Psicothema

Psicothema 2020, Vol. 32, No. 3, 390-398
doi: 10.7334/psicothema2019.306

ISSN 0214 - 9915 CODEN PSOTEG
Copyright © 2020 Psicothema
www.psicothema.com

International Test Commission guidelines for test adaptation: A criterion checklist

Ana Hernández[1], María Dolores Hidalgo[2], Ronald K. Hambleton[3], and Juana Gómez-Benito[4]

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

Abstract

Background: To improve the quality of test translation and adaptation, and hence the comparability of scores across cultures, the International Test Commission (ITC) proposed a number of guidelines for the adaptation process. Although these guidelines are well-known, they are not implemented as often as they should be. One possible reason for this is the broad scope of the guidelines, which makes them difficult to apply in practice. The goal of this study was therefore to draw up an evaluative criterion checklist that would help test adapters to implement the ITC recommendations and which would serve as a model for assessing the quality of test adaptations. **Method:** Each ITC guideline was operationalized through a number of criteria. For each criterion, acceptable and excellent levels of accomplishment were proposed. The initial checklist was then reviewed by a panel of 12 experts in testing and test adaptation. The resulting checklist was applied to two different tests by two pairs of independent reviewers. **Results:** The final evaluative checklist consisted of 29 criteria covering all phases of test adaptation: planning, development, confirmation, administration, score interpretation, and documentation. **Conclusions:** We believe that the proposed evaluative checklist will help to improve the quality of test adaptation.

Keywords: Test translation, test adaptation, guidelines, International Test Commission, evaluative checklist.

Resumen

Directrices de la Comisión Internacional de Test para la adaptación de test: un listado de verificación. **Antecedentes:** la Comisión Internacional de Test (ITC) propuso una serie de directrices para mejorar la calidad de la traducción y adaptación de los test y, consecuentemente, mejorar la comparabilidad de las puntuaciones a través de distintas culturas. Aunque estas directrices son bien conocidas, no se aplican tan frecuentemente como sería deseable. Este trabajo propone un listado de verificación de los criterios de cumplimiento asociados a las directrices de la ITC, que faciliten su implementación y sirvan de modelo para evaluar la calidad de las adaptaciones realizadas. **Método:** cada directriz de la ITC se operacionalizó a través de distintos criterios. Para cada criterio se propusieron niveles de aceptabilidad y excelencia. El listado inicial propuesto fue revisado por un panel de 12 expertos en el área de medición y adaptación de test. La versión resultante fue aplicada a dos test por dos pares de revisores independientes. **Resultados:** el listado final de verificación del grado de cumplimiento de las directrices consistió en 29 criterios que cubren todas las fases del proceso de adaptación de un test: planificación, desarrollo, confirmación, administración, interpretación de las puntuaciones y documentación. **Conclusiones:** se espera que el listado propuesto ayude a mejorar la calidad de los test adaptados.

Palabras clave: traducción de test, adaptación de test, directrices, Comisión Internacional de Test, listado de verificación.

Standardized psychological and educational tests are essential tools for professionals across clinical, educational, and work and organizational settings (e.g., Miller & Lovler, 2018; Muñiz et al., 2001; Muñiz, Hernández, & Fernández-Hermida, 2020). They are used to make diagnostic decisions, hire or promote individuals, and assess students' competence level, etc. Tests are also essential for research purposes. Consequently, the use of tests has important consequences, not only for the individuals tested but also for knowledge generation (e.g., ITC, 2014).

When there is a need to measure a specific trait or competence, researchers often translate and adapt tests that have been developed and validated in other cultures, rather than developing a new test from scratch (e.g., Hambleton & Lee, 2013). Adaptation not only saves researchers time and other resources but also facilitates cross-cultural comparisons. This is a crucial feature given that individuals are nowadays embedded in multicultural and multilingual contexts (Duarte & Rossier, 2008).

The number of test adaptations has risen exponentially in recent decades and it has now become a common practice (Epstein, Santo, & Guillemin, 2015; Hambleton & Zenisky, 2011). In this context, the quality of the translation/adaptation process is crucial for ensuring the validity and usefulness of the adapted test. Far from being a trivial task, adaptation is a rigorous process whose aim is to maintain equivalence in content and cultural meaning between the original and the translated/adapted test, thus fostering the comparability of scores across individuals from these different cultural groups.

