Chapter 9

Best Practices in Translating and Adapting DMQ 18 to Other Languages and Cultures

Fajrianthi, Jun Wang, Stephen Amukune, Marcela Calchei and George A. Morgan

Introduction

This chapter provides guidelines developed by the International Test Commission: *The ITC Guidelines for Translating and Adapting Tests* (ITC, 2017). We recommend that future researchers who want to translate and use DMQ 18 follow these guidelines to both translate the questionnaire and assess its cultural appropriateness in their language and culture; and also provide evidence for the reliability and validity of resulting data. In addition, the chapter is an overview about validity issues and biases in regard to making such adaptation. It also provides a step-by-step approach to what we believe are best practices for doing adaptation, using examples based on a proposed translation from English into a Southeast Asian language. Furthermore, the chapter provides detailed examples of how to provide evidence for the reliability and validity of hypothetical data from the use of the

DMQ in this Southeast Asian culture. The examples utilized in this chapter are based on a previous research conducted by Rahmawati et al. (2020).

Translation of psychological questionnaires developed and normed in other countries is a common practice. For example, in the research literature there are other translations and adaptations of the Dimension of Mastery Questionnaire (DMQ). Using the decentering procedure (Marín & Marín, 1980), DMQ 18 was developed in English, Chinese, and Hungarian for children 6 months to 19 years (see **Chapter 2**). Research using translations of DMQ 18 has been published in Turkish (Özbey & Daglioglu, 2017), Persian/Farsi (Salavati et al., 2018), and Bangla (Shaoli et al., 2019). In **Chapter 3**, there are several tables showing the characteristics of samples from other more recent translations and adaptations.

Reasons for and Cautions about Adapting Tests and Questionnaires

When adaptation are made, rigorous assessment of the equivalence of the original and adapted versions of the questionnaire is essential. There are many good reasons and considerable advantages for adapting a questionnaire. Hambleton and Patsula (1999) identified at least five reasons found in the literature for adapting tests or questionnaires:

1. It is usually cheaper and faster to adapt a questionnaire, compared to developing a new one in a second language.
2. Adapting a questionnaire is the most effective method in producing an equivalent questionnaire in a second language, when the purpose is cross-cultural or cross-national assessment (for example: credentialing exams).
3. Developing a new questionnaire in a second language demands expertise which may be lacking.
4. An adapted questionnaire of an already well-known questionnaire offers a greater sense of security, compared to developing a new questionnaire.
5. Providing multiple language versions of a questionnaire offers more fairness to examinees.

By adapting a questionnaire, in particular when adaptation is used for cross-cultural studies, the major issue is obtaining tests for cross-cultural populations that produce valid and comparable results, so that the researcher is able to compare data from cross-lingual populations. This enables greater fairness in the evaluation because the same instrument assesses the construct based on the same theoretical and methodological perspectives. The use of adapted instruments naturally enables a greater ability to

generalize and also enables one to investigate differences within cross-lingual populations (Borsa et al., 2012; Hambleton, 2005).

Adaptation processes aim to yield instruments that are equivalent across different cultures (Hambleton, 2005). Unfortunately, in practice the questionnaire adaptation process is often viewed as a simple task that can be completed by anyone who knows the target languages. Researchers have incorrectly assumed that finding a good translator would be sufficient for obtaining equivalent cross-linguistic or cross-cultural questionnaires and surveys. Failing to follow-up the translation process by providing a compilation of empirical evidence, which supports the intended uses of the questionnaire scores in its target languages and cultures is a fundamental mistake in the practice of test adaptation (Rios & Hambleton, 2016).

Common Issues Related to Test Adaptation

Test and questionnaire adaptation is a scientific and professional activity that refers to the development of a derived questionnaire; the adapted questionnaire is obtained by transferring the original questionnaire from its source language or culture to a target language or culture. The adaptation process should offer proof of the psychometric appropriateness and similarity ("equivalence") of the adapted questionnaire, in the new language and culture, to the original questionnaire (Greiff & Iliescu, 2017).

"Equivalence" (or "invariance") refers to score compatibility obtained from the administration of the versions of a questionnaire (original vs. adapted), and is considered to be a specific source of validity. One version of questionnaire being equivalent to another has two important implications. First, the scores of the two versions are directly comparable. Second, evidence generated by a version is also valid for the other version, as validity evidence is transferable.

"Equivalence" and "bias" are closely connected. "Bias" is associated with errors, often used as an expression of "non-equivalence". When the original and adapted versions of a questionnaire are not equivalent, responses collected using the two versions cannot be directly compared, and conclusions based on evidence from the original version cannot be advanced for scores from the adapted version (Greiff & Iliescu, 2017; Rios & Hambleton, 2016). van de Vijver and his colleagues identified three potential sources of measurement bias in cross-cultural assessment: (1) construct bias, (2) method bias, and (3) item bias (van de Vijver & Hambleton, 1996; van de Vijver & Leung, 1997; van de Vijver & Poortinga, 2005). Construct bias occurs because of differences in conceptual definitions or in behaviors that are deemed indicative of the construct. Methodological-procedural bias happens when the assessment procedure causes unfavorable difference between groups. Item content bias can take place because of poor translation

or use of items that are not suitable in a particular cultural context (Byrne & Watkins, 2003; van de Vijver & Tanzer, 2004).

For example, the construct of happiness does not have the same meaning across cultural groups (He & van de Vijver,2012). In European American culture (Western culture), with the positive hedonic experience at its core, happiness is imagined to be infinite, attainable, in principle, for everybody if sought. In the United States, then, there is a widespread belief that happiness is an end result of personal pursuit, which in turn is grounded in personal goals and aspirations. In contrast, in Southeast Asian cultural (Eastern culture) contexts, there is a contrasting view of the self as interdependent. Within this interdependent, highly relational model of self, happiness is also likely to take one particular form, wherein interpersonal and social aspects of happiness receive a much greater emphasis (Uchida & Kitayama, 2009). As a result, the tests implemented to measure happiness in Western culture do not capture the same underlying dimensions of the construct in Eastern cultures. This has two implications: the validity of the measurement is lacking, and direct comparisons between samples cannot be made. There are validity concerns whenever an instrument developed in one language and culture is translated and used in another language and culture.

