Achievement Criteria for ITC Guidelines on Test Adaptation

A. Hernández[1], M. D. Hidalgo[2], R. K. Hambleton[3] and J. Gómez-Benito[4].

[1] Universitat de València, [2] Universidad de Murcia, [3] University of Massachusetts at Amherst, and [4] Universitat de Barcelona

The process followed to develop the criterion checklist and the corresponding achievement criteria is described in:

> **Hernández, A., Hidalgo, M. D., Hambleton, R. K. & Gómez-Benito, J. (2020). International Test Commission guidelines for test adaptation: A criterion checklist. *Psicothema, 32* (3), 390-398. doi: 10.7334/psicothema2019.306**

Please, refer to the publication above when referring to this criterion checklist.

Introduction

The second edition of the ITC guidelines on test translation and adaptation (ITC, 2017) can be downloaded at:
https://www.intestcom.org/files/guideline_test_adaptation_2ed.pdf

In accordance with the ITC guidelines, the expressions ‘test’ and ‘testing’ should be interpreted broadly. Thus, in the achievement criteria we propose, the term ‘test’ refers to all sorts of tests (psychological tests, such as intelligence and personality tests; educational and personnel tests, such as competence and cognitive test batteries; clinical screening tests, etc.) and covers a variety of formats (questionnaires, behavioral rating scales, performance ratings, and other assessment tools.)

The 18 guidelines on test adaptation proposed by the ITC (2017, second edition) are organized into six broad categories (pre-condition, development, confirmation, administration, scoring and interpretation, and documentation). In this document the 18 guidelines have been operationalized through a number of criteria (29 in total). All ITC guidelines for test adaptation are considered, although the order in which they are addressed has been changed, in some cases so as to facilitate appraisal of the degree to which the criteria have been accomplished. Each of our proposed criteria has an alphanumerical label which identifies its corresponding category of ITC guideline and the specific criterion within that category. The label also includes, between brackets, two general counters, one referring to the ITC guideline and one to our criterion. Thus, for example, PC3-1 (G3, C4) refers to the first criterion for operationalizing pre-condition guideline PC3, which is the third guideline proposed by the ITC and the fourth criterion proposed by us. Similarly, TD3-2 (G6, C9) refers to the second criterion for operationalizing test development guideline TD3, which is the sixth guideline proposed by the ITC and the ninth criterion proposed by us.

1. PRE-CONDITION GUIDELINES

These guidelines apply to planning and to preliminary issues that must be considered before proceeding to the adaptation:

PC1: Obtain the necessary permission from the holder of the intellectual property rights relating to the test before carrying out any adaptation.

PC2: Evaluate that the amount of overlap in the definition and content of the construct measured by the test and the item content in the populations of interest is sufficient for the intended use (or uses) of the scores.

PC3: Minimize the influence of any cultural and linguistic differences that are irrelevant to the intended uses of the test in the populations of interest.

Operationalization

Before addressing guideline PC1, on copyright permissions, it is necessary to determine whether, on the basis of guideline PC2, translation/adaptation is the best option. Thus we start with PC2 and then proceed to PC1 and PC3, respectively.

PC2-1 (G2, C1): Provide theoretical and empirically-based evidence that the construct of interest is relevant to the target population (van der Vijver & Leung, 2011).

Excellent: There are both theoretical reasons (papers, expert judgments, etc.) and empirical evidence (e.g., by means of interviews, observations or surveys of individuals from the target population) indicating that the construct is relevant to the target population.

Acceptable: There are theoretical reasons (papers, expert judgments, etc.) to believe that the construct is relevant to the target population. However, there is as yet no empirical evidence of the construct's relevance.

PC2-2 (G2, C2): Consider whether the meaning of the construct can be generalized across cultures, and ensure and be able to justify that translation/adaptation is preferable to creating a brand new test for the target population (Hambleton, Merenda, & Spielberg, 2005; van der Vijver & Leung, 2011).

Excellent: A group of experts in the construct and in the cultures and languages involved checks the definition of the construct, the dimensionality, and the items of the source test that capture that definition and dimensionality. They then provide arguments that a) there is a complete overlap in the construct, and b) the source items are adequate for representing the construct in the target population.

Acceptable: A group of experts in the construct and in the cultures and languages involved checks the definition of the construct, the dimensionality, and the items of the source test that capture that definition. They then provide arguments that there is a partial overlap in the construct and that a meaningful number of the source items are adequate for representing the construct in the target population.

PC1-1 (G1, C3)*: If adaptation is the best option, ask the copyright owners for permission to adapt the test, even if the test is going to be used for research purposes only (see ITC statements on using tests for research and check copyright laws; ITC, 2014).

Acceptable=Excellent: Written permission is obtained from the copyright owner.

* Note that in many cases the copyright owner will not be the author but rather the publisher or distributor of the test. In addition, when obtaining permission the following issues need to be considered: whether editors/copyright owners allow changes in the test structure, based, for example, on the results of a pilot study; whether they would accept changes in verbal content and not just translation (e.g., synonyms); whether they agree that research on the test and its possible modifications may be published in another country; and whether they

accept that norms, if any, may be presented in different ways from those corresponding to the original test, in light of the results derived from the test adaptation procedure.

In international projects (e.g., PISA, TIMMS), it is common for tests to be simultaneously developed for use in different languages and cultures. In cases of simultaneous test development, this criterion would not be applicable since there is no source test to be adapted.

PC3-1 (G3, C4): If adaptation is the best option, check for cultural and linguistic differences before starting the adaptation process and take them into consideration in the adapted version so as to prevent bias and to design studies to control for potential bias (Arnold & Smith, 2013; Pena, 2007).

Excellent: The cultural and linguistic differences (item format, material or word familiarity, emic and etic concepts, lifestyles, etc.) on the basis of both expert opinions (experts in the target language and culture) and empirical evidence (interviews, observation or surveys of individuals from the target population, etc.) are systematized and documented.