To provide methodological guidance for the adaptation process and to improve its quality, the International Test Commission

Received: October 7, 2019 • Accepted: May 8, 2020
Corresponding author: Ana Hernández
Facultad de Psicología
Universitat de València
46010 Valencia (Spain)
e-mail: Ana.Hernandez@uv.es

(ITC) began in 1994 to draw up a number of guidelines for test translation and adaptation (see Hambleton, 1994, 1996; Muñiz & Hambleton, 1996; van de Vijver & Hambleton, 1996). These guidelines were consistent with more general ones such as the Standards for Educational and Psychological Testing (AERA, APA & NCME, 1999, 2014). The first edition of the guidelines was published by the ITC in 2005 (ITC, 2005; Hambleton, 2005). To respond to subsequent advances in technology and testing practices, these guidelines were later reviewed and a second edition was published in 2017 (ITC, 2017; see also ITC, 2018) (for a preliminary publication of this second edition in Spanish, see Muñiz, Elosua, & Hambleton, 2013). This second edition includes 18 guidelines organized into six sections: 1) Pre-Condition, which includes three guidelines concerning decisions that need to be made before starting the translation/adaptation process; 2) Test Development, with five guidelines focused on the actual process of test adaptation; 3) Confirmation, with five guidelines referring to the compilation of empirical evidence that supports the equivalence, reliability, and validity of the original and adapted versions; 4) Administration, with two guidelines whose aim is to ensure that the administration process does not affect the reliability, validity, and fairness of the scores obtained; 5) Score Scales and Interpretation, with two guidelines aimed at ensuring that comparisons among people from different cultures are made only at the level of measurement equivalence achieved by the results of empirical analyses and that the differences are interpreted accurately; and 6) Documentation, with two guidelines offering recommendations on how the adaptation process should be documented to provide transparent evidence about the quality of the adaptation.

Although the ITC guidelines have been translated into 13 languages (Rios & Sireci, 2014), and despite the fact that a Google Scholar search in 2016 indicates that they have been mentioned more than 18,000 times (see Iliescu, 2017), the guidelines have not been followed as often as these numbers suggest. A systematic review of test adaptation processes around the world showed that even if studies frequently mention the ITC guidelines, the specific guidelines that underpin the work and the specific steps in which such guidelines are addressed are often missing (Valdivia Vázquez, 2014). In a similar vein, Rios and Sireci (2014) reviewed a number of papers to check whether the methodologies related to test adaptation had improved with the publication of the ITC guidelines. Although they only focused on papers where test adaptation and cross-cultural comparison were carried out in the same study, their results were disappointing: The majority of studies reviewed did not follow the recommendations proposed by the ITC. Rios and Sireci concluded that there is a need for better dissemination mechanisms to improve the implementation of the guidelines. While we agree that dissemination is crucial, our aim in the present study is to go a step further. Our rationale for doing so is that, in our view, the broad scope of the ITC guidelines, which to some extent is necessary, also leads to some ambiguity and can make the guidelines difficult to apply in practice.

Although the second edition of the ITC guidelines (ITC, 2017) has made important progress in this direction by offering suggestions for implementing the guidelines in practice, we aim to build on this by proposing a checklist of criteria linked to the different guidelines. For each criterion, we also suggest acceptable and excellent levels of accomplishment. These criteria, and the corresponding accomplishment levels, should be a useful tool not only for test adapters but also for test users and those tasked with reviewing adapted tests or research papers involving an adaptation process.

The evaluative criterion checklist, including the corresponding requirements for the acceptable/excellent levels of accomplishment, is available at https://www.cop.es/pdf/ITC-guidelines-for-test-adaptation-CRITERION-CHECKLIST.pdf. In this paper we describe the process of developing the evaluative checklist and the main criteria and accomplishment levels proposed. We will conclude with some final remarks about future challenges.

