Routledge
Taylor & Francis Group

International Journal of Testing

ISSN: (Print) (Online) Journal homepage: https://www.tandfonline.com/loi/hijt20

The Test Adaptation Reporting Standards (TARES): reporting test adaptations

Dragos Iliescu, Dave Bartram, Pia Zeinoun, Matthias Ziegler, Paula Elosua, Stephen Sireci, Kurt F. Geisinger, Aletta Odendaal, Maria Elena Oliveri, Jon Twing & Wayne Camara

To cite this article: Dragos Iliescu, Dave Bartram, Pia Zeinoun, Matthias Ziegler, Paula Elosua, Stephen Sireci, Kurt F. Geisinger, Aletta Odendaal, Maria Elena Oliveri, Jon Twing & Wayne Camara (12 Jan 2024): The Test Adaptation Reporting Standards (TARES): reporting test adaptations, International Journal of Testing, DOI: 10.1080/15305058.2023.2294266

To link to this article: https://doi.org/10.1080/15305058.2023.2294266

Published online: 12 Jan 2024.

INTERNATIONAL JOURNAL OF TESTING
https://doi.org/10.1080/15305058.2023.2294266

Routledge
Taylor & Francis Group

The Test Adaptation Reporting Standards (TARES): reporting test adaptations

Dragos Iliescu[a,b], Dave Bartram[c], Pia Zeinoun[d], Matthias Ziegler[e], Paula Elosua[f], Stephen Sireci[g], Kurt F. Geisinger[h], Aletta Odendaal[b], Maria Elena Oliveri[h], Jon Twing[i] and Wayne Camara[j]

[a]Department of Psychology and Cognitive Science, University of Bucharest, Bucharest, Romania; [b]Department of Industrial Psychology, Stellenbosch University, Stellenbosch, South Africa; [c]University of Kent, Canterbury, UK; [d]Gravitas Consultants, Amsterdam, The Netherlands; [e]Institute for Psychology, Humboldt-Universität zu Berlin, Berlin, Germany; [f]Psychology, University of the Basque Country, Bilbao, Spain; [g]College of Education, University of MA Amherst, Amherst, MA, USA; [h]Buros Center for Testing, University of NE, Lincoln, NE, USA; [i]Pearson North America, New York, NY, USA; [j]Law School Admission Council, Newtown, PA, USA

ABSTRACT

The "Test Adaptation Reporting Standards" (TARES), or "TARES statement" was developed to alleviate the problems arising from inadequate reporting of test adaptation procedures. The TARES contains a short preamble and a checklist, that comprises an evidence-based minimum set of information for reporting in test adaptations. The TARES statement was developed by an international group of experts, under the umbrella of the International Test Commission (ITC) to support an increase in the accuracy, transparency, and usefulness of test adaptations documentation. This paper reports on the context and motivation for generating the TARES statement, describes the development process, discusses the TARES checklist structure and components, and suggests potential uses.

KEYWORDS

Assessment; outcome measures; reporting standards; test adaptation

Context and motivation for developing the TARES statement

Numerous initiatives have appeared during the past years in an effort to promote transparent and accurate reporting of research studies in health, education, and the social sciences, with the ultimate goal to enhance the

CONTACT Dragos Iliescu dragos.iliescu@fpse.unibuc.ro Department of Psychology and Cognitive Science, University of Bucharest, Sos. Panduri 90, Bucharest, Romania.

value of the research literature. These initiatives have generated what we now call "reporting guidelines"—some of the more prominent are CONSORT (for randomized trials; see http://www.consort-statement.org/), STROBE (for observational studies; https://strobe-statement.org/), PRISMA (for systematic reviews; http://www.prisma-statement.org/), CARE (for case reports; https://www.care-statement.org/), or SPIRIT (for study protocols; https://www.spirit-statement.org/). In the domain of testing and assessment, the need for more structure in test adaptations was felt to be more acute and was signaled in several ways during the past few years, for example, through the revision of the *ITC Guidelines for Translating and Adapting Tests* (ITC, 2017), as well as other influential publications (e.g., Hernández et al., 2020; Iliescu, 2017; Zeinoun et al., 2021; Ziegler, 2020). Transparency through documentation has been highlighted as an important characteristic of good research. Three reasons make transparency especially relevant for the assessment literature and specifically for test adaptations.

First, science and practice in the behavioral, social, and medical sciences rely on good measurement of outcomes, and a significant part of such measurement is undertaken based on tests that are used in languages, cultures, or contexts other than those for which they were initially intended. Thus, the quality of a significant part of research and practice hinges on the accurate and transparent reporting of derivative work (i.e., adaptation) conducted on tests. This is a fundamental issue related to the quality of measurement, and that the credibility of research depends on the possibility that others are able to critically assess the strengths and weaknesses in not only study design, conduct, and analysis, but also of how measurement was conducted.

Second, the quality of measurement is important in the evaluation of study quality when studies are included in systematic reviews and meta-analyses (e.g., AXIS; Downes et al., 2016). Many of the primary studies are based on adapted forms of the assessments employed, but not enough information is usually available to evaluate the quality of these adaptations. Without adequate and transparent reporting of measurement quality (which oftentimes actually relates to test adaptation quality), the weighting of studies included in such reviews is impossible and the consequent considerations related to the limitations in extant findings become impossible.

Third and finally, the assessment literature itself has begun during the past few years to consolidate knowledge through systematic reviews and meta-analyses focusing on assessments (e.g., Iliescu et al., 2022). Transparent reporting of the relevant test adaptations is needed when consolidating evidence on the measurement quality of any one assessment.

For all these reasons and driven by the need for more transparency in assessment research, the reporting standard presented here is in essence a

detailed set of minimal requirements for the accurate and transparent reporting and documentation of test adaptations, irrespective of the outlet these are published in (e.g., journals, test manuals, theses etc.). This detailed set of requirements is comprised in a checklist (the TARES checklist), that makes up the bulk of the TARES statement.