Acceptable: The most important cultural and linguistic differences on the basis of expert opinions (experts in the target language and culture) are summarized and documented. However, no empirical evidence is yet available.

2. DEVELOPMENT

TD1 Ensure that the translation and adaptation processes consider linguistic, psychological, and cultural differences in the intended populations through the choice of experts with relevant expertise

TD2 Use appropriate judgmental designs and procedures to maximize the suitability of the test adaptation in the intended populations.

TD3 Provide evidence that the test instructions and item content have similar meaning for all intended populations.

TD4 Provide evidence that the item formats, rating scales, scoring categories, test conventions, modes of administration, and other procedures are suitable for all intended populations.
TD5 Collect pilot data on the adapted test to enable item analysis, reliability assessment and small-scale validity studies so that any necessary revisions to the adapted test can be made

NOTE: Before starting the test adaptation process, researchers should ensure compliance with the ethical code for research involving human beings and make sure that participants involved in the adaptation process sign an informed consent document, as defined in their own countries.

Operationalization

TD1-1 (G4, C5): Form a multidisciplinary team composed of: a) professional translators who are proficient in the source and target languages (if different languages are involved) and have knowledge of the cultures involved, b) experts in the construct to be measured, c) experts in the cultures involved, and d) experts in test construction. In some cases, the same team member may be an expert in more than one of these aspects, for example, the languages and cultures, the construct and cultures, etc. (Epstein, Santo, & Guillemin, 2015).

*Excellent**: The team includes a minimum of five experts: two professional translators with knowledge of the cultures involved and who have received some training in item writing principles, one expert in the construct to be measured, one expert in the cultures involved, and one expert in test construction.

*Acceptable**: The team includes a minimum of three experts: two professional translators with knowledge of the cultures involved and one expert in the construct to be measured and/or test construction. In this case some overlap among categories is permitted: translators may also be experts in the construct and/or in the test construction process, for example.

*Note that more experts are required when important decisions for individuals or groups will be made on the basis of score comparisons across cultures. It should also be taken into account that professional translators are not necessary if the adaptation does not require translation from one language to another (e.g., a test developed in Spain that is going to be adapted for administration in Mexico). However, linguistic equivalence is still a crucial goal in this case, and the multidisciplinary team is still needed. To ensure team quality, the procedure for selecting experts, as well as their qualifications and experience, must be documented. Finally, if the adapted test needs to accommodate special populations, the team should incorporate additional professionals (e.g., psychologists or special education professionals (see Leong, Bartram, Cheung, Geisinger, & Iliescu, 2016)

TD2-1 (G5, C6): Use recommended translation designs and justify the choice. Forward, backward or simultaneous translations may be used, depending on the purpose of the adaptation, the scope of the project, the number of cultures involved, and whether or not it will be necessary to compare scores of individuals from different cultures. Independent translators (at least two professionals or two teams) are involved in forward and backward translations (Hambleton, 2005). If a test is intended to be used cross-culturally from its inception, use simultaneous/concurrent development of multiple language versions of the test from the outset.

Excellent: Forward translation is performed by independent translators (see Hagell, Hedin, Meads, Nyberg, & McKenna, 2010) or several designs are combined to obtain the initial version of the adapted test (Wild et al., 2005). The reason for preferring independent forward translation from source to target, as opposed to backward translation, is that potential discrepancies are detected and reviewed directly in the target language (ITC, 2017). In large-scale cross-cultural assessment programs, such as PISA, different language versions may be used as separate sources for translation, which are afterwards reconciled into a single target version (Grisay, 2003). Apart from allowing researchers to identify and review possible discrepancies directly in the target language, using more than one source language helps to minimize the impact of cultural characteristics of the source (ITC, 2017).

Acceptable: Backward translation is adequately used and the choice is justified.

TD2-2 (G5, C7): Have several translators that work independently and form a committee of experts to review and compare the proposed translations in order to compile judgmental reviews, resolve possible discrepancies, and produce a consensus version (Epstein, Osborne, Elsworth, Beaton, & Guillemin, 2015; Koller et al., 2012).

> *Excellent*: At least two translators work on each design phase and an independent team of qualified translators and experts (in the culture, construct, and measurement) reviews all of the versions and works with the original translators to resolve discrepancies.
>
> *Acceptable*: At least two translators work on each design phase and an independent translator who has knowledge of the cultures involved reviews the source and target versions and works with the original translators to resolve discrepancies.

TD3-1 (G6, C8): Ensure that the instructions are clear and comprehensible, using terms that are familiar to the target population.

> *Excellent*: The team of experts has reviewed, approved, and documented the adequacy of the instructions. In addition, pilot studies and/or pre-tests (e.g., cognitive interviews; see Levin et al., 2009; Padilla & Benítez, 2014) with bilingual samples or using samples from the intended target populations have been carried out and the results support the adequacy of the instructions.
>
> *Acceptable*: The team of experts has reviewed, approved, and documented the adequacy of the instructions.

TD3-2 (G6, C9): Ensure that the item content is clear and expressed with similar levels of commonality and difficulty in the source and target cultures. Linguistic elements that could hinder the understanding of the translated version, such as words with different meaning, double negations, etc., should be avoided. Non-linguistic elements (such as images and pictures) must be contextualized for the target population (Hambleton & Zenisky, 2011; van der Vijver & Tanzer, 2004).

*Excellent**: The team of experts has reviewed, approved, and justified the adequacy and comparability of the item content. In addition, pilot studies and/or pre-tests (e.g., cognitive interviews) with samples from the intended populations (ideally bilingual) have been carried out to guarantee that the item content is clear and similarly understood.

*Acceptable**: The team of experts has reviewed, approved, and justified the adequacy and comparability of item content.

