Translation of Instruments for Cross-Cultural Research

HUI-JU C. LIU

Department of English Language, Da-Yeh University
112 Shan-Jiau Rd., Da-Tsuen, Changhua, Taiwan

ABSTRACT

Researchers tend to disregard problems related to the translation of measurement instruments as compared to other conceptual problems in cross-cultural studies. Often little or no information concerning the translation procedures is contained in studies in which translation is involved. A lack of correctness in a test translation can result in invalid conclusions regarding cross-cultural research.

In this study three aspects of translation were investigated: (1) factors that influence the validity of a translated instrument, (2) basic translation techniques, and (3) procedures for establishing equivalence between source and target versions. A study of translation equivalence on the equating of English and Chinese versions of the Statistical Reasoning Assessment (SRA) instrument was also conducted. The SRA was developed to measure statistical reasoning concerning probability and statistical concepts. It was first translated into Chinese by using a combination of methods to ensure the quality of translation.

Key Words: translation of measurement instruments, translation equivalence, back translation

跨文化研究測驗量表的翻譯

劉慧如

大葉大學英美語文學系
彰化縣大村鄉山腳路112號

摘 要

研究學者在作跨國研究時常常較重視觀念方面的問題，而忽略翻譯測驗量表時所產生的問題。在跨國研究中，當研究者必須使用並翻譯量表時，經常很少或是沒有提及任何關於翻譯程序的訊息，量表的翻譯若缺乏正確性可能會導致此類研究作出缺乏效度的不當結論。

本研究從三個方面來探討有關翻譯的問題：(1) 影響翻譯量表效度的因素，(2) 基本的翻譯方法，以及(3) 如何確定量表的原文版與翻譯版本兩者之間的對等性。統計推理測量（the statistical reasoning assessment, SRA），一份用以測量學生機率與統計推理能力的問卷，在本研究中首度被翻譯成中文，並在翻譯過程中以各種方法來達到被大多數研究學者所忽略的翻譯對等性。

關鍵詞：測驗量表的翻譯，翻譯的對等性，反向翻譯法
I. INTRODUCTION

Review of past studies showed that researchers have been interested in the cross-cultural use of measurement instruments for as long as psychological tests and attitude scales have existed. One of the earliest examples was the translation of the Binet-Simon Intelligence test from French into English by Henry Goddard in 1911. By 1916 this test had been translated into seven languages [19]. Today there is still a widespread practice of translating tests, scales, and questionnaires for use in other languages and cultures. For example, the most widely adapted personality inventory is the Minnesota Multiphasic Personality Inventory (MMPI). According to Butcher and Spielberger [5], the MMPI had been used in over 50 countries and there were over 100 translations in existence.

Researchers such as Chapman and Carter [7], Hambleton [11], and Hambleton and Kanjee [13] suggested that translating an instrument for use in a different culture is far less expensive and considerably faster than to construct a new one. Many developing countries have only a short history of empirical research and do not have the technical expertise or qualified personnel to actually develop the instruments they need. Hambleton and Kanjee [13] also suggested that translation problems that remain are considered less serious from a validity perspective than the problems of bias that occur when examinees are being tested in their second language. Hansen [14] found that the development of interest inventories in a second culture is in most cases too expensive, too impractical, and sometimes even impossible due to unavailability of adequate sample sizes for test construction and validation.

Furthermore, translating instruments from one language to another provides a basis for cross-national comparisons. Such studies are important because an increasing number of researchers are interested in understanding the similarities and differences that exist between members of different countries. An example is the Third International Mathematics and Science Study (TIMSS) conducted by the National Center for Education Statistics (NCES) and the National Science Foundation (NSF). This international study conducted an assessment of mathematics and science achievement of half a million students from 42 countries at three different grade levels (fourth, eighth, and the end of secondary school) during the 1995 school year. Items on the TIMSS achievement test were translated from English to different languages. The TIMSS results are used to examine U.S. students’ performance in an international context, and to improve U.S. mathematics and science education to a more internationally competitive level. In 1999, a follow-up study, TIMSS-R, was conducted to assess eight-grade students in 38 countries.