Aims of the TARES statement

The cultural and linguistic adaptation of a test from a source to a target culture and/or language requires a sophisticated and work-intensive scientific and technical process. Such processes are featured under different labels, among others "translation," "indigenization," "adoption," "adaptation," "transadaptation," "development," or "assembly." These standards are applicable not only for pure "test adaptations," but for the whole family of endeavors mentioned above (van de Vijver, 2015).

While the TARES may seem more applicable to "pure" reports of test adaptations, we also acknowledge that papers that are solely directed toward reporting the adaptation of a specific test are relatively rare. In many cases, authors adapt a test in order to address some substantive question of interest, e.g., to provide measurement that is necessary for the investigation of a relationship or phenomenon. We believe that the TARES is just as applicable in these cases: they *are* cases of test adaptation even though the adaptation may not be focal to the paper, and we urge authors, reviewers and editors to consider these recommendations of transparency in their papers where possible, or at the very least in electronic supplementary materials to their papers.

The elements of the Test Adaptation Reporting Standards (TARES) are prescriptive insofar as they are basic and minimal requirements that, as we consider, need to be featured in published papers and test manuals that report on the adaptation process. For example, Zeinoun et al. (2021) have observed that in some articles authors do not go into details for some of these important issues related to measurement quality, but instead prefer to refer to previous articles for the adaptation process; unfortunately in many cases those secondary articles, when published at all, appear in relatively unknown journals, were unavailable or were unclear in what and how they reported, which made it difficult to judge the quality of the original translation or adaptation work. We argue against such a practice and for the need regarding all relevant information prescribed by the TARES checklist to be reported transparently in the main manuscript and preferably in English so as to make the information available to an international audience. But we acknowledge at the same time that this may not be possible for all studies and all manuscripts: the basis of the TARES is flexibility in its application. In this

sense, the absolute minimum expectation would be an explicit statement that no information exists in a specific area that is outlined by TARES, and such lack of information should not be considered a penalty but a transparent report of what is and what is not available. This is also consistent with the processual manner in which evidence for validity is gathered over time: authors may simply not have all forms of evidence available. The TARES statement does not force evidence to be given where none exists: while acknowledging the incomplete nature of validity evidence, these guidelines still require explicit statements on what can be shown and what not.

The concept of equivalence is often at the heart of any test adaptation. Equivalence has been conceptualized in many ways and can be evaluated with much place for methodological innovation. From all the various ways in which to approach it, we have opted for the approach championed in such papers as Byrne (2015), van de Vijver and Leung (1997), or van de Vijver and Tanzer (1997), that discuss bias vs. equivalence under three large headings: construct bias, method bias, and item bias. These are reflected in components of the TARES checklist.

The TARES adheres to the contemporary approach taken by the AERA et al. (2014) *Standards for Educational and Psychological Testing* and by the various ITC *Guidelines* - e.g., the International Guidelines for Test Use (ITC, 2001), the ITC Guidelines for Translating and Adapting Tests, Second edition (ITC, 2017), or the Guidelines for Technology-Based Assessment (ITC & ATP, 2022) - that reinforce the fact that the validity of a test score interpretation is linked to the intended test use and emphasize the need to offer justification (i.e., validity evidence through data and analyses) related to the purpose of testing. We therefore urge authors to keep in mind the need to provide validity evidence of adequate translation, and in many cases of score comparability, in their reports and documentation, in order to support valid test score interpretations and uses.

Development of the TARES statement

We established the TARES initiative in early 2020, based on a proposal made by the first three authors of this paper to the Council of the International Test Commission. The proposal was accepted, and a work group was established, with all the authors of this paper.

We began our work by searching the literature for relevant material, including previous recommendations that were made in the domain of test adaptations. Of special importance in this work were the second edition of the ITC *Guidelines for Translating and Adapting Tests* (ITC, 2017), the companion criterion checklist published by Hernández et al. (2020),

the recommendations issued by Ziegler (2020), and such empirical reviews looking into the divide between good practice recommendations and actual state of the psychometric reports in test manuals and journal articles as Elosua and Iliescu (2012), and Zeinoun et al. (2021).

The work group has met several times (online only, due to the pandemic situation that made traveling difficult in 2020–2022). The resultant draft went through three rounds of extensive internal consultations: in February 2022, November 2022, and January 2023. The fleshed-out TARES statement went out for a public consultation of 30 days in April and May 2023, after which a final revision was undertaken based on the comments and suggestions that were received. The final document was accepted by the ITC Executive Board in June 2023.

Components of the TARES statement

The TARES statement comprises a short preamble and a checklist of 43 items in 11 categories. Table 1 presents the structure of the TARES and the 11 categories that we consider essential for transparent reporting of test adaptations. These categories relate to the article's title and abstract (category 1, with 2 items), the introduction (category 2, with 8 items), three categories related to the methods employed—translations (category 3, with 3 items), materials (category 4, with 6 items) and participants/sample (category 5, with 4 items) –, four categories related to the results obtained—equivalence (category 6, with 5 items), reliability (category 7, with 1 item), validity (category 8, with 6 items), and norms (category 9, with 2 items) –, discussion (category 10, with 4 items) and supplementary materials (category 11, with 2 items). The categories are numbered from 1 to 11, and inside each category the checklist items are labeled with lowercase letters, e.g., 1a, 1b, 2a, 2b, 2c etc.

Table 1. Structure of the TARES.

	Section	Items
1	Title and abstract	1a-b
2	Introduction	2a-h
	Methods	
3	Translation	3a-c
4	Materials	4a-f
5	Participants/Sample	5a-d
	Results	
6	Equivalence	6a-e
7	Reliability	7a
8	Validity	8a-f
9	Norms	9a-b
10	Discussion	10a-d
11	Supplementary	11a-b

Each of the various items is identified in the checklist by a short title and is accompanied by a short explanation and a long explanation. The two types of explanations are led in their phrasing by active verbs. The complete TARES checklist with explanations is provided in Table 2, and in the following we provide short descriptions and explanations for each of the 11 categories. Some of these 11 categories can be further grouped—for example translation, materials and participants/sample group into a "Methods" section, and equivalence, reliability, validity and norms group into a "Results" section. A visual representation of categories, their groupings and the corresponding checklist items is provided in Figure 1.