* It is recommended that experts use a quality control checklist for adapted items, such as the one proposed by Hambleton and Zenisky (2011), focusing on items that are relevant for the type of test under adaptation (this checklist is also available in Appendix C of the ITC guidelines (ITC, 2017).

TD4-1 (G7, C10): Try to ensure that the item format, response options, scoring rubrics, if any, and administration mode are similar in both versions (Beaton, Bombardier, Guillemin, & Ferraz, 2000).

Excellent: The experts agree (and justify) that the format, response options, scoring rubrics, if any, and administration mode are highly similar.

Acceptable: There are some meaningful differences in format, response options, etc., but the experts agree (and justify) that the differences do not substantially alter the meaning or difficulty of the items.

TD4-2 (G7, C11): Ensure that the target population is sufficiently familiar with the procedures used (item format, response scales, scoring rubrics (if any), test conventions, and administration mode) in the adapted test (Hambleton, Merenda, & Spielberg, 2005; van der Vijver & Leung, 2011).

Excellent: There are previous studies that use the procedures involved (item format, administration mode, etc.) and they show there is no problem in this respect. In addition, either the results of a pilot study show that individuals from the target population do not have difficulties using the proposed procedures, or, if difficulties are detected, there are enough trial items to ensure that individuals become familiar with all aspects of the procedures.

Acceptable: Either there are previous studies that use the procedures involved and show there is no problem in this respect or the results of a pilot study show that individuals from the target population do not have difficulties using the proposed procedures. If a problem is detected, there are enough trial items to ensure that individuals become familiar with all aspects of the procedures.

TD5-1 (G8, C12): Check the psychometric quality (item analysis, reliability, and validity) of the scores from the adapted test in a pilot sample of the target population and, based on the results, make any necessary revisions for the final version of the test. Pilot samples should be large enough to carry out the statistical analysis involved in the pilot study.

*Excellent**: Item analysis (difficulty, discrimination, and, when necessary, distracter analysis) has been carried out and reliability coefficients (or their IRT counterparts, that is, information function) have been estimated. In addition, at least one validity study based either on the relationships between the test scores and other variables (e.g., intercorrelations between different dimensions assessed or correlations with variables from the nomological network to assess criterion, convergent or discriminant validity or group differences) or the internal structure of the test** has been carried out. Either the results are excellent according to standard criteria (see, for example, the EFPA test review model; Evers et al., 2013; Furr, 2017; Mellenbergh, 2011) or problematic items are improved.

*Acceptable**: Item analysis has been carried out (difficulty, discrimination, and, when necessary, distracter analysis) and reliability coefficients (or their IRT counterparts) computed. Either the results are at least acceptable according to standard criteria (see, for example, the EFPA test review model) or problematic items are improved.

* Validity evidence based on the content of the items and/or validity evidence based on the response processes (e.g., cognitive interviews) has already been considered in TD3-2 (G6, C9).

** Note that when measurement equivalence is tested across groups, configural invariance is assessed and used to determine if the internal structure is the expected one and is invariant in the different comparison groups.

3. CONFIRMATION [EMPIRICAL ANALYSIS] GUIDELINES

C1 Select sample with characteristics that are relevant for the intended use of the test and of sufficient size and relevance for the empirical analyses.

C2 Provide relevant statistical evidence about the construct equivalence, method equivalence, and item equivalence for all intended populations.

C3 Provide evidence supporting the norms, reliability and validity of the adapted version of the test in the intended populations.

C4 Use an appropriate equating design and data analysis procedures when linking score scales from different language versions of a test.

Operationalization

C1-1 (G9, C13): Ensure that the target sample is large enough to carry out the necessary statistical analyses and to adequately represent the population*

Excellent: The target sample size has been chosen using probabilistic sampling and is appropriate according to the population size and the variability of the measured characteristics within the population. In addition, the sample is large enough to carry out the statistical analysis and have stable estimates of the parameters investigated in the study.

Acceptable: The target sample size is appropriate according to the population size and the variability of the measured characteristics within the population. The sample has not been chosen using probabilistic sampling, but the most important sociodemographic characteristics of the population have been considered. In addition, the sample is large enough to carry out the statistical analysis and have stable estimates of the parameters investigated in the study.

*It is important to bear in mind that some techniques, such as item response theory (IRT) or confirmatory factor analysis (CFA), may require large sample sizes (Byrne & Van de Vijver, 2014; DeMars, 2010; Gagne & Hancock, 2006; Wolf, Harrington, Clark, & Miller, 2013).

C1-2 (G9, C14): When the focus of interest is on cross-cultural comparisons, ensure that the source and target samples are comparable for all relevant variables except for language and/or cultural background (van der Vijver & Leung, 1997).

Excellent: The target sample has the same characteristics as the source sample (i.e., there are non-significant differences in relevant variables: sociodemographic, academic, clinical, etc.), except for language and/or cultural background.

Acceptable: The existing differences between the target and source samples in relevant variables have been controlled for using appropriate matching designs or statistical techniques.

C2-1 (G10, C15): When there is interest in comparing the source and target populations, use statistical procedures to ensure that construct equivalence holds across populations (van der Vijver & Poortinga, 2004; van der Vijver & Tanzer, 2004).

Excellent: Procedures based on the internal structure of the test (i.e., factor analysis) and on the nomological network (i.e., comparing statistically the pattern of relationships between the construct and relevant variables) have been applied with adequate sample sizes. The results obtained across procedures support construct equivalence.

Acceptable: One adequate procedure, based either on the internal structure of the test or on nomological network analysis, has been applied with adequate sample sizes and the results support construct equivalence.

C2-2 (G10, C16): When there is interest in comparing the source and target populations, check for method equivalence (instrument characteristics, administration process, and sample characteristics).

