,* 304–312.

Guillemin, F., Bombardier, C., & Beaton, D. (1993). Cross-cultural adaptation of health-related quality of life measures: Literature review and proposed guidelines. *Journal of Clinical Epidemiology*, *46,* 1417–1432.

Hambleton, R. (2001). The next generation of the ITC Test translation and adaptation guidelines. *European Journal of Psychological Assessment*, *17,* 164–172.

Hambleton, R. K., & Patsula, L. (1999). Increasing the validity of adapted tests: Myths to be avoided and guidelines for improving test adaptation practices. *Journal of Applied Testing Technology*, *1,* 1–30.

Harter, S. (1982). The Perceived Competence Scale for Children. *Child Development*, *53,* 87–97.

Harter, S. (1988). *Manual for the Self-Perception Profile for Adolescents*. Denver, CO: University of Denver.

Herdman, M., Fox-Rushby, J., & Badia, X. (1997). 'Equivalence' and the translation and adaptation of health-related quality of life questionnaires. *Quality of Life Research*, *6,* 237–247.

Heubeck, B. G. (2000). Cross-cultural generalizability of CBCL syndromes across three continents: From the USA and Holland to Australia. *Journal of Abnormal Child Psychology*, *28,* 439–450.

Jeanrie, C., & Bertand, R. (1999). Translating Tests with the International Test Commission's Guidelines: Keeping validity in mind. *European Journal of Psychological Assessment*, *15,* 277–283.

Jöreskog, K. G., & Sörbom, D. (1993). *LISREL 8*. Chicago: Scientific Software International.

Kovacs, M. (1992). *Children's Depression Inventory (CDI) manual*. New York: Multi-Health Systems.

March, J. S., Parker, J. D., Sullivan, K., Stallings, P., & Conners, K. (1997). The Multidimensional Anxiety Scale for Children (MASC): Factor Structure, Reliability and Validity. *Journal of the American Academy of Child Psychiatry*, *36,* 554–565.

Muris, P., Schmidt, H., Merckelbach, H., & Schouten, E. (2001). Anxiety sensitivity in adolescents: factor structure and relationships to trait anxiety and symptoms of anxiety disorders and depression. *Behaviour Research and Therapy*, *39,* 89–100.

Nunnally, J. C., & Bernstein, I. H. (1994). *Psychometric theory* (3rd ed.). New York: McGraw-Hill.

Ollendick, T. H. (1983). Reliability and validity of the Revised Fear Survey Schedule for Children (FSSC-R). *Behaviour Research and Therapy*, *21,* 685–692.

Ollendick, T. H., King, N. J., & Frary, R. B. (1989). Fears in children and adolescents: Reliability and generalizability across gender, age, and nationality. *Behaviour Research and Therapy*, *27,* 19–26.

Ollendick, T. H., Yule, W., & Ollier, K. (1991). Fears in British children and their relations to manifest anxiety and depression. *Journal of Child Psychology and Psychiatry*, *32,* 321–331.

Parke, R. (2000). Beyond white and middle class: Cultural variations in families- assessments, processes, and policies. *Journal of Family Psychology*, *14,* 331–333.

Perneger, T. V., Leplege, A., & Etter, J. F. (1999). Cross-cultural adaptation of a psychometric instrument: Two methods compared. *Journal of Clinical Epidemiology*, *52,* 1037–1046.

Perris, C. Jacobsson, L., Lindstrom, H., von Knorring, L., & Perris, H. (1980). Development of a new inventory for assessing memories of parental rearing behavior. *Acta Psychiatrica Scandinavica*, *61,* 265–274.

Rogler, L. H. (1999). Methodological sources of cultural insensitivity in mental health research. *American Psychologist*, *54,* 424–433.

Ronan, K., Kendall, P. C., & Rowe, M. (1994). Negative affectivity in children: Development and validation of a self-statement questionnaire. *Cognitive Therapy and Research*, *18,* 509–528.

Rubin, K. H. (1998). Social and emotional development from a cultural perspective. *Developmental Psychology*, *34,* 611–615.

Saarni, C. (1998). Issues of cultural meaningfulness in emotional development. *Developmental Psychology*, *34,* 647–652.

Schneider, B. H. (1998). Cross-cultural comparison as doorkeeper in research on the social and emotional adjustment of children and adolescents. *Developmental Psychology*, *34,* 793–797.

Siebelink, B., & Treffers, P. D. A. (2001). *Dutch version of the Anxiety Interview Schedule for children for DSM-IV, child and parent versions by W. K. Silverman & A. M. Albano*. Lisse, the Netherlands: Swets & Zeitlinger.

Silverman, W., & Albano, A. M. (1996). *The Anxiety Interview Schedule for children for DSM-IV, child and parent versions.* San Antonio, TX: Psychological Corporation.

Silverman, W. K., Fleisig, W., Rabian, B., & Peterson, R. A. (2001). Childhood anxiety sensitivity index. *Journal of Clinical Child Psychology*, *20,* 162–168.

Sprangers, M. A. G., Cull, A., Bjordal, K., Groenvold, M., & Aaronson, N. K. (1993). The European Organization for Research and Treatment of Cancer approach to quality of life assessment: Guidelines for developing questionnaire modules. *Quality of Life Research*, *2,* 287–295.

Sudman, S., & Bradburn, N. M. (1982). *Asking questions*. San Francisco, CA: Jossey-Bass.

Treffers, P. D. A., Goedhart, A. W., Veerman, J. W., van den Bergh, B. R. H., Ackaert, L., & de Rycke, L. (2002). *Competentiebelevingsschaal voor adolescenten* (CBSA) [Dutch adaptation of the Self-Perception Profile for Adolescents]. Lisse, the Netherlands: Swets & Zeitlinger.

Van de Vijver, F., & Tanzer, N. K. (1997). Bias and equivalence in cross-cultural assessment: An overview. *European Review of Applied Psychology*, *47,* 263–279.

Van de Vijver, F. J. R., & Poortinga, Y. H. (1997). Towards an integrated analysis of bias in cross-cultural assessment.

European Journal of Psychological Assessment, *13,* 29–37.

Veerman, J. W., Straathof, M. A. E., Treffers, P. D. A., Van den Bergh, B. R. H., & ten Brink, L. T. (1997). *Competentiebelevingsschaal voor kinderen* (CBSK) [Dutch adaptation of the Perceived Competence Scale for Children]. Lisse, the Netherlands: Swets & Zeitlinger.

Wang, Y., & Ollendick, T. H. (2001). A cross-cultural and developmental analysis of self-esteem in Chinese and western children. *Clinical Child and Family Psychology Review*, *4,* 253–271.

Weisz, J. R., Suwanlert, S., Chaiyasit, W., Weiss, B., Achenbach, T. M., & Walter, B. A. (1987). Epidemiology of behavioral and emotional problems among Thai and American children: parent reports for ages 6–11. *Journal of the American Academy of Child Psychiatry*, *26,* 890–898.

Werner, O., & Campbell, D. T. (1970). Translating, working through interpreters, and the problem of decentering. In R. Naroll & R. Cohen (Eds.), *A handbook of method in cultural anthropology* (pp. 398–420). New York: American Museum of Natural History.