Method

Based on the recommendations made in the ITC Statement on Test Adaptation Guidelines (ITC, 2017), a literature review of test adaptation and cross-cultural research (e.g., Hambleton & Lee, 2013; Hambleton, Merenda, & Spielberger, 2005; Iliescu, 2017; Matsumoto & van de Vijver, 2011), and several test evaluation models and checklists such as the EFPA (European Federation of Psychological Associations) Test Review Model (Evers et al., 2013) and the COSMIN (Consensus based Standards for the selection of health status Measurement Instruments) checklist (see Terwee et al., 2012), each ITC guideline was operationalized through a number of criteria. For each criterion, acceptable and excellent levels of accomplishment were proposed. The original proposal, comprising 34 criteria and accomplishment levels, was reviewed by a panel of experts on testing and test adaptation.

A number of strategies were used to identify potential experts from around the world. Specifically, we checked who had comprised the committees that had participated in the development of the ITC guidelines, we reviewed authorship of key publications on test adaptation, we identified ITC members by means of the ITC newsletter, and, finally, we relied on personal recommendations. In order to have a broad view, we contacted 27 experts from across five continents: 11 from the Americas (eight from North America and three from central and south America), three from Africa, seven from Europe, five from Asia, and one from Oceania. They were each invited to participate in the review via email. After explaining the purpose of the study and sending the proposed checklist, we asked them to give their opinion on the following specific issues:

1) Are all the proposed criteria appropriate for operationalizing the specific ITC guidelines to which they are linked? Should any of the proposed criteria be deleted or changed?
2) Are there any important criteria missing that should be included to facilitate the operationalization and implementation of the guidelines?
3) Are the proposed levels of accomplishment (acceptable/excellent) adequate or are they too demanding or too lenient? Should these acceptable/excellent levels be changed for any of the criteria?
4) Do you have any other suggestions for improving the proposed checklist?

Although more than 60% of the experts contacted expressed a willingness to participate in the project, only 44% sent us their feedback after several reminders. The list of final respondents is provided in Table 1. The feedback received was reviewed by the authors. The most frequent concerns of the reviewers had to do with the difficulty of grading the acceptable/excellent levels for some of the criteria, depending on, for example, the target population or the decisions to be made with the scores obtained after adaptation. After analyzing and discussing the feedback provided, some of the criteria were eliminated or integrated with others, while others (as well as the acceptable/excellent levels of accomplishment) were rewritten. We describe the main changes made in the Results section.

After incorporating the experts' feedback, we assessed the utility and validity of the resulting checklist. Specifically, two independent pairs of reviewers applied the checklist to 1) the Programme for International Student Assessment (PISA) project, an exemplary project for guiding test adaptation practices according to the ITC (2017), and 2) the Spanish adaptation of the State-Trait Anxiety Inventory (STAI) (Buela-Casal, Guillén-Riquelme, & Seisdedos, 2015). The reviewers were four academic experts in psychometrics with 8, 5, 15, and 25 years of experience, respectively. A more senior and a more junior reviewer independently assessed each test and examined the degree to which the proposed criteria had been achieved, following which inter-rater agreement was computed. They also provided feedback about the difficulties they had encountered when applying the criterion checklist, which helped to explain the observed disagreements. Further modifications to the proposed checklist were made based on this information. The results from application of the checklist to the PISA project and the STAI adaptation are presented in the following section.

Results

The final evaluative checklist, after having integrated the experts' feedback, consisted of 29 criteria which serve to operationalize the ITC guidelines for test adaptation across all the phases considered: planning, development, confirmation, administration, score interpretation, and documentation. The specific criteria and the guidelines to which they refer are presented in Table 2. The guidelines incorporate the feedback received from the experts and the reviewers who applied the checklist to the PISA project and the STAI adaptation.