Excellent: Excellence has been achieved for the criteria referring to sample bias (C1), instrument characteristics (TD4), and the administration process (A1 and A2)*. Additional analyses are carried out to test if the potential methodological threats identified in the design, development, and administration phases are affecting the final results. Results show that method effects are not an issue.

Acceptable: An acceptable level (as a minimum) is achieved for the criteria referring to sample bias (C1), instrument characteristics (TD4), and the administration process (A1 and A2)*. Additional analyses are carried out to test if the potential methodological threats identified in the design, development, and administration phases are affecting the final results. If results show that method effects are an issue, the influence is controlled for.

* The achievement criteria can be fully addressed after assessing criteria A1 and A2 of the administration guidelines.

C2-3 (G10, C17): When there is interest in comparing the source and target populations, assess DIF between the cultural groups to be compared using statistical procedures appropriate to the item format, sample size, and test dimensionality* (Hidalgo & Gómez-Benito, 2010; Sireci & Rios, 2013).

Excellent: The most adequate statistical technique for DIF detection (considering item format, sample size, etc.) is applied using a matching criterion purification procedure. Apart from using significance tests, DIF effect size measures and indicators of differential test functioning (DTF) are reported.

Acceptable: The most adequate statistical technique for DIF detection (considering item format, sample size, etc.) is used. Apart from using significance tests, DIF effect size measures are reported.

* The results of these analyses should be taken into account for subsequent steps in the adaptation process (see, especially, C4-1 and the section on score scales and interpretation guidelines).

C2-4 (G10, C18): In the event that DIF is detected at meaningful levels, carry out analyses to understand the reasons for the DIF (e.g., linguistic or method effects) across cultures (Benitez, Padilla, Hidalgo, & Sireci, 2016; Gómez-Benito, Sireci, Padilla, Hidalgo, & Benitez, 2018).

Excellent: Mixed methods, combining qualitative (e.g., cognitive interviews or expert judges) and quantitative approaches (e.g., experiments or quasi-experiments), are used to understand the reasons for DIF.

Acceptable: Either qualitative or quantitative studies are carried out to understand the reasons for DIF.

C3-1 (G11, C19): Ensure that the type of reliability indicators reported is adequate for the type of test, using adequate statistical analysis and sample sizes. The obtained values must be satisfactory and the standard error of measurement must be reported (AERA, APA, & NCME, 2014; Evers et al., 2013; Terwee et al., 2012).

Excellent: Depending on the purpose of the test, evidence about internal consistency, measurement stability, and/or inter-rater agreement are provided (for more information, see APA standards, 2014). Information about the standard error of measurement, obtained by means of either classical test theory or item response theory is also provided. Results show good or excellent values when considering the importance of the decisions to be made from the test (see EFPA test review model; Evers et al., 2013).

Acceptable: Depending on the purpose of the test, evidence about internal consistency, measurement stability, and/or inter-rater agreement are provided (for more information, see APA standards, 2014). Information about the standard error of measurement, obtained by means of either classical test theory or item response theory is also provided. Results show acceptable values when considering the importance of the decisions to be made from the test (see EFPA test review model; Evers et al., 2013).

C3-2 (G11, C20): Provide validity evidence consistent with the intended use of test scores, using adequate statistical analysis and sample sizes (AERA, APA, & NCME, 2014; Evers et al., 2013; Terwee et al., 2012).

Excellent: Validation hypotheses/objectives are formulated for different types of validity evidence. Evidence based on content validity (partially checked through the 'Excellent' criteria for PC2 and TD1 to TD3 guidelines), the internal structure of the test, and relationships with other variables are reported. When relevant, evidence based on the response process (partially checked through the 'Acceptable' criteria for TD3) and/or consequences of testing are also considered. The results obtained considering all sources of evidence support the intended use of the test.

Acceptable: Validation hypotheses/objectives are formulated for the intended use of test scores. The results of the evidence collected support the intended use of the test.

C3-3 (G11, C21): Ensure and verify that the norms are adequate for interpreting the test scores of the target population.

Excellent: Specific norms have been developed for the target population. The sample has been chosen using probabilistic sampling methods, it is large enough to ensure representativeness, and sampling error is reported. In addition, and apart from a general norm, specific norms are developed to account for relevant variables (age, gender, etc.) or specific populations (e.g., clinical or disabled), when present.

Acceptable: Either there is evidence that the norms developed for the original version can be used for the target population, or specific norms have been developed with the adapted version for the target population. In the latter case, the sample has not been chosen following a probabilistic sampling method, but it has been selected to ensure that the most important sociodemographic characteristics of the population are considered.

C4-1 (G12, C22): When cross-cultural/cross-lingual assessment is the objective, and comparability of scores across groups is necessary but some items are functioning differentially, use appropriate linking designs and data analysis procedures before comparison (Angoff, 1984; Dorans, Pommerich, & Holland, 2007; Kolen & Brennan, 2004).

Excellent: To choose the anchor items that will be used to equate the samples, judgmental criteria (i.e., items that experts deem to be the easiest to translate and adapt) and empirical criteria (i.e., using bilingual samples or a subset of items that have shown to be DIF-free in DIF analysis) are combined. Data equating analysis and sample sizes are adequate.

Acceptable: To choose the anchor items that will be used to equate the samples, either judgmental criteria (i.e., items that experts deem to be the easiest to translate and adapt) or empirical criteria (i.e., using bilingual samples or a subset of items that have shown to be DIF-free in DIF analysis) are used. Data equating analysis and sample sizes are adequate.

4. ADMINISTRATION GUIDELINES

A1 Prepare administration materials and instructions to minimize any culture- and language-related problems that are caused by administration procedures and response modes that can affect the validity of the inferences drawn from the scores.

A2 Specify testing conditions that should be followed closely in all populations of interest.

Operationalization

A1-1 (G13, C23): For all administration materials and instructions the requirements specified in the development guidelines have been checked (TD3 to TD5). The experience accumulated when administering the original version of the test in the source population should be taken into account to prevent possible administration problems in the target population.