Although there is widespread use of test translation in the second culture, Chapman and Carter [7] suggested that many researchers tend to disregard the problems of translation as compared to the other conceptual or methodological problems involved in cross-cultural studies. Brislin [1] reviewed 80 articles in two publications for cross-cultural research: *Journal of Social Psychology* and *Public Opinion Quarterly*, in which he found that 61 manuscripts contained little or no information about the translation procedures used in the studies. Researchers including Triandis [20] and Butcher and Garcia [4] noted that there are problems due to translation of instruments in cross-cultural research. It is important for cross-cultural researchers to determine whether the translation of the test is valid [9].

II. FACTORS THAT INFLUENCE THE VALIDITY OF TRANSLATED INSTRUMENTS

The lack of correctness of test translations can result in invalid research. According to Hambleton and Kanjee [13], three technical factors tend to influence the validity of translated instruments. They are: selection and training of translators, the instrument itself, and the process of translation.

1. Selection of Translators

Hambleton and Kanjee [13] suggested three criteria for selecting qualified translators. First, qualified translators should be those who are familiar and competent with both the target and the source languages. Second, they should have knowledge about the subject matter. Finally, they should have received some training in instrument development and know the principles of writing good test items. Butcher and Garcia [4] insisted that translators should not only be intimately familiar with both languages, but also should have a high degree of familiarity with both the target and the original cultures. It is best if translators are not only native speakers of the target language but also a native of the country in which the translated instrument is going to be administered.

2. The Instrument Itself

Hambleton and Kanjee [13] suggested that if a researcher knows that the test or scale will be translated into another language, it is advantageous to take this into account at the instrument development stage. Campbell [6] suggested that every concept under investigation should be represented by at least two questions. And they should be worded differently if possible.

Very often these tests or scales contain everyday expressions which enhance their meaningfulness in one...
language but introduce problems in the translation process into another language. This can be a threat to the validity of the translated instrument. There are several guidelines for writing English that is easier to translate. These include: using short and simple sentences; avoiding passive voice in a sentence; repeating nouns rather than using pronouns; avoiding metaphors and colloquialisms since they are less likely to have equivalent meaning in the target language; avoiding words that are vague regarding events or things; employing specific rather than general terms; and employing words or concepts that are familiar to the translators where possible [1-4, 22].

3. The Translation Process
Every effort must be made to preserve the original meaning of the test items. It is also desirable to convey accurately the meaning in a way that will be understood by the prospective test users. Translators frequently encounter linguistic problems in the translation process. Words and phrases that are difficult to translate should be avoided at the outset of the instrument development process. Sometimes, if an item is found to contain some words, phrases, and expressions that do not have clear equivalents in the other language, it may be eliminated or subject to “decentering”.

Decentering is a translation concept first developed by Werner and Campbell [22]. It refers to a process in which both the source and the target language versions are equally important and open to revision. The basis of decentering is back translation. A test is translated to and back from another language; after the two source language versions are compared, both the original and the target versions are revised if necessary. In decentering, the research project does not center on one of the cultures or languages. It is aimed at both loyalty of meaning and equal familiarity and colloquialness in both languages. It is important that materials in the source language are translated so that they are natural-sounding and meaningful to native speakers of the target language. Brislin [1] maintained that the best translations result when decentering is used. For the success of the decentering procedure, it is important that both the original and the target versions are open to revision.

III. TRANSLATION METHODS
When adapting an instrument for use in another culture, achieving translation equivalence is a crucial point in the process. Items translated should not only have linguistic equivalence in the target language, but also should have the same conceptual and psychological meaning of the original items. Several basic translation methods were suggested by Werner and Campbell [22], and Brislin [3].

1. Back Translation
This is a very useful technique for revisions of the original as well as the first translation effort. It refers to a process in which the test instrument is first translated into the target language and then back translated into the original language. It is one of the most commonly used procedures for verifying the translation of an instrument, and it is highly recommended by many researchers [1, 7, 22]. Brislin, Lonner, and Thorndike [2] recommended the use of back translation at least three times by multiple independent translators. After comparing the original with its back-translated version, if the two source language forms are not identical, subsequent modifications are made.

