The Cultural Element in the Adaptation of a Test: Proposals and Reflections on Internal and External Influences

Ana Isabel Ponce Gea and Francisca José Serrano Pastor

1 Department of General and Specific Didactics, University of Alicante, 03690 Alicante, Spain
2 Department of Research Methods and Diagnostics in Education, University of Murcia, 30100 Murcia, Spain; fjserran@um.es
* Correspondence: anaisabel.ponce@um.es

Abstract: In the framework of instrumental research, the adaptation of tests has been increasing. The aim of this paper is to explain the procedures followed for the adaptation of a test with a construct of a cultural nature to another context, reflecting on the difficulties and limitations related to its equivalence and validity. For this purpose, we start from the Test on the Construction of Historical Knowledge, originally designed for a Spanish context and targeted at elementary school students, in order to design and validate its Portuguese version. The process of “emic-etic” adaptation, in which the technical, semantic, conceptual, and metric equivalence of a test is sought, was carried out through group translation with post-translation empirical processes (expert judgement, a pilot study, and external criteria). It highlighted some issues closely linked to internal and external cultural factors. On one hand was the predominance of analytical-rational processes in the adaptation process, with an essential deep reflection on the construct and the meanings intimately linked to the context. On the other hand, were the difficulties of empirical procedures due to the qualitative nature of the construct. This led us to reflect on the possibility of minimizing the "emic" aspects in eminently cultural constructs without renouncing reliable and valid results with respect to the construct.

Keywords: adaptation of a test; cultural relevance; evidence-based practices; design of a test

1. Introduction

In the face of empirical research, instrumental research focuses on the design of instruments and the analysis of the psychometric features of tests [1] in order to obtain the best measurements for each domain adjusted to each receiver.

In the framework of this type of research, the adaptation of tests to other linguistic and cultural contexts has a long history that can be traced back to the work of Binet and Simon [2]. However, in recent decades there has been a notable increase in the adaptation of tests [3], largely because of the need arising from the continuous interchange of people and ideas across the planet, the utility of applying a result to different contexts, and the proliferation of comparative studies. Thus, test adaptation as a cross-cultural product [4] is becoming increasingly popular and useful.

A key aspect of these decades of development is the progressive distinction between the translation and adaptation of a test [5,6]. Adapting a test is much more than a linguistic issue; it involves cultural, conceptual, and metric elements [3] that have an impact on the equivalence between the original and the resulting tests. The adaptation process involves attending to the internal and external validity of the instrument [1], perceiving the process as flexible, and transcending literal translation.

The International Test Commission, in response to the proliferation of test adaptations, started a project in 1994 [7,8] to develop guidelines for test adaptation. The aim of these guidelines is that “the final product of the adaptation process should achieve the highest possible level of linguistic, cultural, conceptual and metric equivalence to the original test” [3], p. 152. These guidelines are divided into different categories related to the process.
of test design and application: before development (analysis of the legal framework, design, and construct evaluation), during development (linguistic and cultural adaptation and pilot studies), confirmation (data collection, equivalence, reliability, and validation), application (administration), scoring and interpretation (interpretation of the results and comparability), and documentation (changes between versions and appropriate use). The adaptation process is, therefore, presented as a complex process, distant from the initial idea of translation.

In the context of increasing recognition of cultural influences when adapting a test [6,9–12], advances have been achieved in methodological and technical developments for cross-cultural equivalence [13,14]. Several authors and institutions have recently offered guidelines for the translation and adaptation of tests [15–19]. Each approach assumes that technical, semantic, and cultural equivalence must be sought in the construction of a scale because the meaning of terms can differ according to context. The search for conceptual equivalence [20] implies the recognition that a concept may differ according to cultural context and that equivalence in the meaning of an item is not enough to ensure equivalence in its conceptualization and, consequently, the equivalence of the scale. In the same vein, it implies accepting the existence of some general elements in tests that influence the technical equivalence itself and that are culturally dependent. For example, this is the case for condescension, which is directly related to the response to items [21,22].