Title and abstract

The title (1a) and abstract (1b) of the paper are important in many ways. Among others, the title identifies and positions the paper, and the abstract gives preliminary information for the interested reader, but should also offer easily collectible data for reviews, meta-analyses, and other generalization research that makes use of data mining. Therefore, the TARES recommends that the title identifies, at the minimum, the name of the test, the target language and/or culture of the adaptation, and the fact that the paper presents a test adaptation. Abstracts vary significantly from one journal to another: some journals allow for longer, and others require shorter abstracts, some require structured and others narrative abstracts. Given all this variety, there are no prescriptive recommendations that the TARES can offer, but we recommend that the abstract should cover at least the source and target languages and populations, the adaptation design employed, and the sample used. If possible, it is recommended that the abstract covers also other elements of the TARES checklist, in order to make this information easy to mine for future reviews.

Introduction

The introduction to the paper is a critical part that is often ignored, with authors oftentimes looking upon their work as purely empirical, and therefore delving directly into the data and statistical treatment. The TARES recommends that authors take time to set the stage and cover in their introduction a number of eight areas (2a-2h), that help both prepare the reader and acknowledge some of the rationales, expectations, and assumptions of the authors. The TARES therefore argues for a comprehensive introduction to a test adaptation, in which various decisions are presented and justified, in the context of the language and culture to which the adaptation is targeted.

Table 2. The TARES checklist.

Category	ID	Checklist item	Short explanation	Long explanation
Title and abstract	1a	Title	Identify as a test adaptation in the title, pointing out the test name, as well as the target language/culture	Indicate in the title the fact that the paper refers to a test adaptation (or another approach of the same family, such as translation, or assembly). Identify the name of the source test. Identify the target language and/or culture.
	1b	Abstract	Provide a summary of the study (structured, if the journal accepts it)	Provide a summary of the study, in which you clearly identify the elements of this checklist, but at the very least the test, the target language or population, the design employed, the sample used, and results obtained. The structure of the abstract may differ depending on the journal.
Introduction	2a	The test	Identify and briefly present the test	Indicate the test name and acronym, test author(s), existing forms (or variations, e.g., item pools and item selection procedures), specific version(s) that is being adapted, other existing test adaptations, domain of application, existing knowledge base (e.g., test reviews by BUROS, EFPA, COP etc.).
	2b	Intended population	Identify and briefly present the intended population	Indicate the intended target population and describe it in terms of cultural, linguistic, and other characteristics that may be relevant for the test. Explain any difference in terms of the target population between source and target forms of the test.
	2c	Purpose and intended use	Identify the purpose of the assessment and briefly present the test's domain of use	Indicate the purpose of the assessment. Indicate the intended target construct and use of the test. Define the construct(s) measured. Specify resulting requirements for evaluation.
	2d	Copyright	Identify copyright status of your work	Indicate if the original test is protected by copyright, if you have obtained the necessary permissions for your derived work (or if this is not needed) and indicate the copyright status of the adapted test.
	2e	Need for adaptation	Clarify the need for the test adaptation	Explain why the test adaptation is needed in the specific target context (e.g., cultural, professional). E.g., explain why this specific test is needed, explain the gap in practice or research (i.e., are there already tests in the target culture covering the same territory).
	2f	Appropriateness of adaptation	Clarify appropriateness of test adaptation as an approach	Explain why test adaptation is the most appropriate approach compared with other approaches (e.g., test adoption, test assembly, test development).
	2g	Coverage of adaptation	Clarify if the entire original test is adapted or if any part of it were omitted from/included in the adaptation	Explain any aspect of the original test which may not have been adapted or known limitations to the adaptation that may impact scores. If the translation doesn't encompass the entire blueprint, note which subtests of domains are omitted and how this may impact equivalence.

(*Continued*)

Table 2. Continued.

Category	ID	Checklist item	Short explanation	Long explanation
	2h	Likelihood of biases	Assess likelihood of biases	Explain the expected impact of any cultural and linguistic differences on construct bias, method bias and item bias, in the intended population.
Methods				
Translation	3a	Translators	Describe translators	Indicate the number and credentials of the translators, outlining their fluency in the languages, cultures, and concepts involved.
	3b	Translation design	Describe translation design and process	Indicate the translation design that was used and justify the choice. Describe the procedure in detail, including all steps, verifications and checks conducted. Acknowledge any limitations in the approach taken.
	3c	Similarity and suitability of test components	Clarify similarity and suitability of test components across source and target use	Provide arguments for the fact that test instructions and item content, as well as item formats, scales, scoring categories, test conventions, modes of administration, and other procedures are suitable and have similar meaning for all intended populations. These arguments can refer to studies developed, and data collected during the translation process (e.g., piloting, expert focus groups), or expert judgment.
Materials	4a	Test stimuli	Explain if and how you adapted the test stimuli	Provide documentation of any changes in the test stimuli.
	4b	Instructions and scoring rubrics	Explain if and how you adapted the instructions and scoring rubrics	Provide documentation of any changes in the test instructions and scoring rubrics.
	4c	Test manual	Explain if and how you adapted the test manual	Indicate if a test manual is provided for the adaptation, how it was developed and how to obtain it.
	4d	Test reports	Explain if and how you adapted any test reports	Indicate if test reports are provided for the adaptation, how they were developed and how to access them.
	4e	Training materials	Explain if and how you adapted any training materials	Indicate if training is provided for the test adaptation and how to access it
	4f	Testing conditions	Outline testing conditions that should be followed	Specify testing conditions that should be followed closely in all populations of interest.
Participants/ Sample	5a	Sample size	Indicate the sample sizes for all analyses relevant to the adaptation	Indicate and justify the size of the sample(s) used for data analysis, including any subsamples (e.g., clinical vs. community), and pilot samples.
	5b	Sample collection procedure	Describe sample collection procedure	Indicate the procedure used for sample(s) collection.