However, it is important to note that, even back translation is a very useful technique to discover translation errors, this technique can only be used as an initial check of translation equivalence. When two language versions seem identical, several factors may suggest the equivalence between the two versions [1, 2]. First, there are some shared rules for translating certain non-equivalent words and phrases. Second, the two languages under study are of similar grammatical structures (i.e., English and Spanish). Either forward translation or back translation is more difficult for languages of different structures (i.e., English and Chinese). Thirdly, the grammatical form is that of the source language, not the target language. The translated version of this kind is easy to back translate but may not be meaningful for target-language monolinguals. Lastly, some back-translators may be able to make sense out of a poorly translated version. Brislin [1] also considered other factors affecting translation quality: content area and difficulty of materials.

Researchers [13, 15] mentioned several criticisms of back translation. The first criticism is that the comparison of the instrument in two versions is carried out in the original language. It may be that the findings in the original language do not generalize to the target language version of the instrument. The second criticism is that no examinees of the intended population ever take two versions of the test under true testing conditions and therefore they are often operating at different cognitive levels than the translators. Third, back translators with limited second language capability can produce translations that sound awkward and unnatural. The result tends to limit the usefulness of the back translation procedure because one cannot tell whether the discrepancy between the original and back translation is due to a problem associated with translation or a lack of language fluency of back translators. Finally, the quality of translation could be poor while the evidence generated by back translation suggests that
the two versions are equivalent. Hulin [15] suggested that even high quality of back translations does not ensure psychometric equivalence of two language versions. Hulin and Mayer [16] said that back translators may generate excellent quality back translations from poorly translated target language versions by insightful guesses, assumptions and inferences by the translators about what the term must have meant in the original language. Sometimes translators may even retain the same grammatical structure of the original language because it facilitates back translation, but this may mask serious problems in the target version of the instrument and may not be meaningful to the examinees in the target population.

2. The Bilingual Technique

Bilinguals take the same test in both the source and the target language versions. Items which yield discrepant responses or different frequency of responses can be easily identified. The advantage of this technique is its potential for using sophisticated statistical methods and concepts. The disadvantage is that the research instrument developed on the basis of bilinguals’ responses may not be generalizable to the population of interest.

Tyson, Doctor and Mentis [21] maintained that different responses of bilinguals to the same items in different languages may be due to reasons other than inaccurate translation. They suggested some social psychological explanations, including the theory of accommodation, ethnic affirmation, and social desirability. According to the theory of accommodation, subjects tend to give a response which is appropriate in the culture of that language. It is because the learning of language also involves the acquisition of the values and attitudes of that particular culture. The other explanation, ethnic affirmation, is proposed by Yang and Bond [23]. Subjects are more likely to respond in a manner that affirms their ethnicity when the questionnaire is presented in the nonnative language. This type of affirmation is especially salient when questions are concerned with certain values or beliefs that are important to the respondents. When questions are concerned with less important issues to the respondents, cross-cultural accommodation is more likely to occur. A third explanation, social desirability, is that subjects tend to give more socially desirable answers especially when responding to questions in a nonnative language. All these explanations are based on the assumption that bilinguals’ discrepant responses are the result of their engaging in impression management, that is, to project an ethnic image.

Other researchers such as Ervin and Osgood [8] considered the discrepancies of bilinguals’ responses a function of language acquisition rather than an attempt to present a certain ethnic image. They distinguished between compound and coordinate bilinguals based on the way in which they acquired their languages. Compound bilinguals learn their languages in an environment where both languages are used interchangeably. Coordinate bilinguals, on the other hand, learn their languages in separate contexts. They suggested that compound bilinguals’ responses to questions in two languages should be the same, whereas coordinate bilinguals’ responses are likely to differ.

Some researchers proposed different ways of defining bilingualism. For example, Lambert [17] operationally defined compound bilinguals as those brought up in a thoroughly bilingual environment, whereas coordinate bilinguals are those who acquired their second language after age 10. In sum, bilinguals who are coordinate are more likely to differ in their responses to the same items than their compound counterparts.