Questionnaire Adaptation and Instrument Validity

In adapting a questionnaire, issues about the validity of the translated instrument must be considered and dealt with. Validity is a theoretical concept that has evolved considerably over time. In modern validity theory, it is often referred to as a *unitary validity framework*. Validity is an 'integrated evaluative judgement of the degree to which empirical evidence and theoretical rationales support the adequacy and appropriateness of inferences and actions based on test scores' (Wolming & Wikstrom, 2010). The search for evidence of an instrument's validity is subdivided into two main areas: the instrument validation for the new context and validation for cross-cultural studies (Borsa et al., 2012).

Instrument Validation for the New Language and Context

Instrument validation begins by evaluating the factorial structure. Instruments are generally designed to measure multifaceted constructs, so instruments should have a relatively organized factor structure, even when latent (Borsa et al., 2012). For example, factorial validation for the Dimensions of Mastery Questionnaire (DMQ) has focused on of five main dimensions or factors: Cognitive/Object Persistence (COP), Gross Motor Persistence (GMP), Social Persistence with Adults (SPA), and Social Persistence with

Children (SPC), and Mastery Pleasure (MP). The DMQ is considered the most widely used instrument for measuring mastery motivation; it assesses those five dimensions, plus Negative Reactions to Challenge (NRC) and General Competence (COM). The competence scale is a quick way to assess a child's ability and, thus, is not considered a measure of mastery motivation. Factorial structures that are relatively similar to the original proposal are expected in DMQ validation studies for use in new contexts. Otherwise, discrepancies will affect the understanding of the evaluated construct. Possible changes, which occur in validation studies in light of quantitative and qualitative discrepancies, should be discussed. By doing so, researchers can identify possible reasons for changes in the questionnaire's factorial structure. Certain changes are to be expected, especially in complex questionnaires with high number of items and factors, as a result of sampling characteristics. The techniques of confirmatory factor analysis (CFA) should be used to assist the researcher in their choice of a factorial structure that is most plausible for the sample. Evaluating the factorial structure of the instrument is only one aspect of a validation study. Other evidences of validity are to be collected, including the evaluation of the instrument's content and criterion validity through comparing its results with those of equivalent measures. See **Chapter 5** for discussion of the several types of evidence for evaluating the validity of the Dimensions of Mastery Questionnaire. The analysis of internal consistency among items (i.e. internal consistency reliability) is often also a part of the evaluation process. See **Chapter 4** for a discussion of the several types of evidence for evaluating the reliability of the DMQ.

Instrument Validation for Cross-Cultural Studies

Researchers must simultaneously assess the compatibility of a measure within the various groups when conducting cross-cultural studies (Hambleton & Patsula, 1998; Sireci, 2005). Through comparative analyses, researchers ensure that the same construct in different populations is similarly evaluated, ensuring the assumption of measurement invariance (Reise et al., 1993). Multi-Group Confirmatory Factor Analysis (MGCFA), Differential Item Functioning (DIF) proposed by the Item Response Theory (IRT), and Multidimensional Scaling (MDS) can be considered as valuable ways of assessing measurement invariance (Rios & Hambleton, 2016; Sireci, 2005). The validity of the assumption of factorial invariance between groups is paramount for psychometric instrument development and adaptation, and also for group comparisons in cross-cultural studies. Unless thoroughly tested, researchers cannot claim that an instrument has similar structures and parameters in different populations. If the instrument measurements are not comparable between different groups, any differences in group scores or correlation patterns with external variables tend to be measurement errors,

not reflecting the actual differences between groups (Tanzer, 2005). See **Chapter 2** for discussion of the measurement invariance of DMQ 17 in preschool children whose parents spoke Chinese, Hungarian, and English (Hwang et al., 2017) and also discussion of the measurement invariance for self-reports by school-age children in China, Hungary, in the US (Wang et al., 2014).

ITC Guidelines for Translating and Adapting Tests

In order to avoid common translation biases, the International Test Commission (ITC, 2017) developed guidelines for test adaptation. These guidelines were organized into six categories: (1) pre-condition, (2) test development, (3) confirmation, (4) administration, (5) score interpretation, and (6) documentation. This section summarizes 10 specific ITC guidelines from the first three categories. The description here is based on the current version (2.4) of the second edition of the ITC Guidelines, published on the International Test Commission website in 2017. Researchers should endeavor to use the most recent editions when they become available. After the description of these 10 guidelines, we will provide a hypothetical example of the process for translation and adaptation of DMQ 18 into a Southeast Asian language and culture.

Pre-Condition (PC) Guidelines

PC-1 (Guideline 1) Request Permission

Obtain the necessary permission from the holder of the intellectual property rights relating to the test before carrying out any adaptation.

Intellectual property rights refer to a set of rights people have over their creations, inventions, or products, to protect the interests of creators by providing moral and economic rights over their creation. An agreement from the intellectual property owner should be obtained before starting test adaptation. The agreement should specify the modifications which are acceptable regarding original test characteristics and the property rights of the developer of the adapted version.

PC-2 (Guideline 2) Evaluate Overlap

Evaluate whether the amount of overlap in the construct's definition and content measured by the test is sufficient for the intended use(s) in the population of interest.

The items assessed should be understood in the same way in both the source or original language and in the new or target language and cultural group into which it is being translated. This is the foundation of valid cross-cultural comparisons. In this stage, the test or questionnaire has not been adapted, so it is good to compile previous empirical evidence with similar

tests and make judgments of the suitability of the construct, including the item content, in the new language.

In order to make valid interpretations of scores, the scope of the test has to be described thoroughly. To do so requires an adequate working definition of the construct to be measured. Psychologists and other knowledgeable persons in the new culture should determine if the construct exists and if the same definition applies equally well in both language and cultural groups. Persons with expertise about the construct and about the cultural group should be recruited to evaluate the legitimacy of the measured construct in each cultural/linguistic group, and to answer the question as to whether the construct makes sense in both cultures. Focus groups, interviews, and surveys can be utilized to obtain structured information regarding the degree of construct overlap.

The goal of any analyses is to confirm the equivalence of the structure of the test across the two languages; e.g., English vs. Southeast Asian in the example in the next section. This process is conducted to avoid construct bias, which occurs when the studied constructs are non-equivalent across language or cultural groups. Non-equivalence can occur when there is partial overlap in conceptualizing the construct or when the behaviors associated with the construct manifest themselves differentially across cultures (van de Vijver & Hambleton, 1996). As a result, the tests do not capture the same underlying dimensions of the construct across groups; there are two implications: validity of the measurement is lacking, and direct comparisons between samples cannot be made.