	5c	Sample composition	Describe sample composition	Indicate the sample(s) composition in terms of relevant variables, including at the very least age and gender, and also other variables considered relevant to the test function or to equivalence testing (e.g., level of education, language proficiency). Compare the sample composition to other studies (e.g., the studies conducted for the development of the test). Discuss the impact of any deviations in sample(s).
	5d	Sample relevance	Describe sample relevance	Provide arguments for the relevance and representativeness of sample for intended analyses and intended target population and use.
Results				
Equivalence	6a	Construct equivalence	Provide evidence for construct equivalence	Provide quantitative evidence for construct equivalence (i.e., invariance analyses) using robust statistical measures; if construct bias is detected explain reasons and consequences. Provide judgmental evidence for construct generalization (i.e., that the construct associated with the original instrument is appropriate for the target population).
	6b	Method equivalence	Provide evidence for method equivalence	Provide evidence for method (sample, instrument and administration) equivalence; if method bias is detected explain reasons and consequences. This rubric should be included only if relevant - if not, then reason for lack of relevance should be provided instead.
	6c	Item equivalence	Provide evidence for item equivalence	Provide evidence for item equivalence (e.g., differential item functioning); if item bias is detected explain reasons and consequences. This rubric should be included only if relevant - if not, then reason for lack of relevance should be provided instead.
	6d	Limits of equivalence	Explain limits of equivalence	Indicate the limits of equivalence (i.e., the level of equivalence that was established). Explain consequences (e.g., for the comparability of scores at the individual and group level).
	6e	Solutions for nonequivalence	Outline solutions for nonequivalence	Provide solutions if nonequivalence is present (e.g., partial equivalence), if relevant.
Reliability	7a	Reliability	Describe evidence for reliability (e.g., measurement error) for the test score interpretation	Report indicators of reliability (e.g., reliability indices, standard error of measurement, error or precision, decision consistency) for test scores (and subscores, if applicable); explain why indicators are adequate (in terms of statistical analysis and sample size) for the type of test; compare with source form of the test.

(*Continued*)

Table 2. Continued.

Category	ID	Checklist item	Short explanation	Long explanation
Validity	8a	Validity evidence	Summarize the body of validity evidence and describe its appropriateness	Summarize the body of evidence that supports the adaptation and the use of the test for its intended purposes (i.e., validity argument drawing from evidence based on test content, response processes, internal structure, relations to other variables, and consequences of testing). Explain why the validity evidence provided is consistent with the intended use of the test scores. Indicate if evidence relies on new studies with the adapted instrument and new samples, or on transporting validity evidence. If the latter, explain.
	8b	Validity evidence based on test content	Describe evidence to support the content of the test is congruent with the testing purpose.	Describe the test content of the adapted form. Explain if and how this is different between the target and source form of the test. Provide evidence that this content is congruent with the testing purpose.
	8c	Validity evidence based on response processes	Describe any evidence that indicates the intended cognitive processes are being used by respondents who take the assessment.	Describe the intended cognitive processes of test takers. Explain if and how these are different between the target and source form of the test. Describe any evidence that these are used in the adapted form of the test.
	8d	Validity evidence based on internal structure	Describe evidence that the hypothesized dimensionality (factor structure) is consistent with the theory underlying the assessment.	State the hypothesized dimensionality (factor structure) and any alternative structures. Describe how the hypothesized dimensionality is consistent with the theory underlying the assessment. If applicable, report both item-to-factor and inter-factor relationships. Describe the evidence using appropriate confirmatory statistical methods.
	8e	Validity evidence based on relations to other variables	Describe evidence that scores from the assessment display the hypothesized relationships with other variables, in a manner consistent with the construct theory.	Indicate and describe each of the variables external to the test (e.g., constructs used for convergent and discriminant validity evidence) that were used. Indicate and explain the expected and obtained relationships of the test scores with these variables.
	8f	Validity evidence based on testing consequences	Describe evidence that the intended consequences of the assessment are occurring, and any unintended negative consequences do not occur or are minimized.	Indicate the intended consequences. Describe how they occur. Describe likely unintended negative consequences and explain their occurrence. Describe any actions taken to minimize the impact of these negative consequences.

Norms	9a	Norming procedure	Describe the statistical procedure used for norming	Indicate the statistical approach taken to norming the test; explain why it is appropriate for the intended use; indicate and justify any deviation from the original norming procedure, describe and give evidence for any effects (e.g., gender or age effects) that had an impact on norming. Clarify any differences (from the source form) in the norms reported.
	9b	Norm tables	List norm tables	Report the norm tables or provide automated formulas for deriving norms based on relevant demographics (e.g., in the case of continuous norming).
Discussion	10a	Practical relevance	Clarify practical relevance of current test adaptation	Describe the practical relevance of the test adaptation and its likely impact (e.g., size of target population, frequency of usage etc.).
	10b	Theoretical relevance	Clarify theoretical relevance of current test adaptation	If the test adaptation process uncovered conclusions that are of larger theoretical relevance, describe them.
	10c	Limitations	State limits for usage of the adapted test	Outline the limits for usage of the adapted test, in terms of test components, materials, testing conditions, score interpretation etc.
	10d	Future research	Describe future research	Clarify what specific future research should be needed on the adapted test, and what other areas of future research have emerged from the test adaptation process.
Supplementary	11a	Registration	Registration number and name of registry	Clarify if the study was pre-registered or not; if it was pre-registered offer the relevant coordinates to identify the pre-registration. If the study differs from the pre-registration, outline the differences, and the reasons for their implementation.
	11b	Funding	State sources of funding (and role of funders) and other support	State the sources of funding and the roles of funders. If the study did not receive funding, state this explicitly.