The results of studies focusing on test adaptation—[10,23–27] to mention a few examples—support the same idea, recognizing the differences between cultures, even to the extreme of invalidity of the scales [23,26].

In education, the rise of cultural adaptations finds paradigmatic examples in measurements through the Programme for International Student Assessment (PISA) or the Trends in International Mathematics and Science Study (TIMSS), among the most relevant. In addition to the influence that cultural factors exert on all tests and the intrinsic difficulties of measurement in education (the singularity of each school environment, the prevalence of a qualitative paradigm for certain constructs, etc.), there are also close relations with the cultural sphere established by certain constructs. In this sense, epistemological theories on the construction of historical knowledge [28–30] necessarily imply the consideration of first- and second-order contents [31]. Therefore, if we understand epistemological theories as an organization of beliefs that assume a specific way of understanding knowledge, reflecting both its nature and its construction process [32,33], the notions of how to construct in history—in other words, the second-order content (causality, elaboration of evidence, historical perspective, etc.)—cannot be separated from the historical processes on which the construction is planned (first-order or substantive content). Thus, factual events directly related to the historical culture in which they are embedded [34,35] have an impact on the way of understanding the construction of historical knowledge within the same cultural context, especially when this context is modified.

Given a construct wherein most research is based on a qualitative approach [36] and very few instruments with their own name are available for the generalization of results [37–39], we adapted the Test on the Construction of Historical Knowledge (CONCONHIS) [40], originally designed for a Spanish context, for a Portuguese context.

The purpose of this paper is to explain the procedures followed for the adaptation of a test that has a construct of a cultural nature to another context, reflecting on the difficulties and limitations regarding its equivalence and validity. To this end and in accordance with the singularity of the construct, we pay as much attention to the analytical-rational processes as to the empirical.

2. Original Test Design: Description of the Construct and of the Instrument

The CONCONHIS was designed to assess the construct of epistemological theories on the construction of historical knowledge, understood as those beliefs related to the nature and process of the construction of historical knowledge. Its aim is to evaluate subjects’ conceptions of each of the operations involved in the process of constructing historical
knowledge. To this end, its design is based on the definition of a construct composed of three dimensions in which ten variables and twenty-one indicators are distributed.

The first dimension, inference making, includes epistemological theories relating to the choice of what is relevant and the formation and evaluation of inferences. The second dimension, structuring operations, includes the formulation and evaluation of problems; the search for, suitability, and reading of sources; the adoption of perspectives; the obtaining of evidence; the resolution and justification of questions; and the elaboration of discourses. A third dimension, alternative operations, involves variables related to theories around the observation of reality and its questioning; the search for alternatives; the retention and identification of perspectives; and the formulation of divergent questions about sources.

Based on these conceptual foundations, the CONCONHIS was designed as a standardized, objective, nonverbal test [41,42] consisting of 42 items (two for each indicator) distributed in Part I and Part II of the test and validated for Spanish students aged between twelve and fifteen years. It is a test for individual or collective application, requiring about 50 min to complete each part. Its items are multiple-choice questions, with three answer choices corresponding to three levels of development for the indicator measured. The wording of an item consists of a situational question in which the participant is asked about his or her perception of the way in which a particular historical procedure is conducted associated with a specific historical event; it includes an image in the style of a vignette, which may or may not be essential in order to answer the item.

The CONCONHIS has undergone various validity and reliability improvement procedures ensuring its content validity (with the participation of 17 experts), construct validity (confirmatory factor analysis based on the responses of 222 subjects; CFI = 0.99; TLI = 0.99; RMSEA = 0.01; SRMR = 0.04), and internal consistency (alpha = 0.63), which is understood as moderate in accordance with the nature of the construct [43].


The adaptation process conducted for the CONCONHIS can be defined as an emic-etic adaptation [20]. Accordingly, attention was paid both to the aspects that were common to the cultures involved (etic) (variables, general structure of the scale, items, etc.) and to the specific elements of the culture to which the test was to be adapted (emic)—in essence, to its social-historical context.

Figure 1 includes the analytical-rational processes for test adaptation in the search for technical, semantic, and conceptual equivalence [10,44] based, among others, on Gana et al. [18].