Excellent: The criteria specified in the development guidelines (TD3 to TD5) are all met to the level of 'Excellent'.

Acceptable: The criteria specified in the development guidelines (TD3 to TD5) are met to at least the level of 'Acceptable'.

A2-1 (G14, C24): When cultural comparisons are of interest, ensure that the testing conditions (administration mode, time restrictions, information about the test purpose, etc.) are standardized across groups. If changes are necessary, data should be collected to evaluate the possible impact of different testing conditions.

Excellent: Testing conditions are specified to be the same, or there is a pilot study showing that the changes introduced do not affect the results.

Acceptable: Differences in testing conditions are unavoidable but are clearly described, such that cautious comparisons can be made across populations. Some evidence is provided suggesting that the differences should not have a substantial impact.

A2-2 (G14, C25): Ensure that the interviewers or test administrators have the credentials required for the type of test to be administered. Test administrators should submit a signed pledge to conduct their activities in accordance with the code of ethics and principles of professional practice established by the relevant national professional associations and bodies.

Excellent: All administrators have the required credentials (if any) and documented experience in administering the type of test that has been adapted. Test administrators have signed the aforementioned pledge to conduct their activities in accordance with ethical and professional codes of behavior.

Acceptable: The test administrators have the required credentials (if any). All or some of them may have no experience in administering the type of test that has been adapted, but training is provided and its outcome is evaluated. Only individuals whose training performance is adequate are selected as test administrators. Test administrators have signed the aforementioned pledge to conduct their activities in accordance with ethical and professional codes of behavior.

5. SCORE SCALES AND INTERPRETATION GUIDELINES

SSI1 Interpret any group score differences with reference to all relevant available information.

SSI2 Only compare scores across populations when the level of invariance has been established on the scale on which scores are reported.

Operationalization

Before addressing guideline SSI1, on the interpretation of group score differences, it is necessary to determine whether the scores are comparable on the basis of guideline SSI2. Thus, we start with SSI2 and then proceed to SSI1.

SSI2-1 (G16, C26): To compare individual scores of people belonging to different cultures, and/or mean scores across cultures, ensure that measurement equivalence (a.k.a. lack of DIF) is assessed and supported, at least for a meaningful number of items* (Byrne, 2008; Byrne & van der Vijver, 2017; Dimitrov, 2010; van der Vijver & Leung, 2011).

Excellent: Measurement invariance has been tested and supported: the differences in parameters are not significant, or are negligible, according to effect sizes.

Acceptable: Measurement invariance has been tested and only partial invariance has been supported. In this case, only invariant items should contribute to the group comparisons, provided that content validity is not compromised, or alternatively corrected scores (after applying an adequate linking design) should be used.

* Note that if the focus of interest is solely the comparability of relationships between the construct of interest and other relevant constructs, then only metric invariance needs to be supported.

SSI1-1 (G15, C27): When score comparisons are justified on the basis of measurement invariance analysis, consider a number of interpretations of cross-cultural differences, taking into account the information that has been systematized and documented in PC3-1 (G3, C4) regarding cultural and linguistic distance. To understand the differences in the observed scores, the role of these variables (e.g., religiosity, individualism, different response tendencies) should be considered (SSI1) (van der Vijver & Leung, 2011).

Excellent: The information systematized and documented in PC3-1 (G3, C4) has been considered and the role of key variables has been evaluated empirically.

Acceptable: The information summarized in PC3-1 (G3, C4) has been considered in order to provide reasonable explanations for the differences obtained.

6. DOCUMENTATION GUIDELINES

Doc-1 Provide technical documentation of any changes, including an account of the evidence obtained to support equivalence, when a test is adapted for use in another population.

Doc-2 Provide documentation for test users that will support good practice in the use of an adapted test with people in the context of the new population.

Operationalization

Doc-1-1 (G17, C28): Create a number of documents and make them accessible to relevant stakeholders, providing information about the following issues:

a) Evidence (theoretical and empirical) that the construct of interest is relevant to the target population and has the same meaning as in the original population (as considered in PC2-1 (G2, C1) and PC2-2 (G2, C2)).

b) The cultural distance and linguistic differences between cultures on the basis of expert opinions (experts in the target language and culture) and empirical evidence (interviews, observation or surveys of individuals from the target population, etc.) (as considered in PC3-1 (G3, C4)).

c) The steps taken to prevent bias (e.g., cultural decentering, translation design, development of a detailed and standardized administration protocol across cultures, etc.) (see A1 and A2).

d) All deviations from the original test and the original administration conditions (e.g., number of items, administration conditions, time), explaining the reasons for the changes and, when the main purpose is to compare scores across cultures, the collected evidence that changes do not seriously compromise comparability (see A1 and A2).

e) Information about the team members who have worked on the adaptation process and the adaptation design, justifying the specific design choice.

f) When score comparison across cultures is required, the characteristics of the source and target populations, the sample selection procedure, and the sample characteristics, as well as the main differences (if any) and the possible influence that these differences may have on comparability.

g) Technical information about the psychometric properties of the scores obtained with the adapted test (item analysis, reliability, and evidence of validity).

h) When the objective of the adaptation is to compare individuals from different cultures, technical documentation including the results of the measurement equivalence studies carried out and the conclusions about the level on which the cultures can be compared.

Excellent: The documentation includes fully comprehensive information about all of the above issues. This is necessary for users to determine the adequacy of any changes that have been made and to allow researchers to replicate the adaptation process. It is stated where the documentation is available and how it is accessible.

Acceptable: The documentation includes sufficient information about the above issues to judge the adequacy of the adaptation process. It is stated where the documentation is available and how it is accessible.