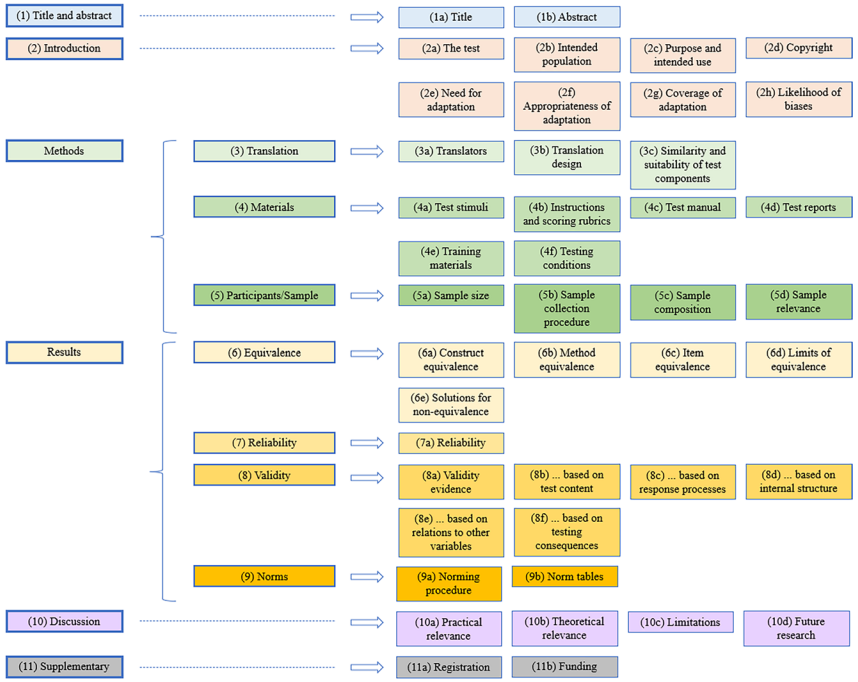

Figure 1. The structure of the TARES checklist.

- The test (2a). Authors should take their time to describe the focal test - first and foremost in terms of name and author(s), and of the specific version that is being adapted. Also, the extant knowledge base on the test should be described, e.g., mentions and descriptions of any qualitative reviews of the test, of versions or variations in the test, of previous adaptations (and their results), of empirical studies using the test (and their conclusions).
- Intended population (2b). Authors should clearly indicate the intended target population and describe it in terms of cultural, linguistic, and other characteristics that may be relevant for the test. This point is not trivial: this part is oftentimes overlooked, and it is assumed that the reader understands these characteristics. Also, a description of these characteristics should lead the authors into briefly describing and explaining any differences between the target and source populations, that would mandate and justify the test adaptation work.
- Purpose and intended use (2c). A test cannot really be adapted, nor can the adaptation itself be justified (let alone the decision taken as part of the adaptation process) if the target construct and the intended uses are not explicitly specified. The *Standards* require all

these to be described, and request that the author outright specifies the requirements for the evaluation that result from these fundamental inputs, i.e., to explain what a successful test adaptation in their specific case would look like. This point is bound to make authors of test adaptation to reason in a more lucid manner about components of their process and to decide where effort is justified and where it is less so.

- Copyright (2d). Tests are results of a creative process and as such are protected as intellectual property. In spite of a pervasive culture of free test usage in research settings, such usage is not warranted (ITC, 2014). The copyright status of the test and of the resulting adaptation should be stated, as should be the necessary permissions – both for the authors to develop the adaptation, and for interested parties to access and use the resulted adaptation.
- Need for adaptation (2e). In its famous *Guidelines for Test Use* (ITC, 2001), the ITC stated as the first recommendation in the chapter dedicated to the technical skills needed in test use (2.1.1) the need to "produce a reasoned justification for the use of tests." Indeed, a competent test user knows when to test and when not to test. Similarly, a test adaptation needs to be justified in a lucid way, by explaining why it is needed in the specific target context, or for coverage of a specific gap in practice or research.
- Appropriateness of adaptation (2f). As argued in the previous criterion, not all test adaptations cover an actual need and not all needs can be covered by a test adaptation. Sometimes the actual need can be better served through a different approach, or through a specific turn or emphasis in the adaptation process: adoption, assembly, development etc. That the adaptation is appropriate needs to be explicitly justified, as well as the fact that the specific road taken is appropriate, in light of the actual need that the adaptation is supposed to serve.
- Coverage of adaptation (2g). Authors should clarify what parts of the original test are adapted and which parts are not. It is surprising how often authors of test adaptations focus exclusively on the items, suggesting in some way that the items are the test, and ignoring instructions, scoring rubrics, norms, test materials, manuals, and other components, that are actually part of the test (Greiff & Iliescu, 2017). If the translation does not encompass the entire test, authors should explicitly note which subtests, domains or components are omitted - and they should note how this may impact the equivalence and usability of the adapted form of the test.
- Likelihood of biases (2h). Finally, the introduction should already prepare the reader for – and show that the author has dedicated

time and energy to think about – the likely biases that may show up in the test adaptation. The likelihood of biases should be presented for the three categories of construct bias, method bias, and item bias – in conjunction with the linguistic and cultural context of the target population.

Methods – translation

Discussion about the actual translation procedure is mandatory—this is oftentimes ignored in papers by authors just briefly stating that a translation was prepared, or maybe that a translation-backtranslation procedure was followed. Several authors (e.g., Iliescu, 2017) have argued convincingly for the fact that the current concentration on statistical results in test adaptation papers is detrimental to experience accumulation about how effective test adaptations can actually be achieved. Statistical analyses are post-facto, they can only show if the procedure applied was successful or not—what specifically was successful, i.e., the craft of actually developing a good test adaptation is based on activities that come before the statistical analysis and should be covered and described in detail. The "translation" category in the TARES comprises three subdomains (3a-3c), regarding the actual people who have prepared the translation, the design that was employed and the similarity and suitability of test components.

- Translators (3a). Specifically, who does the translation is important: is it the researchers themselves? Are there any other people involved, maybe a panel? How many, and what are their specializations or credentials in translation in general and in the specific focal concept or test? Was software or artificial intelligence-based translation processes used, were those approaches used in conjunction with humans or on their own? These aspects need to be briefly outlined in the manuscript. It should be described whether they worked independently or collaboratively, and which roles different individuals took if there were such roles (such as forward and backward translators).