3. The Committee Approach

A group of bilingual translators are used to translate the instrument from the source to the target language. Any mistake of one member can be caught by others on the committee. The disadvantages of the committee approach are that committee members may (1) have shared misconceptions, and (2) not be willing to criticize one another. This technique is best used in conjunction with other methods.

4. The Pretest Procedure

After a translation is completed, the translated instrument should be field tested to insure that it is equivalent to the original version and can be used for future studies. Brislin [1] suggested two pretest techniques. One is to randomly select subjects and ask questions about a sample of test items. Another similar procedure is to have interviewers rate the items after questioning the respondents. These procedures provide insight into the quality of translation.

Researchers usually use a combination of these methods. Errors missed by one method may be identified by another method. For example, in Chapman and Carter’s [7] study, the use of the back translation procedure is combined with the administration of both versions to bilingual respondents. The disadvantages of one method are offset by the strengths of the other.

IV. VALIDATION OF TRANSLATION EQUIVALENCE

Brislin [1] first developed five criteria to test the equivalence of source and target versions. The various methods of establishing equivalence can be further classified.
into two categories: judgmental methods and empirical methods [2, 11-13]. Judgmental methods are based on a subjective viewpoint and decision of an individual or a group on the issue of translation equivalence. They include (1) source language monolinguals check for errors, (2) bilinguals check for errors, (3) post-translation probes, and (4) performance criteria. Criterion 1 is the most used and popular judgmental design, which is a back translation design.

1. Judgmental Design

A. Comparisons of meaning between the original and back translated form

Source language monolinguals examine the two forms, and discover errors that would lead to differences in meaning in the two versions. This criterion needs the test to be back translated before source language raters can check for errors. This back translation design only provides one of many types of evidence in a translation study. Many researchers suggest the use of the back translation technique as an initial step to establish translation equivalence before applying another method [14, 16]. Some shortcomings of the back translation technique have already been discussed previously.

B. Comparisons of meaning between the original and translated form

Bilinguals examine the two versions and discover meaning errors as in criterion 1. These meaning errors found by bilinguals can be compared with those found by monolingual raters. There are also problems associated with criterion 2. First, it is difficult to find bilingual judges who have the same degree of familiarity with both languages and cultures. Second, bilingual judges may not think about a test item in the same way as the source or target language monolinguals. The results may not be generalizable to their monolingual counterparts. Thirdly, bilingual judges may also use “insightful guesses” to infer equivalence of meaning as back translators use it to generate satisfactory back translation.

C. Post-translation probe

Target language monolingual subjects are asked questions about the content of the original version when they have only read the target version. Evidence of translation equivalence is obtained based on their correct responses. The use of this method provides insight into the translation quality. However, there are also some weaknesses with this criterion: (1) possible prober-examinee communication problems due to their cultural and linguistic differences, (2) labor intensive and time-consuming compared to other methods.

D. Performance measure

If the original version asks for some sort of performance, subjects are requested to perform a task with instructions written in the target language. If the subject can complete the task then both versions are functionally equivalent. The problems associated with this criterion are that it can only be used with items that can be evaluated through bodily movements, it is labor intensive, and it is sensitive to prober-examinee interactions.

2. Empirical Design

In contrast to judgmental methods, empirical designs are based on the examinees' responses to the test items. They are often used in combination with some statistical technique. The use of judgmental and empirical methods is not necessarily independent. Judgmental methods are often used as a general check of translation equivalence before empirical methods are applied.

Based on Brislin's criterion 5, bilinguals' responses to both versions are used and compared. It is an extension of the bilingual technique as described previously. Four groups of bilingual subjects equated by randomization respectively take (1) the original language version, (2) the target language version, (3) the first-half original, second-half target, and (4) the first-half target, second-half original. If both versions are equivalent, the item frequencies and subjects' responses should be similar across groups.

Based on (1) whether bilingual subjects or monolingual subjects take the test, and (2) whether forward translation or back translation of the test is administered, Hambleton and Bollwark [12] suggested three types of empirical methods for collecting data and establishing item equivalence. These include: (1) comparison of bilinguals' responses to both source and target versions, (2) comparison of source language monolinguals' responses to both original and back-translated versions, and (3) comparison of results from source and target language monolinguals taking the instrument in their own languages.

