1

What Is Test Adaptation?

ADAPTATION VS. TRANSLATION

Test translation is oftentimes used as a synonym for test adaptation. However, the two processes are very different from each other. Test translation, though immensely important, is only a part of the more encompassing test adaptation process. Test adaptation includes test translation but is much more, not only in terms of activities but also in terms of general attitude and awareness of various issues. Test translation is concerned with the transformation of a text from one language to another. Test adaptation is a thorough scientific process, and as a result is guided by the principles of the scientific method, most prominent of all being the need to offer proof for the appropriateness of said linguistic transformation not only in terms of language but also in terms of other psychometric characteristics.

Test translation is linguistically driven (content over intent). In test translation, capturing the content of the original text is more important than staying true to its intent. When speaking about translation, the focus is on the linguistic transformation of a text: We transform test items formulated in one language into test items formulated in another language. Oftentimes when doing translations, most of the attention and effort goes into the pure linguistic translation: ensuring that the target-language form is acceptable from a linguistic point of view. This shifts attention to how well the text sounds in the target language, how well it is adapted to the specific ways of formulating ideas in that new language and to its specific colloquialisms or specific forms of address. The accent will be placed on aesthetics and on personal interpretation.

Many a time, translators will approach the translation of a test as they would approach the translation of a novel or a poem: They will begin to

translate the first item, then the second, and so on up to the last, arguing as proof of work well done the fact that the target-language form sounds well, is understandable, or is reasonably close to the original.

There are two approaches a translator may take to a test *translation*. Though very different from each other, both are equally inappropriate for a proper test adaptation.

One approach would be one in which every effort is made for the translation (the new form of the test, in the target language) to be as close as possible to the original form of the test. Comparisons are constantly done with that original form, and an indicator for a well done translation is its close relation with the original.

Another approach would be one that acknowledges that languages are not isomorphic, that one cannot translate well and at the same time stand close to the original, and that, as a result, a good translation will use the intricacies of the target language to convey the meanings intended by the original text. This approach acknowledges that even the most trivial translation sometimes goes beyond factual information and may invoke, sometimes unwillingly, sound effects, emotions which are attached to specific words or other specifics of the language. An indicator for a well done translation is the easy readability of the text (e.g., test items) in the target language and the fact that it is not obviously a foreign text, but sounds indigenous.

Test adaptation is validity-driven (intent over content). In test adaptation, staying true to the intent of the original text is more important than capturing the actual original content. Indicators of a good translation, such as easy readability of the test items or aesthetic characteristics, are inappropriate in light of the objective of the translation/adaptation process. Indeed, the efficacy of any work should only be considered in light of its objectives. When translating or adapting a test from one language to another, the intent is to use an original test, which was proven valuable in its original form, in another language, culture, and context. But the assumption behind the entire process is that the new language form of the test will capitalize on all the value of the original test: If the original test has been proven valid in the source language, culture, and context we expect it to be thus also in the target language, culture, and context.

If this is the objective of the whole process, then it requires more than spurious evidence of being aesthetically pleasing in the target language. If we accept this objective as an overarching objective for the whole adaptation process, then the entire process needs to be done based on ways that ensure a reasonable chance for success (having proven efficient before),

and needs to encompass proof for the fact that the new language form of the test indeed capitalizes on the advantages of the original form.

This is the single major difference between test translation and test adaptation: Test adaptation takes responsibility for offering proof of the fact that the target-language form is close enough (equivalent, as we will see) to the source-language form, not only in language but in its intended use and consequences.

As a result, test adaptation is a veritable, work-intensive scientific process, including not only many or all activities which have been done in the initial development of the test but several that have not been considered initially. Test adaptation may sometimes be as or more labor-intensive than the initial test development process.

Test adaptation includes decisions about whether the test can measure the same construct in the new language and culture, if adaptation is even possible, about the selection of appropriate translators, about the process which will be used by the translator, and the process which will be used to offer evidence of the quality of the translation, about which test materials will need to be adapted (e.g., test items, instructions, administration procedures, items, formats), about any supplementary research which will need to be undertaken in order to make the test usable in the new language form, such as norming, validity studies in the new context, etc. The decision to adapt rather than adopt or assemble a test should also be based on a preliminary examination of the "adaptability" of the test, i.e., the degree to which it is actually possible to adapt the test. Information about whether other similar tests have been adapted to the target culture and how well this succeeded, or about whether the focal test was adapted to other cultures and how well this succeeded, are important inputs in this decision.