Table 1
List of experts on testing and test adaptation

Experts	Affiliation
Dave Bartram	Independent consultant and University of Kent (honorary professor) (UK)
Fanny Cheung	Chinese University of Hong Kong (China)
Francesca Chiesi	University of Florence (Italy)
Brian F. French	Washington State University (USA)
Jaques Gregoire	Catholic University of Louvain (Belgium)
Musab Hayatli	Capstan (USA)
Dragos Iliescu	University of Bucharest (Romania)
Tania Moreira-Mora	Costa Rica Institute of Technology, Open State University (Costa Rica)
José Muñiz	University of Oviedo (Spain)
Stephen Stark	University of South Florida (USA)
Alina A. Von Davier	ACTNext by ACT (American College Testing) (USA)
Solange Wechsler	Pontifical Catholic University of Campinas (Brazil)

Note: Permission to make their names public was obtained from all the experts. Names are listed alphabetically by surname

Experts' feedback

Some of the concerns were general and not linked to specific guidelines. Accordingly, we grouped some of the initial criteria and incorporated the following general suggestions made by the experts: a) explicitly link the criteria to the specific guidelines by means of clear codes and numbers (two experts), b) contextualize the criteria so that they refer to 'tests' in the broad sense of the word (including scales, questionnaires, etc.) and to all types of assessments (e.g., personality, competences, skills, attitudes, achievement, intelligence, knowledge, etc.) (two experts), c) improve the reference section and base achievement criteria on scientific references (three experts), d) review the wording so as to refer to the properties of the scores obtained and not the test itself (one expert), e) provide a link to the ITC guidelines (one expert), f) make explicit reference to the need to comply with the ethical code for research involving human beings (one expert), and g) provide examples of cases when adaptation would not meet the acceptance criterion and/or would not be applicable (one expert). Finally, the most common general concern (four experts) had to do with the apparent arbitrariness or ambiguity of the requirements that differentiated the acceptable and excellent achievement criteria. To address this concern, we revised the initial proposed requirements based on the relevant literature (and provided references that supported our proposal). However, we decided not to follow some of the general suggestions made by the experts. For example, we did not include a "yes/no" checklist for the guidelines themselves or a dictionary of psychometric terms because both are provided in the second edition of the ITC Guidelines (ITC, 2017; pp. 37-41). Neither did we include an appendix illustrating psychometric methods related to test adaptation because this goes beyond the purpose of the study.

Other concerns raised by the experts referred to specific criteria. Because of space limitations we cannot describe all these issues and the way in which we addressed them. Instead, we indicate the most common concern for each of the six broad categories that are differentiated in the guidelines. Regarding the *pre-condition guidelines*, the most frequent concern was related to permission from the copyright holder and the need to include additional aspects (e.g., acceptance of changes to item content, format, etc., if necessary). For the *development guidelines*, the most criticized criterion in the initial proposal was TD1-1, referring to the composition of the team involved in the translation/adaptation. Specifically, three experts asked us to suggest a specific number of participants in the team and to make explicit the need to document the qualifications of the professionals involved. These suggestions were incorporated into the final TD1-1 criterion with the achievement criteria. With respect to the *confirmation guidelines*, the most common concern had to do with C2-3 (referring to DIF assessment across cultural/linguistic groups) and C3-2 (providing validity evidence). We had initially proposed the combined use of different procedures for detecting DIF and the collection of different types of validity evidence to achieve the excellent criterion for C2-3 and C3-2, respectively. However, three experts pointed out that to detect DIF it would be better to choose the most suitable method by considering the type and size of data analyzed. As for assessing validity, the experts suggested the need to make explicit that the collected evidence

Table 2

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.				

Table 2
ITC guidelines with proposed criteria: The evaluative checklist

	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.				

Table 2
ITC guidelines with proposed criteria: The evaluative checklist

	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).				

* The requirements for concluding that the level of accomplishment is acceptable or excellent are described in the document that can be downloaded at https://www.cop.es/pdf/ITC-guidelines-for-test-adaptation-CRITERION-CHECKLIST.pdf. 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"