First, following the guidelines established prior to the development of the adaptation of the International Test Commission [3,45], we analyzed the issues related to the legal framework and the relevance of construct assessment in the new culture. We concluded that curricular reforms have been carried out in parallel in Spain and Portugal, following the same direction [46,47]; that, in both contexts, textbooks continue to be the predominant didactic resource, which implies a gap between the approach to historical thinking collected by the instrument and the school reality [48]; and that multiple research projects have been carried out in the Portuguese context with similar results to those in the Spanish context [49–52]. In addition, the original design was analyzed by identifying linguistic and cultural features relevant or irrelevant to the adaptation.

This allowed an adaptation to be developed where the translation process was presented as essential. A back-translation process was initially considered [53], and we finally opted for a group translation. The group consisted of two researchers who were knowledgeable about the object of the research and familiar with the scientific approach [54], as well as with the two languages and cultures [11,17]. Specifically, each of them is a native speaker of one of the languages (Spanish or Portuguese) and has completed long-term research stays in the other country.
In the group translation process, three differentiated phases were distinguished. In the first, a group phase, the variables, indicators, and items of the test were reviewed, considering that they should all be maintained in the new version in order to achieve technical equivalence. It was also agreed that, while maintaining the initial structure as far as possible, some changes in terms or examples were acceptable in order to ensure semantic and conceptual equivalence. In a second individual phase, the experts developed independent adaptations of the test, taking into account the elements identified by Hambleton and Zenisky [55] for translation quality control. Finally, in a third phase, a consensus version was obtained, reflecting on the differences between the independent versions for each item.

Table 1 shows the main changes made to the items.

As can be seen, the modifications were mainly due to historical context, specifically to historical events of national influence whose meanings differed for Portuguese participants. Therefore, in the process of adaptation, priority was given to the search for a comparable historical event, both in terms of its meaning for the subject and in terms of its usefulness for measurement of the operation to which the item referred. This was the case for items 10, 14, 23, 24, 34, and 42. The change made to item 12 can also be explained by the context, although in this case it was related to the landscape or even the weather. Another type of modification related to the linguistic level (items 22 and 38), moving away from literal translation to consider connotative or historical aspects of the language. Moreover, it is
interesting to highlight that modifications were not only limited to the verbal statements, but alterations were also necessary in the illustrations. As can be seen in Figure 2, for those items in which the illustrations were not essential for the response to the item, neutral illustrations were chosen to allow future adaptations.

Table 1. Comparison between the items in the Spanish and Portuguese versions.

<table>
<thead>
<tr>
<th>Item</th>
<th>Spanish Version</th>
<th>Portuguese Version</th>
</tr>
</thead>
<tbody>
<tr>
<td>10</td>
<td>El Quijote and Amadís de Gaula are used as relevant pieces of literature to assess the level of criticism of reality; the illustration refers to El Quijote</td>
<td>The works are changed to Os Lusiadas and A Mensagem, looking for the same meaning for the pupils; explicit references to the works in the illustration are eliminated</td>
</tr>
<tr>
<td>12</td>
<td>Illustration of blanket irrigation</td>
<td>Illustration with the characters talking, avoiding the Spanish version because of its resemblance to a river</td>
</tr>
<tr>
<td>14</td>
<td>Based on a research study on Carlos V</td>
<td>Based on a research study on Carlos V</td>
</tr>
<tr>
<td>22</td>
<td>In one of the response options to the item, tomato indigestion is included</td>
<td>Tomato indigestion is changed to potato indigestion because of the connotative meaning that the former may have in a Portuguese context</td>
</tr>
<tr>
<td>23</td>
<td>Contains a censorship process in the post-war context dated 1942</td>
<td>The problem is transferred to the Portuguese context to the period of the Dictatorship (1941)</td>
</tr>
<tr>
<td>24</td>
<td>Based on the loss of Granada by King Boabdil</td>
<td>The historical event is changed by the arrival of D. Alfonso III in the Algarve</td>
</tr>
<tr>
<td>34</td>
<td>The text of the illustration includes “rey y reina de Castilla [King and Queen of Castile]”</td>
<td>This is amended to “reis de Portugal [Kings of Portugal]”</td>
</tr>
<tr>
<td>38</td>
<td>The text of the illustration is presented in old Spanish</td>
<td>The reference to Galician-Portuguese is included in the text of the vignette</td>
</tr>
<tr>
<td>42</td>
<td>The research focuses on the beginning of the reign of the Bourbons with Felipe V</td>
<td>The research focuses on the dynasty of Bragança with D. João IV</td>
</tr>
</tbody>
</table>