Construct bias has two main sources:

Source 1: Differential construct manifestation. Bias could result from the fact that although the construct does exist in both cultures, there are differences in how it is defined and exhibited (Byrne & Watkins, 2003).

Source 2: Construct under-representation. This is characterized by insufficient sampling of the behaviors describing the construct (Messick, 1995); this is similar to the concept of content validity in classical test theory. The test should be fully representative of the construct (Kline, 1993). Construct under-representation means that it does not cover all the essential dimensions and facets of the construct (Messick, 1995). A construct is under-represented when the original test is too short to provide valid deductions or the items are too poorly written for the reader to comprehend the intended construct (Downing, 2002). As with the first source, if the construct is not fully investigated in the target culture, the items from the original version may not be inclusive of the behaviors defining the construct in the target culture.

PC-3 (Guideline 3) Minimize Irrelevant Differences

Minimize the influence of any cultural and linguistic differences that are irrelevant to the intended uses of the test in the population of interest.

This guideline relies mainly on qualitative methods and experts familiar with the research on specific cultural and language differences. For a questionnaire measure like DMQ 18, special emphasis is placed on the selection of content experts and translators, who are native to the target language and culture; knowing the target language is insufficient for identifying possible sources of method bias. The guidelines clearly suggest that a well-designed translation procedure should emphasize conceptual similarity instead of literal similarity of the translation as a necessary step toward a valid adaptation. Consequently, the use of systematic procedures by experts is necessary to complement the use of statistical analyses. The choice of translators and development of the translation procedures are also critical to meet the ITC guidelines concerning test development, so they are described in greater depth in the next section.

Test Development (TD) Guidelines

TD-1 (Guideline 4) Choose Experts for the Translation

Ensure that the translation and adaptation processes consider linguistic, psychological, and cultural differences in the intended populations by choosing experts with relevant expertise.

It is important to use at least two translators; the ITC guidelines note that the older practice of using a single translator, however qualified, is no longer considered acceptable. Expertise in the target culture results from using translators native in the target language who are also living in the target locale, with the former being essential and the latter highly desirable. "Expert" is a person or a team with sufficient combined knowledge of: (1) the languages involved, (2) the cultures involved, (3) the content of the (original) test, and (4) general principles of testing. These are paramount to produce a professional quality translation/adaptation. In practice it may be effective to use teams of people with different qualifications (e.g., translators with and without expertise in the specific subject, etc.) in order to identify areas that may be overlooked (rather than just relying on a single expert). It is also desirable to provide training for translators in item writing principles for the formats utilized.

TD-2 (Guideline 5) Translation

Use appropriate translation designs and procedures in order to maximize the suitability of the adaptation for the intended populations.

This guideline requires that decisions made by translators maximize the adapted version's suitability for the intended population, meaning that the language should feel natural and acceptable, focusing more on functional rather than literal equivalence. Popular designs to achieve these goals are forward and backward translations. Brislin (1986) and Hambleton and Patsula (1999) provide full discussions of the two designs, including definitions, strengths, and weaknesses.

Two (or more) translation and a reconciliation procedure aim to address the shortcomings and risks of relying on the idiosyncrasies resulting from a single translation. A third independent translator or expert panel could then identify and resolve the discrepancies between alternative forward translations, resulting in a single version to be utilized.

TD-3 (Guideline 6) Evidence for Equivalence

Provide evidence that the test instructions and item content have similar meaning for the intended populations.

The evidence required by the guideline can be collected using various strategies. For example, the strategies recommended by van de Vijver and Tanzer (1997) included: (1) using reviewers native to local culture and language to evaluate the translation; (2) using samples of bilingual respondents to provide suggestions about the equivalence of instructions and items; (3) using local surveys to evaluate the test and interview the administrators and respondents post-administration for feedback; and (4) using adapted test administration procedures to increase acceptability and validity, when following the original instructions would make less sense or be misunderstood by respondents of the target language/culture group. Trying out the translation on a small scale can be valuable.

TD-4 (Guideline 7) Appropriateness of the Procedure

Provide evidence that the item formats, rating scales, scoring categories, and modes of administration are suitable for the intended population.

Researchers should ensure that respondents are familiar with any novel item formats or test administration procedures in the testing process. Qualitative and quantitative evidence both have a role in assessing this guideline. Several features of an adapted test may be checked, such as the reading level required for respondents to provide valid responses.

TD-5 (Guideline 8) Pilot Data

Collect pilot data on the adapted test version to enable item analysis, reliability assessment, and small-scale validity studies so that any necessary revisions can be made.

It is important to have confirming evidence regarding the psychometric qualities of the adapted test before conducting large-scale studies of reliability, validity, and/or norming, which are usually time-consuming and expensive. There are many psychometric analyses (such as coefficient alpha to examine the internal consistency of the scales) that could be carried out to provide initial evidence of score reliability and validity.

Confirmation (C) Guidelines

The Confirmation Guidelines are those that are based on empirical analyses of full-scale validity studies.

C-1 (Guideline 9) Sample Selection

Select samples with characteristics for the intended use of the test and of sufficient size and relevance for empirical analyses.

The data collection design refers to the way data are collected to establish norms (if needed), to check the equivalence among the language versions of the test, and to conduct validity, reliability, and DIF studies. The first requirement is that samples should be sufficiently large to allow for stable statistical information. The ITC guidelines provide two suggestions regarding the sample. First, to investigate the factorial structure of a test, a sample size of 300 or above is considered sufficient (Wolf et al., 2013). Second, the sample should be as representative of the intended population as possible.

C-2 (Guideline 10) Empirical Analysis

Provide relevant statistical evidence regarding the construct, method, and item equivalence for all intended populations.

Establishing the construct equivalence of the original and target language versions of a test is important, though not the only important empirical analysis to conduct. Approaches for construct equivalence (PC-2) and method equivalence (PC-3) were addressed briefly earlier in the ITC guidelines.