A. Bilinguals take source and target versions

This design is based on two assumptions: first, bilinguals are equally capable of responding to test items as their monolingual counterparts in either language. Second, the use of bilinguals controls for difference in examinee ability. This is also the major advantage of this design that the confounding variable resulting from the difference in examinee ability can be controlled by having the same examinees take both versions of the test. One criticism of this design is that bilinguals may not respond to the test items in the same way as either target or source monolinguals. It would be improper to generalize
results based on this design to the population of interest. In addition, bilinguals may respond differently to the same item of two versions because they differ in the way they acquired their languages. The way bilinguals acquired their languages should be examined carefully [8].

B. Source language monolinguals take original and back translated versions

The merit of this design is that since the same examinees take both forms of the test in the same language, the result is not confounded with examinees’ ability differences. However, there is a main shortcoming of this design, which makes it very weak. That is, because target language monolinguals are not used, there is no empirical data collected on the translated version of the test. The findings may not be generalizable to the population of interest. Another problem of this design is associated with back translation. Errors made in the translated version may not be found due to insightful guesses and assumptions made by the back translators. Therefore, this design is not encouraged to be used.

C. Source language monolinguals take source version and target language monolinguals take target version

Hambleton [11] suggested that this is a very useful design. The primary advantage of this design is that both source and target language monolinguals are used and therefore the findings can be generalized to the two populations of interest. On the other hand, because two different samples of examinees are used, results may be confounded by the ability difference between the two groups of samples. Effort should be made to match examinees in the two groups on demographic variables such as age, year in school and educational background.

V. A TRANSLATION EQUIVALENCE STUDY

In the present study, a combination of judgmental and empirical designs are used to verify instrument translation. First, a judgmental method, comparisons of meaning between the original and back translated form, was used as a general check of translation equivalence. One of the most common and highly recommended procedures, back translation, is used to identify translation errors. The translated version of the instrument, the Statistical Reasoning Assessment, was back translated into the original language by three independent translators. These translators are bilingual graduate students. Two statistics professors then compared the back-translated versions with the original version. Errors in the translated version were identified and corrected before the other empirical methods were applied. Next, to pretest the translated version, both the original and the target language versions of the test were given to five graduate students. These students were familiar with both the source and target languages and knew the content of the instrument. Discrepant responses of these examinees to the same items were identified. Finally, to further demonstrate translation adequacy, the source version was administered to source language monolinguals and the second language version was administered to target language monolinguals. Results were then examined to provide insight into the translation equivalence across groups.

1. The Instrument: The Statistical Reasoning Assessment (SRA)

The Statistical Reasoning Assessment (SRA) [18] is an instrument developed by Garfield and Konold [10] to assess students’ ability to reason about and solve statistics and probability problems. It contains 20 multiple-choice items which were designed to measure eight different correct reasoning skills and eight different misconceptions. Any given item may measure more than one reasoning skill or misconception. There may be more than one correct response for each item. Therefore, an examinee may obtain 19 correct reasoning item scores and 21 misconception item scores depending on whether he or she selects the correct response(s) for each item. The original version of the SRA is in English. Selected items from this instrument have been used in many research projects in the United States and other English speaking countries such as Australia, and the United Kingdom [10]. Because there is growing interest for researchers from the international community in examining students’ conceptual understanding of probability and statistics, an international version of the instrument was produced in French and Spanish in 1994 [10].

2. Translation Technique

First, a major translation technique, “back translation”, was used as the initial step in the translation process. The researcher, as a native Chinese, translated the instrument into Chinese. Two bilingual graduate students who had some statistics background and one bilingual graduate student who was unfamiliar with the content area independently translated the Chinese version of the instrument back into English. Two professors who teach statistics in the Department of Educational Psychology at the University of Minnesota read the three back-translated versions and compared them with the original version item by item to detect any differences in meanings. Words or phrases that were different from the original meanings of the test items were identified. The researcher then examined the discrepancies and determined whether they were due to misinterpretation of meanings, inaccurate, or inadequate translation by either the researcher or
those who did the back translations. Errors found on the translated version were corrected.