ADOPTION, ADAPTATION, AND ASSEMBLY

Some authors have tried to divide the continuum of the test adaptation process into finer grains, depending on the degree of intrusion on the original components of the test (items, item formats, scales, scoring keys, etc.).

In this regard, the literature has discussed differences between adoption, adaptation, and assembly (van de Vijver & Leung, 1997; He & van de Vijver, 2012, 2015a). This classification is especially useful in cross-cultural research, and describes the degree in which a specific measure follows an "imported" or an "indigenous" logic.

Test adoption and test adaptation are concerned with importing a measure which was developed in another language and culture. There are

differences between the basic philosophies of the two: while both wish to achieve a target-language version which is similar to the source-language version, test adoption is guided by the assumption that the fewer interventions are operated on the original, the more similar the target version will be and test adaptation is guided by the assumption that similarity of the two forms is sometimes only achieved by severe transformations of some test components. As a result, test adoption will modify the components of the test as little as possible, while test adaptation will achieve equivalence by any means necessary – even thorough modifications in any component of the test, should they be needed.

Test assembly, on the other hand, is concerned with developing a measure from scratch in the new culture and language: The new measure is assembled without any intention to be equivalent with, or even mimic, another measure developed in another culture. Sometimes this new measure may target a concept that has been proven important by another established measure, or may include principles established in theory (such as a specific measurement approach or structure), but test assembly will actually always develop and not mimic.

Table 1.1 describes the three types of test adaptation. Adoption is the least intrusive procedure on the original test. Adaptation is more intrusive, but keeps original content more or less untouched. Assembly intervenes on the test in significant ways, generating new content. Adoption is equivalent

TABLE 1.1 *Levels or types of test adaptations, according to van de Vijver (2015a)*

Type	Description	Procedure
Adoption	Items are simply translated, and the original test is adopted as is in the target language	Test items are linguistically translated from source to target language, without changes in item content, other than linguistic
Adaptation	Items are modified (adapted) to suit the target cultural context	Cultural references from the source culture are modified to suit the target culture. Currency, length and weight measures, geographic landmarks, and others are changed.
Assembly	Items are replaced with completely rewritten (new) items, because not even adaptation can make them appropriate for the target cultural context	New items are developed to replace those items that are unsuited for the target culture. The new items are not slightly changed versions of the original, but are completely new

with an imposed-etic stance. Adaptation is equivalent with an etic stance with good observance of cultural aspects. Assembly is the most open to an emic stance, without actually being equivalent with what Church (2001) proposed as levels of an "Indigenization-from-within" process.

It is interesting that van de Vijver (2015a, p. 125) does not argue for the necessity to produce adaptation or assembly for every adapted test, but also suggests a utilitarian approach, by explaining that an item *may* need changes that go into less (e.g., transforming dollars to euro) or more (e.g., rewriting content entirely) subtle areas.

Test adoption. Among these three options, test adoption is most influenced by a wish to import the original test with as few changes as possible. Test adoption is also simpler, less effort intensive, and is faster to accomplish. In the case of test adoption, the original test is certainly translated, but as few changes as possible are made to the original setup. These changes rarely go as far as changing item formats and scales, and usually only touch item wording. The main objective of a test adoption is to have a good translation, ensuring linguistic equivalence between the source and the target versions of the test. Other kinds of equivalence are rarely of interest, although some may emerge without the specific intent of the researcher. For example, if two cultures are sufficiently close to one another, a simple linguistic translation may show later, based on data, that the source and target version of the test also show measurement equivalence. Therefore, test adoption does not preclude higher forms of equivalence than linguistic, but the main (and oftentimes sole) interest of the researcher in the case of test adoption is a linguistically equivalent target form of the original test.

Test adoption is simple, fast, and offers a number of other advantages. For example, it may make it easier to introduce good measures in the practice of emerging countries. In such situations, psychometric expertise is oftentimes not very developed and it is thus very difficult to generate valid indigenous measures. In such situations, the access of professionals to even an adopted measure with good documented characteristics is of high impact.

Test adoption, however, also has a number of severe disadvantages. The main disadvantage is the fact that no attention is given to the actual validity of the target-language form of the test. Validity is assumed to be a characteristic of the test, and not of the test form, i.e., the researchers assume that if the test has been proven to be a valid measure of a specific construct in its original form, it will retain that validity no matter what language it is used in. This reasoning is fallacious, and because of it test adoption ignores a number of important questions.