Translation design (3b). The actual design of the translation (i.e., the actual procedure followed) also needs to be described in detail and justified. Oftentimes the translation design is only noted with a label: forward translation or (more often) backtranslation. There are many ways in which a backtranslation process could happen, and the simple label does not meet the need to be explicit about the procedure taken. The TARES recommends authors to describe in detail all the steps, verifications and

checks that were conducted, to justify the various decisions and to acknowledge any limitations in the approach taken—this information brings an important aspect of the test adaptation to the light and subjects it to open scrutiny and peer evaluation.

Similarity and suitability of test components (3c). As previously noted, test adaptations oftentimes only focus on the items of the test and ignore the other components. This TARES criterion urges authors to explicitly focus on all the other components and to provide explicit notice even when these other components do not need to be adapted and remain equivalent (e.g., item formats, scales, scoring categories, test conventions, modes of administration, and other procedures). These components oftentimes need no adaptation are therefore sometimes simply overlooked and go unmentioned; we argue for a systematic mention of every test component, explaining either that no change was operated (and why), or explaining what changes were operated (and why). Explicit arguments for the similarity and suitability of all test components should be provided—and these should not be only general and speculative, but should be, as much as possible, based on actual studies developed, and data collected during the translation process (e.g., piloting, expert focus groups).

Methods – materials

In a test adaptation, the "Materials" section of the Methods section should be about the materials of the test—of the adapted version of the test. This section of TARES imposes on the authors the obligation to explicitly address the different materials of the test, across six subdomains (4a-4f). Authors will need to explain if and how they adapted each of these six components of the test, and to provide documentation of any changes in them: test stimuli, instructions and scoring rubrics, test manual, test reports, training materials, and testing conditions.

Methods - participants/sample

The sample used for analyses is critical in any study—specifically, it is critical in terms of the suitability with the research questions. Despite this truism, in test adaptations samples are often treated with no regard for the actual intent of the adaptation process: no reference is made to how the sample size, composition, or collection procedure would be suitable for the declared intent of the test adaptation. Because significant bias can stem from the sample (a specific form of method bias, van de Vijver, 2015), this section of the TARES asks test authors to explicitly refer to participants/sample across four categories (5a-5d), namely sample size, sample collection procedure, sample composition, and sample relevance.

This also applies to samples used in field testing and pilot studies. They all need to be indicated with clarity and justified; when possible, they should be compared with other studies and arguments should be provided for their relevance—and possible bias or limitations should be explicitly acknowledged.

Results – equivalence

Equivalence is the central concept in test adaptation—the need for equivalence is mandated directly by the adaptation process: there is no logic in adapting (instead of developing) a test unless the adapted form will share some relationship with the original form, and the degree of equivalence demonstrated by the adaptation reflects that relationship. This section of the TARES encourages authors to refer to equivalence explicitly across five indicators (6a-6e): three refer to actual categories of bias/equivalence, as prescribed by van de Vijver (2015), i.e., construct equivalence, method equivalence, and item equivalence, the other two refer to the limits of established equivalence and the possible solutions for observed nonequivalence in the focal test adaptation. The three bias/equivalence rubrics should only be addressed in detail in the paper if relevant—but even if not addressed in detail, they should be mentioned. That is, if they are not considered relevant, then the reason for this lack of relevance should be discussed.

- Construct equivalence (6a). Construct equivalence can be supported in many different ways, both quantitative and qualitative – and the TARES guides authors to explicitly delve into both these two areas of evidence in providing evidence, explaining the limits of this evidence, identifying construct bias (lack of construct equivalence) and the likely consequences of such bias. Quantitative evidence should be provided in the form of measurement invariance analyses, and qualitative (judgmental) evidence should be provided for construct generalization in the target population.
- Method equivalence (6b). Method bias is possibly the most insidious form of bias out of these three. It refers to the equivalence of samples, instrumentation, and administration procedures between source and target populations. Evidence – both quantitative and judgmental should be provided for method equivalence in each of these three areas, and if any bias is detected, the consequences should be outlined.
- Item equivalence (6c). Item equivalence is oftentimes ignored: many studies consider that "good" translations automatically ensure item equivalence and focus in turn on construct equivalence. The TARES

places an obligation on researchers to explicitly address item equivalence through, for example, differential item functioning, cognitive labs, or similar analyses. Similar to other forms of bias, if item bias is detected, its reasons and consequences should be explored, although it is known that bias analyses across translated items may be difficult to interpret (Sireci et al., 2016) without special dedicated efforts to disentangle the effects of culture and language (Bader et al, 2021).

- Limits of equivalence (6d) and solutions for nonequivalence (6e). Rarely, if ever, are tests perfectly equivalent between source and target forms, in all possible types of equivalence and at all possible levels. The level at which equivalence could be established should be noted explicitly, and the limits arising from this should also be analyzed, as well as the likely consequences. The TARES also asks authors to discuss solutions in cases of nonequivalence at a specific level – partial equivalence is one such solution that could be explored, and more general solutions to this issue have also been proposed (Bauer, 2017).

Results – reliability

Reliability is one of the indicators of test quality that is now reflected in reporting of any results. The TARES also asks indicators of reliability to be given (7a), such as reliability indices, standard error of measurement, error or precision of measurement, decision consistency, etc. Supplementary, authors are asked to explain why the indicators given are adequate for the test score, and to explicitly compare them between the source and target forms of the test.

Results – validity

Validity should be reported in six categories (8a-8f). The first indicator is general and summative (validity evidence, 8a), while the other five ask authors to explicitly address each of the five sources of validity outlined by the *Standards for Educational and Psychological Testing* (AERA et al, 2014): based on test content, on response processes, on internal structure, on relations to other variables, and on testing consequences.