should be adequate for supporting the intended inferences made from test scores. We agreed with these suggestions and changed the two criteria accordingly. Regarding the *administration guidelines*, the only comment, apart from a clarification request, related to criterion A2-2 (referring to interviewers and test administrators). One of the experts stressed the importance of explicitly stating in the acceptable/excellent achievement criteria the need to submit a signed pledge that the activities will be conducted in accordance with the code of ethics and principles of professional practice. We followed this suggestion and included this information in the achievement criteria. For the *scoring and interpretation guidelines*, an initial criterion (now integrated within criterion SSI2-1, which refers to score comparisons) was the one that received most comments (four experts raised concerns). The main problem was that we initially suggested a specific percentage of invariant items to make score comparisons. However, the experts did not agree with these percentages because meaningful comparisons may also depend on the magnitude of DIF effects, direction, test length, etc. We accepted this point and removed the specific percentages initially suggested, focusing exclusively on the comparison based on invariant items. The experts also suggested providing specific references (e.g., Dimitriv, 2010), and these are now included. Finally, regarding the *documentation guidelines*, there was only one comment related to the need for a clearer distinction between the amount of information required to judge the documentation as excellent or acceptable. In the final documentation criteria, we now distinguish between fully comprehensive and sufficient information for judging the quality of the adaptation process.

Reviewers' feedback regarding adequacy of the checklist and test adaptation

For the STAI, the inter-rater agreement was 93.1%, with some disagreements over the pre-condition guidelines criteria. For PISA, although achievement levels were higher, the inter-rater agreement was lower (65.55%) and especially problematic for the scoring and interpretation and documentation guidelines. These disagreements were discussed in two consensus sessions (one per test) with the participation of the two reviewers involved and a member of our research team. The goal of these sessions was to determine the reasons for disagreement and to reach a consensus. It is important to note here that PISA documentation is very extensive and spread across different documents, and thus, although some of the disagreements were related to problems in the specific criteria proposed in the checklist, others were due to the fact that one of the reviewers had overlooked specific information in the extensive documentation. After the consensus sessions, agreement was 100%, and the suggestions made by the reviewers allowed us to improve and clarify some of the criteria. Specifically, the discussion of the criteria during the consensus sessions led us to introduce changes in 8 of the 29 criteria (TD2-1, TD4-2, TD5-1, C2-1, A2-2, SSI1-1, DOC1-1, and DOC2-1). The reviewers also made suggestions regarding a further six criteria (TD4-1, C2-2, C2-4, C3-2, A1-1, and A2-1)[1]. Most of the suggested changes referred to clarifications. The most important substantive change had to do with TD2-1 (referring to the translation/adaptation design). The proposed criterion for 'excellent' required that several translation designs had been combined to obtain the initial version of the adapted test. However, the test adaptation reviewers agreed that this was too exacting. Given the drawbacks of backward translation (the target language version of the test is not reviewed, which can produce a rather awkward target language version of the test) (ITC, 2017), we added forward translation to the criterion for excellence. Note that this does not refer to the "naïve forward translations" mentioned by Iliescu (2017), but rather forward translation done by independent teams, with discrepancies being reconciled into a single version by a third independent translator or expert panel — also called "double-translation and reconciliation" (ITC, 2017).

The final criteria presented in Table 2 serves as a checklist in which the level of accomplishment for each criterion can be assessed by considering the requirements for reaching the level of acceptability/excellence, as set out in the full document that accompanies this paper. The full document can be downloaded at https://www.cop.es/pdf/ITC-guidelines-for-test-adaptation-CRITERION-CHECKLIST.pdf

The conditions and requirements for deciding the level of accomplishment for each of the proposed criteria have been established by considering the recommendations made in the ITC guidelines (ITC, 2017), the test adaptation and cross-cultural research literature (e.g., Hambleton & Lee, 2013; Hambleton et al., 2005; Iliescu, 2017; Matsumoto & van de Vijver, 2011), and different test evaluation models and checklists (such as EFPA and COSMIN) (for more detail, see the full document). For example, criterion TD1-1 (i.e., the first criterion linked to the first guideline in the Test Development (TD) section) specifies the need to "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…". To achieve the 'excellent' level, the multidisciplinary team must be larger and more heterogeneous, whereas a smaller team would be sufficient to achieve the 'acceptable' level. However, when the test is going to be translated and adapted from one language to another, a minimum of two qualified translators are required. A translation made by a bilingual student with the support of an expert researcher on the topic in question would not be enough to ensure the quality and linguistic equivalence of the translated version according to the ITC guidelines. In addition, and to ensure team quality, the procedure for selecting experts, as well as their qualifications and experience, must be documented.