This guideline requires researchers to address construct equivalence empirically. There are at least four statistical approaches for assessing construct equivalence across source and target language versions of a test: Exploratory Factor Analysis (EFA), Confirmatory Factor Analysis (CFA), Multidimensional Scaling (MDS), and comparison of nomological networks (Sireci et al., 2005). Researchers are expected to identify any possible sources of method bias in the adapted test. Sources of method bias include: (1) different levels of test motivation in participants, (2) differential experience on the part of respondents with psychological tests, (3) a longer duration needed to take the test in one language group than the other, (4) differential familiarity with the response format across language groups, and (5) heterogeneity of response style, etc. Item equivalence can be analysed with, for example, CFA and IRT approaches to the identification of potentially biased test items.

An Example of the Adaptation and Evaluation Process for DMQ 18

This section provides a detailed example of the process that we used to create a hypothetical Southeast Asian version, from the original English DMQ 18, and to test its reliability and validity. The example is based on the ITC guidelines with a few additions. The sequence of steps in the example used the Precondition (PC) ITC guidelines for Steps 1-3, the Test Development (TD) for Steps 4, 5, 7, 10 and 11, and The Confirmation (C) ITC guidelines for Steps 12 and 13. Steps 6 and 9 are additional steps that we recommend, which involve consulting with the test developers to be sure that the back translation and any later revisions fit with the original conceptualization of the items. Step 8 provided a recommended method to check the content validity of the instrument.

Step 1: PC-1 Request Permission

The process of obtaining the necessary permission to adapt DMQ 18 was conducted through e-mail addressed to Professors George Morgan and Krisztián Józsa, the developers of DMQ 18. By doing so and receiving a reply that permission was granted, the researchers were ready to start the translation and adaptation process into the new language version.

Step 2: PC-2 Evaluate Overlap

In the adaptation process of DMQ 18, the researchers from Southeast Asia collaborated with three content experts in early childhood psychology and education (the focus of the preschool DMQ 18, which was being considered for use in the planned Southeast Asia studies) to conduct a literature review of the concept of mastery motivation in early childhood. The review was also conducted for similar concepts about general ability and the competencies required for children as they progress through developmental tasks. Based on the review, it was agreed that the DMQ items overlap sufficiently with the concept of mastery motivation in the intended Southeast Asian preschool population.

The DMQ was developed and refined since the 1980s by a team of researchers, including Morgan, Busch-Rossnagel, Harmon, and Jennings (Morgan et al., 1983; Morgan et al., 1993). DMQ 18 uses five-point Likert scales, ranging from 1 (*completely unlike this child*) to 5 (*exactly like this child*). Higher scores indicate higher mastery motivation in a child. Each of the five dimensions utilized in this example consisted of five items.

The next step was to review the construct of mastery motivation with experts, utilizing several questions:

1. Does the particular construct to be measured exist in both cultures?
2. Is it logical to compare the two cultures in regards to the particular construct?

3. Would cross-cultural comparison on the particular construct be meaningful?
4. Does the particular construct to be measured have the same meaning in the compared cultures?

Based on the analysis of the construct, it was found that the definition and scope (or operational definition) of mastery motivation are similar in the Southeast Asian culture and in the culture of where DMQ 18 was originally developed, indicating sufficient overlap of the constructs in the two cultures.

Step 3: PC-3 Minimize Irrelevant Differences

In the adaptation process of DMQ 18, it was important that the Southeast Asia content experts were not just natives and proficient in both languages and cultures (English and the Southeast Asian), but they also had educational backgrounds in early childhood development.

Qualitative methods, including interviews, were conducted. Discussions focused on item clarity, test instructions, and the rating scales. The goal was to develop procedures appropriate for the intended population and to minimize potential problems due to cultural differences. Standardized procedures were designed to administer DMQ 18 under consistent procedures so that the test-taking experience would be as similar as possible across examinees and cultures. Feedback from the discussions noted issues of item clarity and revised the instructions about how to respond to the rating scale.

Step 4: TD-1 Choose Experts for the Translation

In the DMQ 18 adaptation process for the Southeast Asian culture, four experts were selected as translators. All of them were considered functionally bilingual; all were able to conduct professional activities in both languages and had an academic background in psychology or education. They were not all equally fluent in both languages, but all met the "functionally bilingual" condition. All were given written information concerning the kind of translation that was expected from them as well as instructions on how to write test items. The four translators were selected because they were considered content experts; their academic backgrounds were closely related to psychology and child development.

The List of Qualifications for the Expert Translators, including their highest degree:

1. Expert/translator 1: Doctorate in Psychology. Teaches Child Education and Developmental Psychology and is familiar with research on mastery motivation and the DMQ.
2. Expert/translator 2: Master's in Psychology as a Profession. Teaches Child Education and Developmental Psychology. Head of Foundation for Childhood Education. Somewhat familiar with the concept of mastery motivation.

3. Expert/translator 3: Master's in Psychology as Profession. Teaches Child Education and Developmental Psychology. Consultant of Foundation for Childhood Education. Not familiar with the concept of mastery motivation.
4. Expert/translator 4: Doctorate in Education. University department head for preschool education. Not familiar with the concept of mastery motivation.

Step 5: TD-2 Translation

In the process of translating and adapting DMQ 18 to the Southeastern Asian language, both a forward and then a backward translation of the DMQ 18 items were used. Two translators were used for the forward translation, and two different translators were used for the back translation. The two forward translators were not only considered technical experts, but also somewhat knowledgeable about the concept of mastery motivation and its measurement. The two backward translators were not knowledgeable about the concept of mastery motivation, but were generally knowledgeable about child development, as noted above in Step 4. All were given written information about the meaning and the use of the rating scales that they were asked to assess the equivalence of the translated items. The two forward translations were synthesized by consensus. Likewise, a synthesis of the backwards items was done. Table 9.1 shows an example of the original English version of two DMQ 18 items with their forward translation in the Southeastern Asian language and backward translations in English.

Table 9.1. Comparison of an Example of DMQ 18 Items from the Original with the Forward and Backward Translations

Item No.	Original version	Forward translation version	Backward translation Version
4	*Tries to do things to keep children interested*	Berusaha melakukan sesuatu agar anak-anak lain tetap tertarik	*Trying to do something so that other children remain interested*
5	*Tries to keep adults interested in talking*	Berusaha agar orang dewasa tetap tertarik dalam pembicaraan.	*Trying to keep adults interested in the conversation.*

Step 6: Consult Original Developer

Although this step is not explicit in the ITC guidelines, we think that it is highly desirable to have the original developer review the back translation to be sure that the items are consistent with their intended meaning related to the concept of mastery motivation. If some items do not reflect the original meaning adequately, suggestions would be made to have the translator use different terms in the forward translation.