Figure 2. Example of the adaptation of an item (item 10).
As a result of all the above, we obtained a Portuguese version of the test of 42 items with linguistic and contextual adaptations, seeking in the translation process technical, semantic, and conceptual equivalence of the different versions.

4. First Steps in Test Validity and the Search for Metric Equivalence

The challenge of adapting a test to another context is not limited to the moment at which the design is adapted and where cultural issues are explicitly raised but includes a further validation process, which is also not free of cultural influences. In fact, the International Test Commission includes pilot studies as part of its development, discussing data collection, equivalence, reliability, and validation as components of test confirmation. Along these lines, the adapted version of the test underwent three processes with the aim of seeking its validity and equivalence.

First, the CONCONHIS was examined by expert judges (n = 10) with a range of profiles (history, didactics, and methodology) and extensive teaching experience (M = 16.70, SD = 12.68 at university (70%) and non-university levels (30%). The evaluation sheet included 29 items divided into three blocks: assessment of the dimensions; assessment of the items; and global assessment of the test. The judges were asked to place themselves on a scale of 1 to 4, from not at all to totally adequate, and answer either relevant or clear for the items and coherent, appropriate, or sufficient for the dimensions. In addition, they were asked to answer yes or no to the following statement: “I consider item X of the Spanish and Portuguese version to be technically, conceptually, and semantically equivalent.” A statistical package, IBM SPSS Statistics (version 24.0.0), was used to calculate the mean, standard deviation, and frequency of the responses, and Kendall’s W test was applied to analyze the concordance between the average traits assigned by the judges.

In line with the objective of the present study, some of the results are highlighted herein. First, good results were obtained in the global evaluation of the test relative to the research objectives (M = 3.20, SD = 0.27), the target audience (M = 3.60, SD = 0.32), and the applicability (M = 3.50, SD = 0.43), with the overall evaluation of the test yielding M = 3.63 and SD = 0.36, which is almost entirely adequate. Second, the mean scores for the dimensions were close to the maximum value for appropriateness (M = 3.97, SD = 0.07), coherence (M = 3.87, SD = 0.31), and sufficiency (M = 3.87, SD = 0.27). Moreover, they were above the value of 3.50 for all the dimensions of the test (Table 2).

Table 2. Scores given by the expert judges for the dimensions.

<table>
<thead>
<tr>
<th></th>
<th>Appropriateness</th>
<th>Coherence</th>
<th>Sufficiency</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>M</td>
<td>SD</td>
<td>M</td>
</tr>
<tr>
<td>Explanation project</td>
<td>4.00</td>
<td>0.00</td>
<td>3.90</td>
</tr>
<tr>
<td>Documentary phase</td>
<td>3.90</td>
<td>0.31</td>
<td>3.80</td>
</tr>
<tr>
<td>Explanatory phase</td>
<td>4.00</td>
<td>0.00</td>
<td>3.90</td>
</tr>
<tr>
<td>Representative phase</td>
<td>4.00</td>
<td>0.00</td>
<td>3.90</td>
</tr>
</tbody>
</table>

Third, in the assessment of the items’ adequacy (“I consider item X adequate”), relevance (“I consider item X relevant”), and clarity (“I consider item X clear”), all three exceeded a mean of 3.50 out of 4, with a standard deviation of less than 0.40 (Table 3).

Table 3. Scores given by the expert judges for the items.