Doc-2-1 (G18, C29): Make sure that the materials and documentation which accompany the test (e.g., the test manual) are clear (instructions, description of the scope of application, practical examples of its use, etc.) so as to ensure that the test is adequate for the intended population, that the test administration is standardized, and that scores can be interpreted adequately (see administration and scoring sections) (Doc2).

Excellent: The International Test Commission Guidelines on Test Use and accepted standards for assessing test quality (according, for example, to the EFPA test review model or the Standards for Educational and Psychological Testing (AERA, APA, & CCME, 2014), among others) have been considered. Results show good or excellent values (Evers, Muñiz, Hagemeisteir, Hstmælingen, Lindley, Sjöberg, & Bartram, 2013).

Acceptable: The International Test Commission Guidelines on Test Use and accepted standards for assessing test quality (according, for example, to the

EFPA test review model or the Standards for Educational and Psychological Testing (AERA, APA, & CCME, 2014), among others) have been considered. Results show adequate values (Evers et al., 2013).

REFERENCES

American Educational Research Association, American Psychological Association, and National Council on Measurement in Education (AERA, APA, & CCME) (2014). *Standards for educational and psychological testing*. Washington, DC, USA: American Psychological Association.

Angoff, W. H. (1984). *Scales, norms, and equivalent scores.* Princeton, NJ: Educational Testing Service.

Arnold, B. R., & Smith, J. L. (2013). Methodologies for test translation and cultural equivalence. In F. Paniagua & A.-M. Yamada (Eds.), *Handbook of multicultural mental health: Assessment and treatment of diverse populations* (2nd ed.). San Diego, CA, US: Elsevier Academic Press.

Beaton, D. E., Bombardier, C., Guillemin, F., & Ferraz, M. B. (2000). Guidelines for the process of cross-cultural adaptation of self-report measures. *Spine*, *25*(24), 3186-3191.

Benítez, I., Padilla, J. L., Hidalgo Montesinos, M. D., & Sireci, S. G. (2016). Using mixed methods to interpret differential item functioning. *Applied Measurement in Education*, *29*(1), 1-16.

Byrne, B. M. (2008). Testing for multigroup equivalence of measuring instrument: A walk through the process. *Psicothema, 20,* 872-882.

Byrne, B. M., & van der Vijver, F. J. R. (2014). Factorial structure of the Family Values Scale from a multilevel-multicultural perspective. *International Journal of Testing, 14,* 168-192.

Byrne, B. M., & van der Vijver, F. J. R. (2017). The maximum likelihood alignment approach to testing for approximate measurement invariance: A paradigmatic cross-cultural application. *Psicothema, 29,* 539-551.

DeMars, C. (2010). *Item response theory*. Oxford University Press.

Dimitrov, D. M. (2010). Testing for factorial invariance in the context of construct validation. *Measurement and Evaluation in Counselling and Development, 43,* 121-149.

Dorans, N. J., Pommerich, M., & Holland, P.W. (Eds.) (2007). *Linking and aligning scores and scales*. New York, NY: Springer.

Epstein, J., Osborne, R. H., Elsworth, G. R., Beaton, D. E., & Guillemin, F. (2015). Cross-cultural adaptation of the Health Education Impact Questionnaire: Experimental study showed expert committee, not back-translation added value. *Journal of Clinical Epidemiology, 68,* 360-369.

Epstein, J., Santo, R.M., & Guillemin, F. (2015). A review of guidelines for cross-cultural adaptation of questionnaires could not bring out a consensus. *Journal of Clinical Epidemiology, 68,* 435-441.

Evers, A., Muñiz. J., Hagemeister, C., Hstmælingen, A., Lindley, P., Sjöberg, A., & Bartram, D. (2013). Assessing the quality of tests: Revision of the EFPA review model. *Psicothema, 25* (3), 283-291. doi: 10.7334/psicothema2013.97

Furr, R. M. (2017). *Psychometrics: An introduction*. Sage Publications.

Gagne, P., & Hancock, G. R. (2006). Measurement model quality, sample size, and solution propriety in confirmatory factor models. *Multivariate Behavioral Research, 41,* 65-83.

Gómez-Benito, J., Sireci, S., Padilla, J. L., Hidalgo, M. D., & Benítez, I. (2018). Differential item functioning: Beyond validity evidence based on internal structure. *Psicothema*, *30*(1), 104-109.

Grisay, A. (2003). Translation procedures in OECD/PISA 2000 international assessment. *Language Testing, 20* (2), 225-240.

Hagell, P., Hedin, P. J., Meads, D. M., Nyberg, L., & McKenna, S. P. (2010). Effects of method of translation of patient-reported health outcome questionnaires: A randomized study of the translation of the Rheumatoid Arthritis Quality of Life (RAQoL) instrument for Sweden. *Value in Health*, *13*(4), 424-430.

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

Hambleton, R. K., Merenda, P. F., & Spielberger, S. D. (eds.) (2005), *Adapting educational and psychological tests for cross-cultural assessment*. New Jersey: Lawrence Erlbaum Associates.

Hidalgo, M.D., & Gómez-Benito, J. (2010). Education measurement: Differential item functioning. In B. McGaw, P. Peterson, & E. Baker (Eds.), *International*

Encyclopedia of Education, 3rd edition, vol. 4, pp 36-44. Elsevier Science & Technology.

International Test Commission. (2014). ITC statement on the use of tests and other assessment instruments for research purposes. [www.intestcom.org]

International Test Commission. (2017). The ITC Guidelines for Translating and Adapting Tests (Second edition). [www.intestcom.org]

Kolen, M. J., & Brennan, R. L. (2004). *Test equating, scaling, and linking: Methods and practices* (2nd ed.). New York, NY: Springer.

Koller, M., Kantzer, V., Mear, I., Zarzar, K., Martin, M., Greimel, E., ... & ISOQOL TCA-SIG. (2012). The process of reconciliation: Evaluation of guidelines for translating quality-of-life questionnaires. *Expert Review of Pharmacoeconomics & Outcomes Research*, *12*(2), 189-197.