Step 7: TD-3 Gather Evidence for Equivalence

of the original English DMQ items and the back translation synthesis. In order to avoid randomness and mere subjectivity in the evaluation of the translated items, the three content experts in early childhood psychology and education (see Steps 2 and 3) were now asked to rate each item using a systematic method developed by the research team. Based on the various definitions of equivalence proposed over the years, our method focuses on the *linguistic* and also the *conceptual* equivalence (Jeanrie & Bertrand, 1999).

The three content experts were first asked to rate the comparability and similarity between original items and the synthesis of the perhaps somewhat revised backward translation (Jeanrie & Bertrand, 1999). **Comparability** refers to the degree of formal *linguistic equivalence* in language, phrases, terms, words, and sentences. To assess *conceptual equivalence*, the experts were asked to rate **similarity**, which concerns the degree to which the two versions of an item are *semantically* similar, having the same meaning despite the use of perhaps somewhat different terminology. The expert review form (shown in Table 9.2) is a rating scale, with a range of 1 to 4. Items with identical meaning were given a score of 4, while those with a very different meaning were assigned a score of 1.

Table 9.2. A Form to Rate the Linguistic Comparability and Conceptual Similarity of the Original DMQ 18 Items with the Back Translation Items

No.	Original item	BT synthesis item	Comparability				Similarity			
			1	2	3	4	1	2	3	4
1.	Tries to figure out what adults like	Trying to find out what adults like								
2.	Tries to understand other children	Trying to understand other children								
Etc.										

** BT= Back Translation.*

We used the criteria suggested by Polit et al. (2007) to evaluate the ratings for Step 7 (evidence for equivalence) and Step 8 (evidence for content validity). That is, relatively good items were those with a rating of 3 or 4,

while relatively poor items were rated 1 and 2. To evaluate equivalence, ratings were divided into a dichotomous score: 1 (for items with scores of 3 and 4) and 0 (for items with scores of 1 and 2). The linguistic and semantic equivalence of each item was estimated by summing up the dichotomous scores for comparability and similarity, respectively, and then dividing them by the number of reviewers. Polit et al. (2007) suggested a cut-off of 0.78. as evidence that the new items shared adequate linguistic or semantic characteristics with the original DMQ item. If no item was below the cut-off of 0.78, there were only marginal linguistic and semantic differences between items of the original scale and those of the adapted version, regardless of minor differences in the terminologies used. This type of equivalence was rated by three experts, and all of the DMQ 18 items in this example were above the cutoff score of 0.78.

Step 8: Gather Content Validity Evidence

We think that it's important to have the expert reviewers rate the original and translated items for content validity, so we have added this step to our example. The content validity of the items within the cultural context of the new language was rated for relevance, importance, and clarity. Content validity assessment was carried out on both the backward and forward translations. Sireci and Faulkner-Bond (2014) state that content validity (using the Content Validity Index, CVI) refers to the degree to which the content of a test is relevant to the measurement objective. The CVI of each item was calculated by asking the three content expert reviewers to rate each item, from 1 to 4, in terms of its: *relevance* (the extent to which the item measures a relevant dimension of the construct of mastery motivation), *importance* (the extent to which the item is critical for a dimension of the construct of mastery motivation), and *clarity* (the degree of clarity and understandability of the item) (Polit et al., 2007). Table 9.3 and Table 9.4 illustrate the rating forms that the expert reviewers were asked to use to rate the forward translation (Table 9.3), and then separately rate the back translation (Table 9.4).

Table 9.3. Form for Expert Reviewers to Rate the Relevance, Importance, and Clarity of the Forward Translation

No.	Original Item	FT synthesis item	Relevance				Importance				Clarity			
			1	2	3	4	1	2	3	4	1	2	3	4
1.	Tries to figure out what adults like	Mencoba mencari tahu tentang apa yang disukai orang dewasa												
2.	Tries to understand other children	Berusaha memahami anak-anak lain												
Etc.														

Note. FT = Forward Translation.

Table 9.4. Form for the Expert Reviewers to Rate the Relevance, Importance, and Clarity of the Back Translation

No.	Original Item	BT synthesis item	Relevance				Importance				Clarity			
			1	2	3	4	1	2	3	4	1	2	3	4
1.	Tries to figure out what adults like	Trying to find out what adults like												
2.	Tries to understand other children	Trying to understand other children												
Etc.														

Note. BT= Back Translation.

As for evidence of equivalence, Polit et al. (2007) suggested that good items are those with a score of 3 or 4, while poor items are rated 1 or 2. Content validity ratings were, similar to Step 7, dichotomized: 1 (for items with scores of 3 and 4) and 0 (for items with scores of 1 and 2). The Content Validity Index (CVI) of each item was estimated by summing the dichotomous scores and then dividing the sum by the number of reviewers. A minimum CVI value of 0.78 was suggested for an item to be deemed good (Polit et al., 2007) and, thus, provide evidence for content validity. In the Southeast Asian example, all the items had content validity indices above 0.78.

Step 9: Revisions and Further Consultation with the Developer

We have added this step, which is not explicitly in the ITC guidelines, because the results of feedback from the original developer of the DMQ, ratings of conceptual and linguistic equivalence, and also ratings of content validity may lead to revisions in the translated questionnaire. When the ratings from Tables 9.2, 9.3, and 9.4, were completed, the main researcher compiled the results and considered the comments made by the experts on some items. This led the researcher to make some changes at this step, often to adapt an item when the preferred wording in Step 7 of the conceptual similarity rating was different from the linguistic comparability rating. This

led to a somewhat revised version of the Southeast Asian DMQ 18. In general, however, the use of these scales provided evidence for both the semantic and the linguistic equivalence of the items and also for their content validity.

Further consultation with original developer could occur if the results of the assessment of equivalence and content validity by the expert reviewers lead to changes in the adapted questionnaire, as was the case in our example mentioned in the previous paragraph. Minor revisions resulted from correspondence with the developer of the DMQ. The purpose of this consultation was to make sure that the original item and the adapted items had the same meaning so that the adapted scale still measured the concepts intended by the original developer. After obtaining the agreement of the original developer, the adaptation was deemed to be appropriate to be used.