- Validity evidence (8a). The validity evidence gathered for the score interpretation derived from the adapted form of the test should be summarized, and an explanation should be provided about why in general the validity evidence for the adapted form is consistent with the intended use of the test scores. It should be indicated which

part of this evidence relies on new studies, developed with the adapted instrument in the target population, or is based on transporting validity evidence, which is too often the case.

- Validity evidence based on test content (8b), on response processes (8c), on internal structure (8d), on relations to other variables (8e), and on testing consequences (8f). These five subsections are from some points of view self-explanatory: evidence can be provided in each of these areas to support the use of the adapted test. At the same time, emphasis on all these areas is unusual in the testing literature (monolingual and cross-lingual), which shows, albeit in an anecdotal manner, that the current conceptualization of validity featured in the 2014 version of the AERA et al. (2014) *Standards* has not been embraced to the extent that some may have hoped. The TARES very strongly urges researchers to follow these five sources by explicitly referring to each of them – even if only to note that no evidence has been accrued for the score from the adapted test from one or another of these five sources. For example, in terms of test content, an explanation is needed regarding if and how the test content is different between the target and source form of the test, and whether this content is congruent with the testing purpose. In terms of cognitive processes, an explanation is needed regarding the intended cognitive processes of test takers and possible differences in this area between the target and source form of the test. In terms of dimensionality (factor structure), where appropriate, the fit between the hypothesized and obtained factor structures should be approached using appropriate confirmatory statistical methods and should then be discussed, also in terms of consistency with the theory underlying the test. In terms of relationships with variables external to the test (e.g., constructs used for convergent and discriminant validity evidence), the expected and obtained relationships of the test scores with these variables should be described and discussed. Finally, in terms of the intended consequences, a statement of intentions of the test and of how they should occur is needed, as is a description of likely unintended (typically) negative consequences any actions taken to minimize the impact of these.

Results – norms

Norms are addressed in two subsections, i.e., norming procedure (9a) and norm tables (9b). In terms of norming procedure, authors should indicate the statistical approach that was taken to norming the test and explain why this approach is appropriate for the intended use. Any deviations from the original norming procedure should be indicated and should be

justified. A complete description of the size and nature of the norming sample and how the sample was drawn should be provided. Also, any differences in norms between the source and the target form of the test should be clarified and discussed. In terms of norm tables, authors should report these tables, or provide a link to such tables - or provide either the formulas or links to software for deriving norms based on relevant demographics (e.g., in the case of continuous norming). If norms are not provided, an explanation should be given of why this is the case and is acceptable.

Results – discussion

The TARES suggests that the Discussion section of a manuscript reporting on a test adaptation should cover at least the four subsections of practical relevance (10a), theoretical relevance (10b), limitations (10c), and future research (10d). In terms of practical relevance, the authors should describe the likely impact of the test adaptation, for example in terms of the size of the target population, frequency of usage. If the test adaptation process uncovered conclusions that are of larger theoretical relevance, these should be described—with the understanding that this could oftentimes not be the case, as test adaptations are more often than not purely empirical endeavors with little or no concern toward theory building. Limitations in both the test adaptation process and the usage of the adapted test should be outlined—and following the previous sections of the TARES should give authors plenty of opportunity to highlight limitations for most adaptation projects. Future research directions related specifically to the focal test, or to other insights connected to the adaptation process should also be clarified.

Supplementary information

Supplementary information that could be given for a test adaptation project may be numerous—from open data and open syntaxes to many other aspects. The TARES points toward only two subsections that are mandatory, in the spirit of openness and potential conflicts of interest: registration (11a) and funding (11b). Study preregistration is certainly a best practice in modern science—and test adaptations should not be an exception from this point of view; manuscripts reporting on test adaptations should clarify if the study was pre-registered or not; and if it was pre-registered authors should offer both the relevant coordinates to identify the pre-registration and identify any deviation from the initially proposed approach. Also, sources of funding and the roles of funders should be divulged: oftentimes test adaptations are developed together with the

original test authors, or test publishers, which should be made clear for readers, in order to avoid conflicts of interest, real or perceived.

How to use the TARES statement

We recommend authors refer to TARES early during study conceptualization and design, because knowledge of the TARES guidelines and an acknowledgement of the need to address them, will likely shape the study design. Many studies fall short of the high requirements for modern test adaptations because critical components have simply been ignored in the design phase and oftentimes nothing can be done in the analysis or writing phase to overcome such gaps. From this point of view, the TARES offers guidance to researchers in structuring their studies.

We also recommend that at the least authors refer to TARES early in the writing process, not only to ensure the structure of the manuscript covers all the checklist items, but also because the long explanations offered for each of the TARES items will guide the writing of the respective sections of the manuscript. We also recommend editors to also use the checklist to help ensure the guidelines are followed and to provide authors feedback on when they are not.

Some of the items in the TARES checklist may not be applicable for some test adaptations. For example, for some tests (e.g., tests based on non-verbal items) no translation of items may be needed; in this case items 3a and 3b are not applicable. Authors are urged even in this case to not simply overlook the respective checklist items, but to make explicit notes in the manuscript regarding the non-applicability of that item. We believe that professional judgment should guide the use of the TARES checklist and that flexibility in its application is important: rubrics should be included only if relevant—but if they are not relevant in a specific case, then the reasons for such a lack of relevance should be provided instead.

This is important especially as it may easily seem that the components of the TARES statement raise the bar regarding test adaptation to an unrealistic level. Discouraging research in test adaptation or, in a more general sense, discouraging cross-cultural research, is not the intent and should not be the effect of the TARES statement. Adhering to these recommendations may motivate researchers to do more in their studies, and certainly to report more details—and if nothing more is to be reported, at least to explicitly state the nonexistence of information on a specific checklist item, in the spirit of transparency. Consistent with the flexibility that we have emphasized repeatedly throughout this paper, we believe that if the TARES were to be applied as a simple checklist on which to tick if a point was fully addressed (“yes”) or not addressed (“no”), the results would not always be positive. Authors, reviewers, editors and

readers of reports of test adaptations should rather look at the depth of the evidence provided for each checklist item on a graded scale, e.g., a) not addressed, b) somewhat addressed, c) mostly addressed, and d) addressed in full. It is conceivable that such a formal and explicit grading of each checklist item will in time begin to be required by journals, such as the PRISMA flowchart is nowadays (and has not always been) a strong requirement in the case of systematic reviews and meta-analyses, but the extent to which the TARES will be formalized is a question that will only be answered in the future, by the testing community and especially by the editors of leading journals in this field.