In the event that one or more criteria are not applicable in a specific situation this must be explicitly justified. For example, SSI2-1 (the first criterion of the second guideline referring to Score Scales and Interpretation) deals with the score comparison of individuals belonging to different cultures. If the purpose of a particular adaptation is not to compare scores of individuals belonging to different cultures, this must be explicitly stated and justified.

Discussion

We live in multicultural and multilingual societies, and as a result test adaptations and cross-cultural comparisons based on those tests are now a common practice. When adapting tests it is crucial to follow good practices in the adaptation process so as to prevent errors and ensure the comparability of test scores. In addition, it is essential to verify that the adapted test is a legitimate and valid version of the original. In this regard, the ITC guidelines for test translation and adaptation are an excellent benchmark for obtaining high-quality test adaptations. However, although these guidelines are well known in applied contexts (Iliescu, 2017), and despite examples that show adherence to them (e.g., Bakker, Ficapal-Cusí, Torrent-Sellens, Boada-Grau, & Hontangas-Beltrán, 2018; Lourido, Arce, & Ponte, 2018; Swami & Barron, 2019), they have not, in practice, been implemented (or documented) as often as would be desirable (see Valdivia Vázquez, 2014; Rios & Sireci, 2014). This is particularly the case for some of the guidelines. For example, Rios and Sireci (2014) have shown that whereas adaptation studies frequently provide information about guidelines such as those related to sample equivalence or construct equivalence, they often ignore other relevant aspects such as the administration or instrument equivalence (e.g., response styles or differential familiarity with response procedures). This suggests that steps need to be taken to increase the implementation of ITC guidelines.

To this end, the present study offers an evaluative checklist comprising a number of criteria that are linked to each of the ITC guidelines. The proposed checklist is intended to help researchers from across different fields to implement the guidelines when adapting tests and evaluating the quality of test adaptation. However, when using the evaluative checklist, researchers should bear in mind that adaptation is a complex process. As Iliescu (2017) points out, the possible variations in the adaptation process are considerable and depend on both substantive and contextual variables. Accordingly, the researcher must make specific choices in each phase. Furthermore, the extent to which a given phase is crucial or not depends on the particular purpose of test adaptation (for example, depending on whether scores of individuals belonging to different cultures are going to be compared or not). Our proposed criterion checklist thus aims to help researchers in navigating this complex process. Specifically, in developing the checklist we have sought to draw researchers- attention to a number of issues: 1) each ITC guideline must be taken into account unless it is not applicable in light of the purpose of the adaptation in a particular study (this non-applicability must be justified)¸ 2) there are key criteria that need to be considered to ensure and verify that the guidelines are followed, 3) there are minimum requirements that must be fulfilled when adapting a test, and the scope of these requirements should be increased when important decisions are made based on test scores, and 4) it is necessary to justify the choices made in every step of the process, providing documentation that supports the quality of the process and the results.

We believe that the proposed criteria, and the corresponding accomplishment levels, will be a useful tool not only for test adapters but also for cross-cultural researchers, test users, and test and research reviewers. However, it is important to bear in mind that our proposed evaluative criterion checklist cannot be used independently of the ITC guidelines themselves. In fact, the checklist is meant to complement the recommendations set out by the ITC (2017).

Our efforts to facilitate the implementation of the ITC guidelines are in line with other actions and strategies, and we particularly wish to highlight the value of the following: a) the recommendations made by van de Vijver and Leung (2011) regarding how to prevent different types of bias and, when necessary, deal with them; b) the review form proposed by Hambleton and Zenisky (2011), focusing specifically on item translation and adaptation; c) the exhaustive checklists proposed by Iliescu (2017), comprising a large number of questions relating to all phases of the adaptation process; and d) the second edition of the ITC guidelines themselves, which includes suggestions for implementing the guidelines (ITC, 2017). Each of these contributions has its own unique strengths. What makes our evaluative checklist unique is that it explicitly links each ITC guideline to one or more specific assessment criteria, and also that it specifies a level of accomplishment (acceptable vs. excellent) for each guideline.