Step 10: TD-4 Small Scale Administration and Parent Feedback

The translated and adapted DMQ was revised in Step 9, so it should be administered to a few parents of children of the intended age for the planned studies, in order to find out whether the items and instructions would be clearly understood by adult raters such as parents/guardians or preschool teachers. (If this had been a translation of the school-age DMQ, it would be desirable to administer it to a few school-age children to be sure that they were able to answer it appropriately.) Feedback from these parents who were considered to be representative of the potential research sample indicated that items and instructions in this final form of the adaptation version were easy to comprehend and use. Thus, no further revisions were made.

Step 11: TD-5 Pilot Data

Collect and analyze pilot data on the adapted test version to enable item analysis, reliability assessment, and small-scale validity studies, indicating whether any necessary revisions should be made. Pilot data were collected from 169 parents who had kindergarten children aged 5-6 years old. Each of the five dimensions demonstrated relatively high levels of internal consistency ranging from .63 to 0.76. In addition, the relevance, importance, and clarity ratings provided by content experts in Step 8 were also a source evidence for content validity.

Because the pilot study did not suggest that further changes were needed, a full-scale validity study was then conducted for the Southeast Asian version of the preschool DMQ 18.

Step 12: C-1 Sample Selection

For the field test of the validity study, a random sample of 20 kindergarten classes was drawn from those in a large Southeast Asian city. All 20 teachers agreed to participate and to encourage parents to complete the DMQ; 75% of the parents signed a consent form and completed the DMQ and a family information form.

Because the intended population for the study was 5-6-year-old kindergarten children in this Southeast Asian country, the sample was probably representative at least of urban children in that country, who were required to attend kindergarten. The sample was also large enough for the statistics used in the planned validity study.

Step 13: C-2 Empirical Analysis

of the field test results. To validate the factor structure and provide further evidence of construct validity, confirmatory factor analysis (CFA) was used with a different sample than in the pilot study. The CFA sample consisted of 300 parents who rated the mastery motivation of their 5-6-year-old kindergarten children.

Second-order CFA (Hwang et al., 2017) was used to provide construct validity evidence for the translated and adapted questionnaire. The criteria specified by Hair et al. (2014) for deciding whether the model fits is based on several model fit indices. These indices include: (a) the chi-square *p* value; (b) Root Mean Square Error of Approximation (RMSEA: is an index of differences between the observed covariance matrix per degree of freedom and the hypothesized covariance matrix); (c) Goodness of Fit Index (GFI: is a measure of fit between the hypothesized model and the observed covariance matrix); (d) Comparative Fit Index (CFI: is an analysis of the model fit by examining the discrepancy between the data and the hypothesized model; CFI also adjusts for sample size issues in the chi-squared test of model fit and the normed fit index); and (e) Adjusted Goodness of Fit Index (AGFI: is a correction of the GFI, based on the number of indicators in each variables). The criteria for judging the fit of each index are presented in Table 9.5, which is a useful way to provide the goodness of fit index values for: the chi-square *p,* RMSA, GFI, CFI, and AGFI. Next to each required value in Table 9.5 is the goodness of fit statistic for our hypothetical example, and then a statement under "decision" about whether the statistic met the criterion value stated by Hair et al. (2014). Note that, except for the adjusted goodness of fit index, the values shown in Table 9.5 were considered to support a good fit with the model.

Table 9.5. Tests of Goodness of Fit Based on Confirmatory Factor Analysis

Fit Indices	Required Value	Obtained Value	Decision
χ^2 *p*-value	> .05	0.950	*Good fit*
RMSEA	< .08	0.045	*Good fit*
GFI	> .90	0.975	*Good fit*
CFI	> .90	0.960	*Good fit*
AGFI	> .90	0.890	*Marginal fit*

Abbreviation: AGFI = Adjusted Goodness of Fit Index; CFI = Comparative Fit Index; GFI = Goodness of Fit Index; RMSEA = Root Mean Square Error of Approximation.

When the model fit results are not a good fit, researchers can modify the model to obtain a parsimonious or better fitting model. However, the modification must be guided by theory and not just to improve the analysis (Shreiber et al., 2006). Based on the model fit results, a diagram or figure of the confirmatory factor analysis for the adapted questionnaire could be presented. In our example, the hypothesized second-order factor model demonstrated adequate fit.

Further evidence for construct validity is obtained from examination of the CFA factors. The minimum CFA factor loadings should be no less than .5, with a preferred value greater than .70. Other calculations that should be taken into account are a minimum construct reliability (CR) score in the range of .60-.70, a recommended Average Variance Extracted (AVE) coefficient of at least .50, and Cronbach alpha coefficients of at least .60. Table 9.6 presents the factor loadings, Cronbach alphas, construct reliability, and average variance extracted for the adapted DMQ 18 questionnaire of our hypothetical example.

In our hypothetical example, all the items had factor loading greater than .70, which implies that construct validity has been fulfilled according to the criteria. If any items had factor loadings lower than .50, they would have been potential candidates for deletion, especially if there was some other evidence that they were problematic. However, their deletion would affect the content validity of the tool (Hair et al., 2014). Because construct reliability (CR) values were all above .70, they were considered satisfactory. The average variance extracted (AVE) values yielded favorable results because all scores were greater than .50. Furthermore, Cronbach's alpha values met the requirements of above .60. Thus, the values shown in Table 9.6 indicate that the factor loadings, construct reliability, average variance extracted, and Cronbach's alphas were acceptable in this example.