We also urge authors, reviewers, and editors to not look upon the TARES as a rigid prescription. The elements of the checklist should of course be addressed, but they can be addressed in many ways, in many places throughout a manuscript and in many formats: neither the order of presentation, nor the format or approach are outlined prescriptively by the TARES but should be based on the preference and personal style of the author and on the recommendations of the journal.

It is possible that journals and publishers will place some constraints on the structure or volume of text in the manuscript or in specific sections. For example, some journals may require a structured abstract, may impose a limit on the number of tables, or may require a low word limit for the manuscript. In such cases, we encourage authors to place the necessary information as required by the TARES checklist in supplementary digital materials and to refer explicitly to these materials in their manuscript. These supplementary digital materials should be reviewed just as the main paper is reviewed and should be deposited in a repository that allows for long-term storage and offers permalinks through which to access these materials (e.g., OSF, figshare or PsyArXiv, to name just a few).

Implications and limitations

The TARES statement was developed at the confluence of two streams: on one hand as a natural continuation of previous work of the ITC, especially its influential *Guidelines for Translating and Adapting Tests*, and on the other hand the general movement in science toward more transparency and openness that has culminated in similar consensus statements delineating reporting standards for various types of research designs.

The TARES will guide researchers in developing test adaptations, will inform authors in writing their manuscripts, will guide reviewers in appraising the quality of studies, and support editors in deciding if a paper should be published or not. More than anything it should help other scientists and the general public to critically assess the articles that they read. Test adaptation is a sophisticated branch of measurement

science and psychometrics, and not all readers are expected to be competent in this domain. To reach reasonably well-grounded decisions, all these potential readers have now at their disposal the TARES, which reflects the current consensus in this field, reached through an iterative process of expert work and public consultation.

We hope that editors will endorse the TARES as a recommendation for manuscripts that are submitted to journals, especially in the domains of assessment and cross-cultural psychology, where test adaptations are routinely published. We also hope that the TARES will be picked up by editors in domains that are less measurement-focused, but where authors make occasional use of test adaptations in their research designs.

Disclosure statement

No potential conflict of interest was reported by the author(s).

References

American Educational Research Association (AERA), American Psychological Association (APA), & National Council on Measurement in Education (NCME) (Eds.). (2014). *Standards for educational and psychological testing*. American Educational Research Association.

Bader, M., Jobst, L. J., Zettler, I., Hilbig, B. E., & Moshagen, M. (2021). Disentangling the effects of culture and language on measurement noninvariance in cross-cultural research: The culture, comprehension, and translation bias (CCT) procedure. *Psychological Assessment*, *33*(5), 375–384. https://doi.org/10.1037/pas0000989

Bauer, D. J. (2017). A more general model for testing measurement invariance and differential item functioning. *Psychological Methods*, *22*(3), 507–526. https://doi.org/10.1037/met0000077

Byrne, B. M. (2015). Adaptation of assessment scales in cross-national research: Issues, guidelines, and caveats. *International Perspectives in Psychology*, *5*(1), 51–65. https://doi.org/10.1037/ipp0000042

Downes, M. J., Brennan, M. L., Williams, H. C., & Dean, R. S. (2016). Development of a critical appraisal tool to assess the quality of cross-sectional studies (AXIS). *BMJ Open*, *6*(12), e011458. https://doi.org/10.1136/bmjopen-2016-011458.x

Elosua, P., & Iliescu, D. (2012). Tests in Europe: Where we are and where we should go. *International Journal of Testing*, *12*(2), 157–175. https://doi.org/10.1080/15305058.2012.657316

Greiff, S., & Iliescu, D. (2017). A test is much more than just the test itself: Some thoughts on adaptation and equivalence. *European Journal of Psychological Assessment*, *33*(3), 145–148. https://doi.org/10.1027/1015-5759/a000428

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. https://doi.org/10.7334/psicothema2019.306

Iliescu, D. (2017). *Adapting tests in linguistic and cultural situations*. Cambridge University Press.

Iliescu, D., Rusu, A., Greiff, S., Fokkema, M., & Scherer, R. (2022). Why we need systematic reviews and meta-analyses in the testing and assessment literature. *European Journal of Psychological Assessment*, *38*(2), 73–77. https://doi.org/10.1027/1015-5759/a000705

International Test Commission and Association of Test Publishers. (2022). *Guidelines for technology-based assessment*. www.InTestCom.org

International Test Commission. (2001). *The ITC guidelines on test use*. www.InTestCom.org

International Test Commission. (2014). *The 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

Sireci, S. G., Rios, J. A., & Powers, S. (2016). Comparing test scores from tests administered in different languages. In N. Dorans & L. Cook (Eds.) *Fairness in educational assessment and measurement* (pp. 181–202). Routledge.

van de Vijver, F. J. R. (2015). Methodological aspects of cross-cultural research. In M. J. Gelfand, C.-Y. Chiu & Y.-Y. Hong (Eds.), *Handbook of advances in culture psychology* (Vol. 5, pp. 101–160). Oxford University Press.

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.), (pp. 257–300). Allyn & Bacon.

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

Zeinoun, P., Iliescu, D., & El Hakim, R. (2021). Psychological tests in Arabic: A review of methodological practices and recommendations for future use. *Neuropsychology Review*, *32*(1), 1–19. https://doi.org/10.1007/s11065-021-09476-6

Ziegler, M. (2020). Psychological test adaptation and development – how papers are structured and why. *Psychological Test Adaptation and Development*, *1*(1), 3–11. https://doi.org/10.1027/2698-1866/a000002