<table>
<thead>
<tr>
<th></th>
<th>Minimum</th>
<th>Maximum</th>
<th>M</th>
<th>SD</th>
</tr>
</thead>
<tbody>
<tr>
<td>Adequacy</td>
<td>2.98</td>
<td>4.00</td>
<td>3.67</td>
<td>0.30</td>
</tr>
<tr>
<td>Relevance</td>
<td>3.30</td>
<td>4.00</td>
<td>3.83</td>
<td>0.22</td>
</tr>
<tr>
<td>Clarity</td>
<td>3.16</td>
<td>4.00</td>
<td>3.63</td>
<td>0.31</td>
</tr>
</tbody>
</table>
For adequacy, 77.27% of the items had a mean higher than 3.50, with only one item (item 1) scoring less than 3; for relevance, 88.63% of the items exceeded 3.50; and for clarity, 61.36% of the items reached 3.50.

Among the participating judges, six were recognized as having sufficient linguistic and contextual proficiency to assess equivalence; all items received 100 percent positive responses for their technical, semantic, and conceptual equivalence.

According to Kendall’s W test ($W = 0.202$, $J_2 = 239.648$, $gl = 9$), there was a statistically significant agreement ($p < 0.01$) between the average traits assigned by the judges ($J_1 = 5.67$, $J_2 = 3.85$, $J_3 = 5.72$, $J_4 = 6.27$, $J_5 = 5.71$, $J_6 = 5.30$, $J_7 = 4.14$, $J_8 = 5.63$, $J_9 = 6.29$, and $J_{10} = 6.43$), although the intensity of the agreement was insignificant.

Second, a pilot study was carried out with 64 eighth grade (i.e., 13- to 14-year-old) Portuguese students (26 boys, 38 girls). Using the IBM SPSS Statistics (version 24.0.0) and MPlus statistical packages, descriptive statistics, and Cronbach’s $\alpha$, the ease and discrimination indices of the items were calculated, and a confirmatory factor analysis was conducted. Although the small sample size did not allow for an in-depth analysis of the psychometric characteristics of the test and item behavior, neither of the descriptions of the procedures mentioned above were relevant in the framework of this paper. The factor structure of the Spanish version (CFI = 0.99; TLI = 0.99; RMSEA = 0.01; and SRMR = 0.04) was conserved, and the model fit indices were very similar (CFI = 0.99; TLI = 0.98; RMSEA = 0.02; and SRMR = 0.04). Although a separate confirmatory factor analysis was necessary, we understood the results as suggesting metric equivalence based on the recognition of the factor structure as an element supporting this equivalence [56].

Third, we searched for the criterion validity of the test, understood as the relationship between the scores obtained by this test and an external criterion of measurement of the same construct it deals with [41]. Specifically, we focused on the concurrent validity [57,58] applicable when the measures of the criterion and the test are obtained, approximately, at the same time. A relevant, valid, and uncontaminated criterion [58] was considered to be the assessment for each of the students by a teacher with in-depth knowledge of the group and the object of the research. Thus, the teacher scored the historical operations relating to each variable using the same scales on which the test was designed according to the perceived development for each of those assessed ($n = 19$). The validity coefficient (Pearson’s correlation coefficient) between the test variables and the external criterion scores was calculated using IBM SPSS Statistics (version 24.0.0). The analysis of the validity coefficient revealed a high, positive, and statistically significant correlation between the total score of the test and the total score of the external criterion and a higher correlation between the variables of a methodical nature than between the more creative variables. Nevertheless, bearing in mind that coefficients of 0.20 or greater were considered useful for predicting criterion and that values greater than 0.60 were considered unusual for criterion validity [41,57], we understood the results as evidence of this type of validity.

Therefore, the analytical-rational and empirical processes involving the Portuguese adaptation of the test allowed us to speak of a technical, semantic, and conceptual equivalence between the versions and suggestions of metric equivalence.

5. Conclusions, Limitations and Future Perspectives

In this paper, we set out to explain the procedures followed for the adaptation of a test with a construct of a cultural nature to another context. The adaptation process developed, based on the guidelines of the International Test Commission [3,4] and in line with other adaptation studies (e.g., [10]), showed the complexity of the object of the study and the process of adaptation to another culture. In this regard, some reflections can be undertaken, by way of conclusions, that are of interest for the adaptation processes of this type of test.