Leong, F.T.L., Bartram, D., Cheung, F.M., Geisinger, K.F, & Iliescu, C. (2016). *The ITC international handbook of testing and assessment.* New York: Oxford University Press.

Levin, K., Willis, G. B., Forsyth, B. H., Norberg, A., Stapleton Kudela, M., Stark, D., & Thompson, F. E. (2009). Using cognitive interviews to evaluate the Spanish-language translation of a dietary questionnaire. *Survey Research Methods, 3* (1), 13-25.

Mellenbergh, G. J. (2011). *A conceptual introduction to psychometrics: Development, analysis and application of psychological and educational tests*. The Hague, Netherlands: Eleven International.

Padilla, J. L., & Benítez, I. (2014). Validity evidence based on response processes. *Psicothema, 26*, 136-144.

Pena, E. D. (2007). Lost in translation: Methodological considerations in cross-cultural research. *Child Development, 78,* 1255-1264.

Sireci, S. G., & Rios, J. A. (2013). Decisions that make a difference in detecting differential item functioning. *Educational Research and Evaluation*, *19*(2-3), 170-187.

Terwee, C. B., Mokkink, L. B., Knol, D. L., Ostelo, R. W. J. G., Bouter, L. M., & de Vet, H. C. W. (2012). Rating the methodological quality in systematic reviews of studies on measurement properties: A scoring system for the COSMIN checklist. *Quality of Life Research, 21,* 651-657.

van de Vijver, F. J. R., & Hambleton, R. K. (1996). Translating tests: Some practical guidelines. *European Psychologist, 1* (2), 89-99.

van de Vijver, F. J. R., & Leung, K. (1997). Methods and data analysis of comparative research. In J. W. Berry, Y. H. Poortinga, & J. Pandey (Eds.), *Handbook of cross-cultural psychology (2nd ed.). Handbook of cross-cultural psychology, Vol 1: Theory and method* (pp. 257-300). Needham Heights, MA, US: Allyn & Bacon.

van de Vijver, F. J. R., & Leung, K. (2011). Equivalence and bias: A review of concepts, models and data analytic procedures. In D. Matsumoto & F. J. R. Van de Vijver (Eds.), *Cross-cultural research methods in psychology* (pp. 17-45). New York, NY: Cambridge University Press.

van de Vijver, F. J., & Poortinga, Y. H. (2004). Conceptual and methodological issues in adapting tests. In *Adapting educational and psychological tests for cross-cultural assessment* (pp. 51-76). Psychology Press.

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

Wild, D., Grove, A., Martin, M., Eremenco, S., McElroy, S., Verjee-Lorenz, A., & Erikson, P. (2005). Principles of good practice for the translation and cultural adaptation process for Patient-Reported Outcomes (PRO) measures: Report of the ISPOR Task Force for translation and cultural adaptation. *Value in Health, 8,* 94-104.

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*, 913-934.

Summary
ITC guidelines with proposed criteria: The evaluative checklist

	ITC guidelines	Assessment criteria	Not applicable*	Not acceptable	Acceptable	Excellent
PRE-CONDITION	PC1: Obtain the necessary permission from the holder of the intellectual property rights relating to the test before carrying out any adaptation.	PC1-1: If adaptation is the best option, ask the copyright owners for permission to adapt the test, even if the test is going to be used for research purposes only.				
	PC2: Evaluate that the amount of overlap in the definition and content of the construct measured by the test and the item content in the populations of interest is sufficient for the intended use (or uses) of the scores.	PC2-1: Provide theoretical and empirically-based evidence that the construct of interest is relevant to the target population.				
		PC2-2: Consider whether the meaning of the construct can be generalized across cultures, and ensure and be able to justify that translation/adaptation is preferable to creating a brand new test for the target population.				
	PC3: Minimize the influence of any cultural and linguistic differences that are irrelevant to the intended uses of the test in the populations of interest	PC3-1: If adaptation is the best option, check cultural and linguistic differences before starting the adaptation process and take them into consideration in the adapted version so as to prevent bias and to design studies to control for potential bias.				
TEST DEVELOPMENT	TD1: Ensure that the translation and adaptation processes consider linguistic, psychological, and cultural differences in the intended populations through the choice of experts with relevant expertise.	TD1-1: Form a multidisciplinary team composed of: a) professional translators who are proficient in the source and target languages (if different languages are involved) and have knowledge of the cultures involved, b) experts in the construct to be measured, c) experts in the cultures involved, and d) experts in test construction. In some cases, the same team member may be an expert in more than one of these aspects, for example, the languages and cultures, the construct and cultures, etc.				
	TD2: Use appropriate judgmental designs and procedures to maximize the suitability of the test adaptation in the intended populations.	TD2-1: Use recommended translation designs and justify the choice. Forward, backward or simultaneous translations may be used, depending on the purpose of the adaptation, the scope of the project, the number of cultures involved, and whether or not it will be necessary to compare scores of individuals from different cultures. Independent translators (at least two professionals or two teams) are involved in forward and backward translations. If a test is intended to be used cross-culturally from its inception, use simultaneous /concurrent development of multiple language versions of the test from the outset.				
		TD2-2: Have several translators that work independently and form a committee of experts to review and compare the proposed translations in order to compile judgmental reviews, resolve possible discrepancies, and produce a consensus version.				
	TD3: Provide evidence that the test instructions and item content have similar meaning for all intended populations.	TD3-1: Ensure that the instructions are clear and comprehensible, using terms that are familiar to the target population.				
		TD3-2: Ensure that the item content is clear and expressed with similar levels of commonality and difficulty in the source and target cultures. Linguistic elements that could hinder the understanding of the translated version, such as words with different meaning, double negations, etc., should be avoided. Non-linguistic elements (such as images and pictures) must be contextualized for considering the target population.				