Secondly, both versions of the instrument were given to five Chinese students who came from Taiwan and had taken one or two statistics course(s) at the University of Minnesota. In addition to taking the test in both languages, they were also asked to comment on the translation. They first took the Chinese version, one week later, they took the English version. Responses to the same items on these two versions of the instrument were compared to see if they were identical. Discrepant responses gave the researcher some idea (descriptive) about translation equivalence of the two versions. Two doctoral students who came from Taiwan and had taken some advanced statistics courses also reviewed the revised Chinese version of the test and provided their comments and suggestions. Some translation errors were also discovered by these students.

Finally, two native Mandarin speakers in Taiwan further contributed suggestions to revise the translated version to make it more natural-sounding in the native language. Two professors who teach business statistics at Cheng-Chi University in Taiwan also gave the researcher additional suggestions about how certain statistical terms have been commonly translated into Chinese in statistics textbooks and used in the classrooms in Taiwan. At this point, the instrument was finally ready to be administered to the subjects in Taiwan.

In summary, items with problems associated with translation were identified either by using the back translation technique or by having different people in both the United States and Taiwan, with or without statistics background, review the translated version of the instrument. Efforts were made not only to identify translations that did not preserve the original meaning, but also to avoid translations that might lead to subjects' confusion or misunderstanding of the original meaning.

3. Items Detected with Translation Problems

The first meaning error in the translation came from one of the alternatives in item 1: “Use 6.15 since it is the most accurate weighing”. The word “accurate” was first translated as “correct” (正確) in the Chinese version. It was then replaced by a word with the more accurate meaning in Chinese (精確). The second translation error was found in item 12: “The Caldwells want to buy a new car, and they have narrowed their choices to a Buick or an Oldsmobile. They first consulted an issue of Consumer Reports, which compared rates of repairs for various cars...” The word “rate” was first translated as “expense” (花費) in Chinese. It was then corrected and changed to its original meaning (比率).

One translation that might result in subjects’ misunderstanding of the original meaning appeared in the alternatives for the following items: 9, 11, 13, 19, and 20. These are all probability problems related to either flipping a coin or throwing dice. Alternatives that showed translation problems all contained the words “equally likely”. They are: “all four sequences are equally likely” in items 9 and 11, “a and b are equally likely” in item 13, and “all three results are equally likely” in item 19 and 20. The problem was first detected in one of the back translations. “Equally likely” was back translated as “all possible” by one of the bilingual graduate students. It was found that the meaning of the word “equally” was not clearly conveyed in the translated version. The test was revised to preserve the original meaning.

Another translation problem that was also discovered in one of the back translations occurred in item 18. Item 18 is as follows: “When two dice are simultaneously thrown it is possible that one of the following two results occurs: Result 1: A 5 and a 6 are obtained. Result 2: A 5 is obtained twice.” The sentence that “it is possible that one of the following two results occurs” was first translated into Chinese by retaining the original language structure. It was back translated as “two outcomes are possible”. It could be misunderstood by students that only two outcomes are possible when in fact more than two outcomes are possible. The test was modified to prevent such misunderstanding.

Some translation problems were identified by the reviewers. The word “typical” in item 4: “...She wants to summarize this data by computing the typical number of comments made that day...” was first ignored in the translation. It was then discovered that it was a key word in the item in terms of helping subjects make the correct decision.

As for the statistical term for “fair coin” in items 9, 10, and 11, one professor at Cheng-Chi University in Taiwan gave the correct translation term in Chinese (公正硬幣). In addition, “head” and “tail” of a coin in English is simply “front” (正面) and “rear” (反面) in Chinese.