Given the importance of test adaptation, it is crucial to ensure the quality of the adaptation process, and in this respect it is essential to adhere to the ITC guidelines. Our evaluative criterion checklist is a step in this direction. However, we acknowledge that the levels of accomplishment in our checklist may be criticized for not being fully objective. In our view, the requirements that define each level should not be interpreted too rigidly, because the level of requirement may depend, among other things, on how test scores will be used and interpreted, as well as on the potential consequences of the decisions made on the basis of these scores. Our recommendations are based on our experience, on a review of the relevant literature, the feedback provided by the experts who assessed the initial version of the proposed checklist, and the feedback provided by the test adaptation reviewers who applied the checklist. Further studies are necessary to confirm the accuracy of the established levels of accomplishment. In this respect, the challenge will be to refine the proposed criteria and reach a broad consensus about their utility. In a different vein, future research should also examine whether the number of studies that follow the ITC guidelines in practice increases as a result of the proposed evaluative checklist, and whether the quality of the adaptation process improves based on the reported documentation. Documenting the adaptation process is, of course, crucial for making informed decisions about its quality.

We hope that the proposed evaluative checklist, together with the other aforementioned contributions (Hambleton & Zenisky, 2011; Iliescu, 2017; ITC, 2017; van de Vijver & Leung, 2011), will lead to greater implementation of the ITC guidelines and to improvements in the quality of test adaptation and cross-cultural research.

Acknowledgments

The authors would like to thank all the participating experts (see Table 1) for reviewing the initial evaluative checklist and for the valuable feedback provided. However, the final criteria and accomplishment levels that feature in the proposed evaluative checklist are solely the authors' responsibility.

The authors also thank Inés Tomás and Laura Galiana (University of Valencia) and Georgina Guilera and Ángela Berrío (University of Barcelona) for their work in applying the criterion checklist and providing suggestions for improving it.

Notes

[1] More details about the specific changes can be obtained from the first author upon request.

References

American Educational Research Association, American Psychological Association, and National Council on Measurement in Education (1999). *Standards for educational and psychological testing*. Washington, DC: American Psychological Association.

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

Bakker, A. B., Ficapal-Cusí, P., Torrent-Sellens, J., Boada-Grau, J., & Hontangas-Beltrán, P. M. (2018). The Spanish version of the Job Crafting Scale. *Psicothema, 30*(1), 136-142. doi:10.7334/psicothema2016.293

Buela-Casal, G., Guillén-Riquelme, A., & Seisdedos, N. (2015). *Cuestionario de Ansiedad Estado/Rasgo*. Madrid: TEA Ediciones.

Duarte, M. E., & Rossier, J. (2008). Testing and assessment in an international context: Cross-and multi-cultural issues. In J. A. Athanasou & R. Van

Esbroeck (Eds.), *International handbook of career guidance* (pp. 489-510). Dordrecht: Springer. doi:10.1007/978-1-4020-6230-8_24

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*(4), 435-441. doi:10.1016/j.jclinepi.2014.11.021

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*, 283-291. doi:10.7334/psicothema2013.97

Hambleton, R. K. (1994). Guidelines for adapting educational and psychological tests: A progress report. *European Journal of Psychological Assessment, 10,* 229-244. doi:10.1027//1015-5759.15.3.270

Hambleton, R. K. (1996). Adaptación de tests para su uso en diferentes idiomas y culturas: fuentes de error, posibles soluciones y directrices prácticas [Adapting tests for use in different languages and cultures: Sources of error, possible solutions, and practical guidelines]. In J. Muñiz (Coord.), *Psicometría* (pp. 207-238). Madrid: Universitas.

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., & Lee, M. K. (2013). Methods for translating and adapting tests to increase cross-language validity. In D.H. Saklofske, C.R. Reynolds, V.L. Schwean (Eds.), *The Oxford handbook of child psychological assessment* (pp. 172-181). New York, NY: Oxford University Press. doi:10.1093/oxfordhb/9780199796304.013.0008

Hambleton, R. K., & Zenisky, A. (2011). Translating and adapting tests for cross-cultural assessments. In D. Matsumoto & F. J. R. van de Vijver (Eds.), *Cross-cultural research methods in psychology* (pp. 46-74). New York, NY: Cambridge University Press.