ITC guidelines	Assessment criteria	Not applicable*	Not acceptable	Acceptable	Excellent

TEST DEVELOPMENT (Cont.)	TD4: Provide evidence that the item formats, rating scales, scoring categories, test conventions, modes of administration, and other procedures are suitable for all intended populations.	TD4-1: Try to ensure that the item format, response options, scoring rubrics, if any, and administration mode are similar in both versions.
		TD4-2: Ensure that the target population is sufficiently familiar with the procedures used (item format, response scales, scoring rubrics (if any), test conventions, and administration mode) in the adapted tests.
	TD5: Collect pilot data on the adapted test to enable item analysis, reliability assessment and small-scale validity studies so that any necessary revisions to the adapted test can be made.	TD5-1: Check the psychometric quality (item analysis, reliability, and validity) of the scores from the adapted test in a pilot sample of the target population and, based on the results, make any necessary revisions for the final version of the test. Pilot samples have to be large enough to carry out the statistical analysis involved in the pilot study.
CONFIRMATION	C1: Select sample with characteristics that are relevant for the intended use of the test and of sufficient size and relevance for the empirical analyses.	C1-1: Ensure that the target sample is large enough to carry out the necessary statistical analyses and to adequately represent the population.
		C1-2: When the focus of interest is on cross-cultural comparisons, ensure that the source and target samples are comparable for all relevant variables except for language and/or cultural background.
	C2: Provide relevant statistical evidence about the construct equivalence, method equivalence, and item equivalence for all intended populations.	C2-1: When there is interest in comparing the source and target populations, use statistical procedures to ensure that construct equivalence holds across populations
		C2-2: When there is interest in comparing the source and target populations, check for method equivalence (instrument characteristics, administration process, and sample characteristics).
		C2-3: When there is interest in comparing the source and target populations, assess DIF between the cultural groups to be compared using statistical procedures appropriate to the item format, sample size, and test dimensionality.
		C2-4: In the event that DIF is detected at meaningful levels, carry out analyses to understand the reasons for the DIF (e.g., linguistic or method effects) across cultures.
	C3: Provide evidence supporting the norms, reliability and validity of the adapted version of the test in the intended populations.	C3-1: Ensure that the type of reliability indicators reported is adequate for the type of test, using adequate statistical analysis and sample sizes. The obtained values must be satisfactory and the standard error of measurement must be reported.
		C3-2: Provide validity evidence consistent with the intended use of the test scores, using adequate statistical analysis and sample sizes.
		C3-3: Ensure and verify that the norms are adequate for interpreting the test scores of the target population.
	C4: Use an appropriate equating design and data analysis procedures when linking score scales from different language versions of a test.	C4-1: When cross-cultural/cross-lingual assessment is the objective, and comparability of scores across groups is necessary but some items are functioning differentially, use appropriate linking designs and data analysis procedures before comparison.

	ITC guidelines	Assessment criteria	Not applicable*	Not acceptable	Acceptable	Excellent
ADMINISTRATION	A1: Prepare administration materials and instructions to minimize any culture- and language-related problems that are caused by administration procedures and response modes that can affect the validity of the inferences drawn from the scores.	A1-1: For all administration materials and instructions the requirements specified in the development guidelines have been checked (TD3 to TD5). The experience accumulated when administering the original version of the test in the source population should be taken into account to prevent possible administration problems in the target population.				
	A2: Specify testing conditions that should be followed closely in all populations of interest.	A2-1: When cultural comparisons are of interest, ensure that the testing conditions (administration mode, time restrictions, information about the test purpose, etc.) are standardized across groups. If changes are necessary, data should be collected to evaluate the possible impact of different testing conditions.				
		A2-2: Ensure that the interviewers or test administrators have the credentials required for the type of test to be administered. Test administrators should submit a signed pledge to conduct their activities in accordance with the code of ethics and principles of professional practice established by the relevant national professional associations and bodies.				
SCORE SCALES AND INTERPRETATION	SSI1: Interpret any group score differences with reference to all relevant available information.	SSI1-1: When score comparisons are justified on the basis of measurement invariance analysis, consider a number of interpretations of cross-cultural differences, taking into account the information that has been systematized and documented in PC3-1 (G3, C4) regarding cultural and linguistic distance. To understand the differences in the observed scores, the role of these variables (e.g., religiosity, individualism, different response tendencies) should be considered.				
	SSI2: Only compare scores across populations when the level of invariance has been established on the scale on which scores are reported.	SSI2-1: To compare individual scores of people belonging to different cultures, and/ or mean scores across cultures, ensure that measurement equivalence (a.k.a. lack of DIF) is assessed and supported, at least for a meaningful number of items.				
DOCUMENTATION	Doc-1: Provide technical documentation of any changes, including an account of the evidence obtained to support equivalence, when a test is adapted for use in another population.	Doc-1-1: Create a number of documents and make them accessible to relevant stakeholders, providing information about the 8 issues listed in the evaluative checklist.				
	Doc-2: Provide documentation for test users that will support good practice in the use of an adapted test with people in the context of the new population.	Doc-2-1. Make sure that the materials and documentation which accompany the test (e.g., the test manual) are clear (instructions, description of the scope of application, practical examples of its use, etc.) so as to ensure that the test is adequate for the intended population, that the test administration is standardized, and that scores are interpreted adequately (see administration and scoring sections).				

* If one or more criteria are not applicable (for example, SSI2-1, when the purpose of the adaptation is not to compare scores of individuals belonging to different cultures, or C2-4, when items are tested for DIF and they show no DIF) this must be explicitly justified. When there is not enough information to judge whether a criterion is acceptable or not (and when this information is relevant and the criterion is applicable given the purpose of the adaptation), then the achievement criterion would be "Not acceptable".