For example, is the target construct the same in the target language and culture as in the source language and culture? If it is not, then the test will lack construct validity: It will simply measure in an incomplete or otherwise erroneous manner the intended construct. Is the item format which was proven to function in the source language and culture also appropriate for the target language and culture? If it is not, then the item format needs to be modified. Is the structure of the items and scales the same in the target language and culture as in the source language and culture? If not, then the scoring keys may need to be rewritten for the target-language form of the test.

These questions are usually disregarded by test adoption and may at the most be considered post-hoc, by amassing evidence that the simply translated target-language form of the test is good enough to be used in the target culture. We would point out that even if the target-language form of the test would ultimately be shown to be appropriate from several of these supplementary points of view, such an approach does not focus on the *best possible* target-language form: It focuses on the best translation and a *good enough* (i.e., usable) target-language form.

Test adaptation. Test adaptation is considered in this framework to be somewhat broader in scope than test adoption. It adheres to the same philosophy, i.e., that a test created in the source culture is imported in the target culture in such a way as to influence the original form of the test as little as possible. The understanding of what is appropriate to modify is, however, different: In the case of test adaptation, of paramount importance is the fact that the two versions remain comparable ("equivalent"). Every change operated on any component of the test is acceptable, as far as it leads to a usable version in the target language and culture, which is equivalent to the source form of the test. This differs from the basic attitude that is fundamental to test adoption, i.e., that no changes should be made to the test unless absolutely necessary – and then if possible only in language.

Therefore, test adaptation does not stop at a translation of the test from source to target language, but includes at least two extra elements. These extra elements are identified by a number of authors (e.g., Hambleton, 2005) as encompassing all the psychometric activities that are undertaken when a test is developed. These psychometric activities are in part analytical and in part developmental: Test adaptation encompasses the psychometric analyses which test for the equivalence of the target-language version when compared with the source-language version, but it also encompasses systematic efforts to change any components of the tests in

such a way that equivalence is obtained. This is oftentimes an iterative process.

Test assembly. Test assembly is a procedure by which a new test is developed entirely from scratch in a specific language and culture.

Test assembly does not follow any importing logic: No test created in another language or culture is adopted or adapted to a new culture. Instead, the test is directly developed in the target culture and language. Test assembly may at the most follow the example, the success story, provided by an established measure developed in another language or culture.

For example, a new test on emotional intelligence may be developed in a new culture by following the example of another, established test of emotional intelligence; in this case the construct is adopted, and the development process may go through the same steps, but the new test is in no manner similar to the original one – except in terms of the target construct.

Or, a new test may be developed by following an established theoretical model; for example, a new test on personality is developed in a culture by following an internationally reputed model of personality structure. Of course, test assembly may choose to not even follow an internationally reputed model but develop even the underlying model of the new test in the target culture.

The only reason test assembly is actually discussed in conjunction with test adaptation is a dilemma in the domains of cross-cultural and cultural psychology in which researchers debate the utility and efficiency of measures that are imported vs. measures that are developed indigenously. Some authors argue for the need to develop indigenous measures, which are true not only to the cultural specifics of a country but in which the constructs are defined in a way that is specific to that culture. This stance is usually argued on behalf of cultural psychology, emphasizing a fundamental non-comparability of constructs and measures from one culture to another. Other authors argue for the need to ensure cross-cultural comparability, based at the least on the acknowledgment that some constructs can have universal components, if not actually be universal. This is part of a larger emic–etic debate that has been covered in another section of this book.

As part of this debate, test assembly may be the ideal activity, offering in the end measures that are not adopted, nor even adapted, but linguistically and culturally completely appropriate (Byrne, 2015). We consider that for many constructs and measures, this position oversells the benefits of developing a new test. There is extensive evidence from scientific literature

and practice that test adaptation, if followed in a professional and diligent manner, can produce target language and culture forms that are linguistically and culturally perfectly appropriate, covering the target construct in a valid and reliable way. The need to look toward test assembly as an optimal, ideal solution in *every* situation is therefore not supported by evidence.

There are several reasons why professionals may be motivated to opt for test assembly rather than test adoption or adaptation.