Table 9.6. Factor Loadings, CR, AVE, and Cronbach's Alpha for Each DMQ 18 Scale

Item No.	Statement	FL	CR	AVE	Cronbach's Alpha
Cognitive/Object Persistence (COP)			0.90	0.65	0.705
1	Repeats a new skill until he can do it	0.90			
8	Tries to complete tasks, even if takes a long time	0.95			
10	Tries to complete toys like puzzles	0.80			
14	Works long to do something challenging	0.75			
18	Will work a long time to put something together	0.80			
Gross Motor Persistence (GMP)			0.85	0.60	0.735
3	Tries to do well at motor activities	0.75			
7	Tries to do well in physical activities	0.80			
16	Repeats jumping/running skills until can do them	0.85			
23	Tries hard to get better at physical skills	0.75			
25	Tries hard to improve throwing or kicking	0.80			
Social Persistence with Adults (SPA)			0.80	0.65	0.720
5	Tries to keep adults interested in talking	0.85			
9	Tries hard to interest adults in playing	0.90			
13	Tries hard to get adults to understand	0,85			
21	Tries to figure out what adults like	0,75			
24	Tries hard to understand my feelings	0,80			
Social Persistence with Children (SPC)			0.90	0.65	0.780
4	Tries to do things to keep children interested	0.75			
15	Tries to understand other children	0.80			
17	Tries hard to make friends with other kids	0.75			
20	Tries to get included when children playing	0.90			
22	Tries to keep play with kids going	0.75			
Mastery Pleasure (MP)			0.90	0.70	0.710
2	Smiles broadly after finishing something	0.95			
6	Shows excitement when is successful	0.90			
11	Gets excited when figures out something	0.75			
12	Is pleased when solves a challenging problem	0.75			
19	Smiles when makes something happen	0.80			
Total			0.85	0.65	0.805

Note. FL= Factor Loading; CR = construct reliability; AVE = average variance extracted.

Discriminant validity must also fulfill the requirement of having an AVE root square greater than the correlation value between dimensions. These validity results could be presented in a correlation matrix similar to that shown in Table 9.7. Note that each AVE root square coefficient, shown on the diagonal, should be larger than the correlations between the dimensions. The logic here is based on the idea that a latent construct should explain more of the variance in its item measures than it shares with another construct (Hair et al., 2014). Table 9.7 shows that the discriminant validity for the hypothetical example would be considered acceptable.

Table 9.7. Discriminant Validity of the Five DMQ 18 Scales

	COP	GMP	SPA	SPC	MP
Cognitive/Object Persistence (COP)	**0.805**				
Gross Motor Persistence (GMP)	0.515	**0.755**			
Social Persistence with Adults (SPA)	0.460	0.215	**0.775**		
Social Persistence with Children (SPC)	0.485	0.205	0.555	**0.825**	
Mastery Pleasure (MP)	0.565	0.400	0.445	0.570	**0.800**

Conclusion

One purpose of this chapter was to describe potential problems and biases related to the translation of questionnaires into a different language and culture. Many of these issues can be addressed through application of the guidelines from the International Test Commission (ITC) guidelines titled *ITC Guidelines for Translating and Adapting Tests*. The chapter applies the guidelines to describe the procedure we used to develop a hypothetical Southeast Asian version of DMQ 18. Finally, we describe statistical analyses, using realistic but hypothetical data, to assess the reliability and validity of such a translated and adapted questionnaire.

The appendices of this book provide complete English, Chinese, and Hungarian DMQ 18 forms, including the items for each of the four age-related versions, plus how to score them. The available DMQ 18 rating forms in other approved languages can be found in the online version of this book. These are open access and available for free for qualified researchers and clinicians. See Appendix C at the end of the book for how to request formal approval to use DMQ 18.

References

Barrett, K. C., & Morgan, G. A. (2018). Mastery motivation: Retrospect, present, and future directions. In *Advances in Motivation Science* (Vol. 5, pp. 1–39). Elsevier. https://doi.org/10.1016/bs.adms.2018.01.002

Borsa, J. C., Damasio, B. F., & Bandeira, D. R. (2012). Cross-cultural adaptation and validation of psychological instruments: Some considerations. *Paideia (Ribeirão Preto)*, *22*(53), 423–432.

Brislin, R. W. (1986). The wording and translation of research instruments. In W. J. Lonner & J. W. Berry (Eds.), *Field methods in cross-cultural research* (pp. 137–164). Sage.

Byrne, B. M., & Watkins, D. (2003). The issue of measurement invariance revisited. *Journal of Cross-Cultural Psychology*, *34*(2), 155–175.

Downing, S. M. (2002). Threats to the validity of locally developed multiple-choice tests in medical education: Construct-irrelevant variance and construct underrepresentation. *Advances in Health Sciences Education*, *7*(3), 235–241.

Greiff, S., & Iliescu, D. (2017). A test is much more than just the test itself: Some thoughts on adaptation and equivalence. *European Journal of Psychological Assessment*, *33*, 145–148. https://doi.org/10.1027/1015-5759/a000428

Hair, J. F., Black, W. C., Babin, B. J., & Anderson, R. E. (2014). *Multivariate Data Analysis* (7th ed.). Pearson Education Limited.

Hambleton, R. K. (2005). Issues, designs, and technical guidelines for adapting tests into multiple languages and cultures. In R. K. Hambleton, P. F. Merenda, & C. D. Spielberger (Eds.), *Adapting educational and psychological tests for cross-cultural assessment* (pp. 3–38). Lawrence Erlbaum.

Hambleton, R. K., & Patsula, L. (1998). Adapting tests for use in multiple languages and cultures. *Social Indicators Research*, *45*(1–3), 153–171. https://doi.org/10.1023/A:1006941729637

Hambleton, R. K., & Patsula, L. (1999). Increasing the validity of adapted tests: Myths to be avoided and guidelines for improving test adaptation practices. *Association of Test Publishers*, *1*(1), 1–13.

He, J., & van de Vijver, F. (2012). Bias and Equivalence in Cross-Cultural Research. *Online Readings in Psychology and Culture, 2*(2). https://doi.org/10.9707/2307-0919.1111

Hwang, A.-W., Wang, J., Józsa, K., Wang, P.-J., Liao, H.-F., & Morgan, G. A. (2017). Cross cultural invariance and comparisons of Hungarian-, Chinese-, and English-speaking preschool children leading to the revised Dimensions of Mastery Questionnaire (DMQ 18). *Hungarian Educational Research Journal, 7*(2), 32–47.

International Test Commission [ITC]. (2017). *The ITC guidelines for translating and adapting tests* (2nd ed., version 2.4). [www.InTestCom.org].

Jeanrie, C., & Bertrand, R. (1999). Translating tests with the International Test Commission's guidelines: Keeping validity in mind. *European Journal of Psychological Assessment, 15*(3), 277–283.