Firstly, and always with the aim of finding the best measurements for each construct [1], the analytical-rational processes were placed in the foreground. The adaptation process was, thus, understood as a complex, reasoned, and analytical procedure in which a deep knowledge of the object of study was essential. In this sense, the analysis of the construct,
its operationalization in items, and reflections on the influence of cultural changes on the conceptualization of the process of constructing historical knowledge were presented as decisive for the equivalence of the versions. Thus, in a context of test design where the empirical prevails over the analytical, where constructs are difficult to define through objective variables and items, and where the contextual and cultural influence is inseparable from the construct itself, the process of adaptation and the search for equivalence between versions must be understood as a continuous, even cyclical, interaction between rational and empirical processes.

Secondly, and following the same line, the difficulties of empirical procedures were explained by the qualitative nature of the construct. We started from a test whose original version had moderate internal consistency. This reality, a priori, explains why any variation in the test implied a difficulty in maintaining the indices relating to the validity and reliability of the instrument. This led us, once again, to rethink the construct, which may have led to oversimplification or excessive opportunity cost [59], which ultimately distorted what we really wanted to measure. However, the necessary commitment to the development of tests that allow generalization of the results of complex constructs inevitably implies assuming these problems and understanding the challenge in the medium or long term. It also implies adopting a flexible judgement from the point of view of metric equivalence.

Thirdly, the effect that culture had on the test, both externally and internally, provoked a reflection that we consider to be of interest for the development of this type of test. On the one hand, the cultural element—more specifically in this case, the historical events—may influence the level of development of the epistemological theories related to each historical operation; on the other hand, including in the items events of special significance for the students implies a particular adaptation for each context. This led us to ask ourselves to what extent was it possible to construct neutral tests, removed from national history, for the measurement of epistemological theories related to the process and not to the fact? Was it possible to minimize the emic aspects in eminently cultural constructs without renouncing reliable and valid results with respect to the construct? In this sense, the adaptation carried out facilitated the elaboration of later versions—especially regarding the illustrations—assuming, however, episodes of Portuguese national history in the search for the greatest equivalence. Apart from the above, we believe that this is a point for reflection, debate, and research.

Finally, to sum up, we can say that the scale was valid for other cultures in contrast to the results of other studies [23]. This can be affirmed by taking into account some limitations. Firstly, the affirmation of its validity was made for closed cultures. The consideration of other cultural contexts, within or outside of a Western framework, would entail added difficulties and the need for further studies. Secondly, the process of adapting the test was still partial. In this respect, while it was possible to speak of technical, semantic, and conceptual equivalence, the metrical equivalence can be understood as an indicium. Thus, larger studies are needed with a greater number of participants to allow an exhaustive analysis of the equivalence of the model and its factorial structure, as well as the psychometric characteristics of the test and the items (e.g., discrimination and difficulty indices, among others).

Despite the above, we consider this version to be valid and the process carried out to be useful for reflection and decisions regarding the adaptation of tests to other cultures. In this sense, the greater the complexity and the more qualitative and cultural the nature of the construct, the more necessary it is to approach the adaptation process through reflection and analysis, rather than as a mechanical procedure.

**Author Contributions:** Conceptualization, A.I.P.G.; methodology, A.I.P.G. and F.J.S.P.; software, A.I.P.G.; formal analysis, A.I.P.G. and F.J.S.P.; investigation, A.I.P.G.; original draft preparation, A.I.P.G.; writing—review and editing, A.I.P.G. and F.J.S.P. All authors have read and agreed to the published version of the manuscript.
References
2. Binet, A.; Simon, T. Méthodes nouvelles pour le diagnostic du niveau intellectuel des anormaux. L’Année Psychol. 1904, 11, 191–244. [CrossRef]
19. Khan, G.; Mirza, N.; Waheed, W. Developing guidelines for the translation and cultural adaptation of the Montreal Cognitive Assessment: Scoping review and qualitative synthesis. BJPsych Open 2022, 8, 1–17. [CrossRef]