Hambleton, R. K., Merenda, P. F., & Spielberger, C. D. (Eds.) (2005). *Adapting educational and psychological tests for cross-cultural assessment*. Hillsdale, NJ: Lawrence Erlbaum Publishers. doi:10.4324/9781410611758

Iliescu, D. (2017). *Adapting tests in linguistic and cultural situations*. Cambridge: Cambridge University Press. doi:10.1017/9781316273203.

International Test Commission (2005). *International Guidelines on Test Adaptation*. Retrieved from www.intestcom.org

International Test Commission (2014). *ITC Statement On the Use of Tests and Other Assessment Instruments for Research Purposes*. Retrieved from www.intestcom.org

ITC Guidelines for Translating and Adapting Tests (Second Edition) (2017). Retrieved from www.intestcom.org

ITC Guidelines for Translating and Adapting Tests (Second Edition) (2018). *International Journal of Testing, 18*(2), 101-134, doi:10.1080/15305058.2017.1398166

Lourido, D. T., Arce, C., & Ponte, D. (2018). Adaptation of the Test of Performance Strategies Competition Subscale to Spanish. *Psicothema, 30*(1), 123-129. doi:10.7334/psicothema2017.124

Matsumoto, D., & van de Vijver, F. J. R. (2011). *Cross-cultural research methods in psychology*. New York, NY: Cambridge University Press. doi:10.1017/CBO9780511779381

Miller, L. A., & Lovler, R. L. (2018). *Foundations of psychological testing: A practical approach*. London: Sage Publications.

Muñiz, J., & Hambleton, R. K. (1996). Directrices para la traducción y adaptación de los tests [Guidelines for test translation and adaptation]. *Papeles del Psicólogo, 66,* 63-70.

Muñiz, J., Elosua, P., & Hambleton, R. K. (2013). Directrices para la traducción y adaptación de los tests: segunda edición [Guidelines for test translation and adaptation: Second edition]. *Psicothema*, *25*, 151-157. doi:10.7334/psicothema2013.24

Muñiz, J., Hernández, A., & Fernández-Hermida, J. R. (2020). Test use in Spain: The psychologists' viewpoint. *Psychologist Papers, 41*(1), 1-15. doi:10.23923/pap.psicol2020.2921

Muñiz, J., Bartram, D., Evers, A., Boben, D., Matesic, K., Glabeke, K., & Zaal, J. N. (2001). Testing practices in European countries. *European Journal of Psychological Assessment, 17*(3), 201-2011. doi: 10.1027//1015-5759.17.3.201.

Rios, J. A., & Sireci, S. G. (2014). Guidelines versus practices in cross-lingual assessment: A disconcerting disconnect. *International Journal of Testing*, *14*, 289-312. doi:10.1080/15305058.2014.924006

Swami, V., & Barron, D. (2019). Translation and validation of body image instruments: Challenges, good practice guidelines, and reporting recommendations for test adaptation. *Body Image, 31,* 204-220. doi:10.1016/j.bodyim.2018.08.014

Terwee, C. B., Mokkink, L. B., Knol, D. L., Ostelo, R. W., Bouter, L. M., & de Vet, H. C. (2012). Rating the methodological quality in systematic reviews of studies on measurement properties: A scoring system for the COSMIN checklist. *Quality of Life Research, 2,* 651-657. doi:10.1007/s11136-011-9960-1

Valdivia Vázquez, J. A. (2014). *Test adaptation activities across languages and cultures (Order No. 3640085)*. Retrieved from https://search.proquest.com/docview/1625043213?accountid=14777

van de Vijver, F. J. R., & Hambleton, R. K. (1996). Translating tests: Some practical guidelines. *European Psychologist, 1*(2), 89-99. doi:10.1027/1016-9040.1.2.89

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. doi:10.1017/CBO9780511779381.003