1. First, it should be noted that sometimes both adoption and adaptation processes, no matter how well designed and diligently conducted, fail to produce a target-language version that is appropriate from a number of points of view, such as linguistic, cultural, and psychometric (Byrne, 2015). For most tests, providing a linguistically appropriate form is not very difficult even through adoption or adaptation. However, for some tests (e.g., a language achievement test), adapting the measure is almost akin to developing it anew. And no matter what the difficulties in the area of linguistic appropriateness, cultural and psychometric appropriateness is usually much more difficult or even impossible to obtain. In such cases, test assembly is the only choice to provide the target language and culture with a workable, useful, valid measure of the target construct (He & van de Vijver, 2012). In this case, test assembly is driven by a failure to adopt or adapt a specific test.
2. Second, in some other cases, a culture has amassed consistent evidence that in a certain domain (e.g., personality or quality of life), etic (universal) approaches do not work. This may be based on previous trials to adapt measures, by research into the cultural specifics of a construct, by qualitative studies or other sources of evidence. In such cases, test assembly is warranted by a belief that the target culture is unique in some way and test adaptation should not even be attempted, as it is doomed to fail, and that only developing a test from scratch will capture the specifics of this culture.
3. Third, in yet other cases, the test that is the potential target of a test adaptation, may have proven difficult to adapt in other such attempts, in other cultures. This may suggest that the test has strong cultural ties to its source culture, and hints at similar difficulties in this target culture as in other attempts at adaptation. Previous failures to adapt a test in other cultures, and an impossibility to replace that test with another measure, may thus motivate the effort of test assembly.

4. Fourth, of course, sometimes, due to rather ideological reasons, or a strong belief on the part of the researcher, such as a strong cultural stance, adoption or adaptation of a measure may not even be considered: existing measures or models may be considered inappropriate in principle, and test authors may proceed directly to test assembly without bothering to even look at the possibility of adapting another measure.

As seen, the discussion on whether to adapt a measure or develop one in the target culture is complicated, and we will continue to approach it in the following section from two different points of view. The first point relates to a philosophical stance on the usability *in principle* of imported (adapted) vs. culturally developed measures. The first point relates to economic reasoning, i.e., the costs and benefits of imported (adapted) vs. culturally developed tests.

Why Assembly (Local Development) Is Oftentimes Not a Realistic Option

Developing a good test is a difficult feat. While the test is credited to an author or group of authors – as it should be, as it reflects creative scientific work – we advance the suggestion that test production is not only the effect of an author's proficiency and determination but also of an ecosystem comprising researchers, test publishers, test users, and other stakeholders. In effect, this all means that tests tend to be developed in countries where psychology is well developed.

A test is the result of a cluster of competences demonstrated by its authors. In order to produce good tests, an author or team of authors needs at least two critical competences: substantive knowledge and psychometric expertise.

First, authors need substantive knowledge: They need to be experts in the substantive topic addressed by the test. A good test, which will be accepted by the scientific community, will be used in independent research (an important point if evidence of validity is to be generated), and will later be absorbed in practice by specialists cannot be developed by just anybody.

It is more likely that a good test is developed by an established scientist than by an early career researcher or even a student. In order to be able to propose a new measure for a construct, authors need a good knowledge of the domain. Good tests are not developed very early in a scientific career,

as they require more than just psychometrics. This is especially valid for areas where established tests are already flourishing: In order to propose a new test that is better than a number of others that already exist, or is able to fill an existing gap, authors need a good understanding of the domain targeted by the test. When authors reach this level of competence, they may already be themselves established as scientists.

It is also more likely that a test, once developed, will also be accepted by the community if developed by an established scientist. We are aware that the authority argument is a sophism, but both the scientific and professional community are sensitive to this argument. Both buy trust as much as scientific fundamentals in a test. For example, questions that may arise when a new test of depression is published are questions such as "Who are the authors?," "Did they publish before in the area of depression?," "Are they reputed in this domain?" The overarching question here is: "Do they have enough experience or reputation to propose a measure?"

Second, authors need psychometric knowledge: They need to understand modern theory and practice in the domain of measurement. Psychological and educational measurements have specialized heavily during recent decades. Some of the procedures and approaches taken by early test writers may seem simple today, and may actually be considered unacceptable. For example, Classical Test Theory in general has faced a sharp decline, and Structural Equations Modeling and Item Response Theory are omnipresent in modern test development. Even aside from the more technical statistical approaches, psychometrics has become a science in itself, with new research being published almost daily on scaling, dimensionality, constructs, faking, and many more. These procedures are reasonably difficult and competence is required of test authors not only in the substantive domain but also in the psychometric area.

The ecosystem of testing. However, individual competence is not enough in order to generate a good indigenous production of tests. Psychology needs to be sufficiently developed in a country in order to have given birth to the entire ecosystem of testing. Test authors are only one piece of this ecosystem: Test takers, test users, test publishers, and other stakeholders are equally important.

Good tests are developed with a great investment of scientific and financial resources. Test authors may be ready to invest their time and scientific competence, but they usually do not have the financial resources needed for test development. This investment may be supported by an interested party, such as a public institution or a policy maker. But public institutions cannot support the entire test production of a country – they