Kline, P. (1993). *Personality: The psychometric view*. Routledge.

Marín, G., & Marín, B. V. (1980). *Research with Hispanic populations (Applied social research methods series)* (Vol. 23). Sage.

Messick, S. (1995). Standards of validity and the validity of standards in performance assessment. *Educational Measurement: Issues and Practice*, *14*, 5–8.

Morgan, G. A., Harmon, R. J., Pipp, S., & Jennings, K. D. (1983). *Assessing mothers' perception of mastery motivation: The utility of the MOMM questionnaire*. Colorado State University, Fort Collins. https://sites.google.com/a/rams.colostate.edu/georgemorgan/mastery-motivation.

Morgan, G. A., Maslin-Cole, C. A., Harmon, R. J., Busch-Rossnagel, N. A., Jennings, K. D., Hauser-Cram, P., & Brockman, L. (1993). Parent and teacher perceptions of young children's mastery motivation: Assessment and review of research. In D. Messer (Ed.), *Mastery motivation in early childhood: Development, measurement and social processes* (pp. 109–131). Routledge.

Özbey, S., & Daglioglu, H. E. (2017). Adaptation study of the motivation scale for the preschool children (DMQ18). *International Journal of Academic Research*, *4*(1–2), 1–14.

Polit, D. F., Beck, C. T., & Owen, S. V. (2007). Is the CVI an acceptable indicator of content validity? Appraisal and recommendations. *Research in Nursing & Health*, *30*(4), 459–467. https://doi.org/10.1002/nur.2019

Rahmawati, A., Fajrianthi, Morgan, G. A., & Józsa, K. (2020). An adaptation of DMQ 18 for measuring mastery motivation in early childhood. *Pedagogika, 140*(4), 18–33. https://doi.org/10.15823/p.2020.140.2

Reise, S. P., Widaman, K. F., & Pugh, R. H. (1993). Confirmatory factor analysis and item response theory: Two approaches for exploring measurement invariance. *Psychological Bulletin, 114*(3), 552–566. https://doi.org/10.1037/0033-2909.114.3.552

Rios, J. A., & Hambleton, R. K. (2016). Statistical methods for validating test adaptations used in cross-cultural research. In N. Zane, G. Bernal, & F. Leong (Eds.), *Evidence-based psychological practice with ethnic minorities: Culturally informed research and clinical strategies* (pp. 103–124). American Psychological Association.

Salavati, M., Vameghi, R., Hosseini, S., Saeedi, A., & Gharib, M. (2018). Mastery motivation in children with Cerebral Palsy (CP) based on parental report: Validity and reliability of Dimensions of Mastery Questionnaire in Persian. *Materia Socio-Medica, 30*(2), 108. https://doi.org/10.5455/msm.2018.30.108-112

Schreiber, J. B, Amaury, N., Stage, F. K, Barlow, E. A., & King, J. (2006). Reporting Structural Equation Modeling and Confirmatory Factor Analysis results: A review. *The Journal of Educational Research*, *99*, 323–338. https://doi.org/10.3200/JOER.99.6.323-338

Shaoli, S. S., Islam, S., Haque, S., & Islam, A. (2019). Validating the Bangla version of the Dimensions of Mastery Questionnaire (DMQ-18) for preschoolers. *Asian Journal of Psychiatry*, *44*, 143–149. https://doi.org/10.1016/j.ajp.2019.07.044

Sireci, S. G. (2005). Using bilinguals to evaluate the comparability of difference language versions of a test. In R. K. Hambleton, P. F. Merenda, & C. D. Spielberger (Eds.), *Adapting educational and psychological tests for cross-cultural assessment* (pp. 117–138). Lawrence Erlbaum.

Sireci, S., & Faulkner-Bond, M. (2014). Validity evidence based on test content. *Psicothema*, *26*(1), 100–107.

Sireci, S. G., Patsula, L., & Hambleton, R. K. (2005). Statistical methods for identifying flaws in the test adaptation process. In R. K. Hambleton, P. Merenda, & C. Spielberger (Eds.), *Adapting educational and psychological tests for cross-cultural assessment* (pp. 93–116). Lawrence Erlbaum.

Tanzer, N. K. (2005). Developing tests for use in multiple languages and cultures: A plea for simultaneous development. In R. K. Hambleton, P. F. Merenda, & C. D. Spielberger (Eds.), *Adapting educational and psychological tests for cross-cultural assessment* (pp. 235–264). Lawrence Erlbaum.

Uchida, Y., & Kitayama, S. (2009). Happiness and unhappiness in East and West: Themes and variations. *Emotion*, *9*, 441–456.

van de Vijver, F. J. R., & Hambleton, R. K. (1996). Translating tests: Some practical guidelines. *European Psychologist*, *1*, 89–99.

van de Vijver, F. J. R., & Leung, K. (1997). *Methods and data analysis for cross-cultural research*. Sage.

van de Vijver, F. J. R., & Poortinga, Y. H. (2005). Conceptual and methodological issues in adapting tests. In R. K. Hambleton, P. F. Merenda & C. D. Spielberger (Eds.), *Adapting educational and psychological tests for cross-cultural assessment* (pp. 39–63). Lawrence Erlbaum.

van de Vijver, F. J. R., & Tanzer, N. K. (1997). Bias and equivalence in cross-cultural assessment: An overview. *European Review of Applied Psychology*, *47*(4), 263–279.

van de Vijver, F. J. R., & Tanzer, N. K. (2004). Bias and equivalence in cross-cultural assessment: An overview. *European Review of Applied Psychology*, *54*(2), 119–135.

Wang, J., Józsa, K., & Morgan, G. A. (2014, May). *Measurement invariance across children in US, China, and Hungary: A revised Dimensions of Mastery Questionnaire (DMQ).* [Summary] Program and Proceedings of the 18th Biennial Developmental Psychology Research Group Conference, Golden, CO.

Wolf, E. J., Harrington, K. M., Clark, S. L., & Miller, M. W. (2013). Sample size requirements for structural equation models: An evaluation of power, bias, and solution propriety. *Educational and Psychological Measurement*, *73*(6), 913–934. https://doi.org/10.1177/0013164413495237

Wolming, S., & Wikstrom, C. (2010). The concept of validity in theory and practice. *Assessment in Education: Principles, Policy & Practice*, *17*(2), 117–132.