4. Administration of the SRA

After considerable efforts were made in the translation process, an empirical design which compares results from source and target language monolinguals taking the instrument in their own languages was used. The test was administered to 245 college students in Taiwan and 267 college students in the United States. The two groups of samples were matched on demographic variables. All students were at the end of an introductory business statistics course when they took the test. They all majored in business-related field.
Frequencies of each item response were then examined. Comparison of the results for each group is presented in Table 1. According to Brislin [1], the item frequencies should be similar across groups if the original and the translated versions are equivalent. Comparisons of students’ responses to item 1 show that 87.8 percent of the Taiwan samples selected alternative D, the correct answer, when 78.3 percent of the US samples selected the same alternative. Only 0.8 percent of the examinees in Taiwan and 1.1 percent of their US counterparts selected alternative B. Similarly, comparisons of the examinees’ responses to item 2 shows that about 88 percent of the Taiwan samples selected alternative D, the correct answer, while 92 percent of the US samples made the same selection. No examinees in both countries selected alternative B for the

<table>
<thead>
<tr>
<th>Country/ Item response(%)</th>
<th>a</th>
<th>b</th>
<th>c</th>
<th>d</th>
<th>e</th>
<th>f</th>
</tr>
</thead>
<tbody>
<tr>
<td>1. Taiwan USA</td>
<td>5.3</td>
<td>0.8</td>
<td>5.7</td>
<td>87.8*</td>
<td></td>
<td></td>
</tr>
<tr>
<td>2. Taiwan USA</td>
<td>3.3</td>
<td>0</td>
<td>0.4</td>
<td>87.8*</td>
<td>8.6</td>
<td></td>
</tr>
<tr>
<td>3. Taiwan USA</td>
<td>34.7</td>
<td>4.1</td>
<td>10.2</td>
<td>48.6*</td>
<td>2.0</td>
<td></td>
</tr>
<tr>
<td>4. Taiwan USA</td>
<td>26.5*</td>
<td>11.4</td>
<td>60.0</td>
<td>2.0</td>
<td></td>
<td></td>
</tr>
<tr>
<td>5.1. Taiwan USA</td>
<td>5.3</td>
<td>1.6</td>
<td>2.9</td>
<td>78.4*</td>
<td></td>
<td></td>
</tr>
<tr>
<td>5.2. Taiwan USA</td>
<td>1.6</td>
<td>2.0</td>
<td>4.1</td>
<td>3.3</td>
<td></td>
<td></td>
</tr>
<tr>
<td>6. Taiwan USA</td>
<td>39.2</td>
<td>68.6*</td>
<td>49.4</td>
<td>61.2</td>
<td>3.3</td>
<td></td>
</tr>
<tr>
<td>7. Taiwan USA</td>
<td>67.8*</td>
<td>16.3</td>
<td>22.9</td>
<td>24.1</td>
<td>29.0*</td>
<td>65.7*</td>
</tr>
<tr>
<td>8. Taiwan USA</td>
<td>3.3</td>
<td>2.4</td>
<td>93.9*</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>9. Taiwan USA</td>
<td>0.4</td>
<td>3.7</td>
<td>0</td>
<td>3.3</td>
<td>93.5*</td>
<td></td>
</tr>
<tr>
<td>10. Taiwan USA</td>
<td>35.9</td>
<td>11.4</td>
<td>74.3</td>
<td>57.1*</td>
<td>4.5</td>
<td>65.3*</td>
</tr>
<tr>
<td>11. Taiwan USA</td>
<td>6.5</td>
<td>0.4</td>
<td>20.8</td>
<td>13.5</td>
<td>56.3*</td>
<td></td>
</tr>
<tr>
<td>12. Taiwan USA</td>
<td>7.3</td>
<td>83.7*</td>
<td>8.6</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>13. Taiwan USA</td>
<td>38.8*</td>
<td>32.2</td>
<td>29.0</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>14. Taiwan USA</td>
<td>8.6</td>
<td>30.6*</td>
<td>60.8</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>15. Taiwan USA</td>
<td>9.7</td>
<td>37.1</td>
<td>52.4</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>16. Taiwan USA</td>
<td>0.8</td>
<td>5.7</td>
<td>9.0</td>
<td>18.4*</td>
<td>9.0</td>
<td>61.6</td>
</tr>
<tr>
<td>17. Taiwan USA</td>
<td>13.9</td>
<td>17.6</td>
<td>76.7*</td>
<td>22.4</td>
<td>17.6</td>
<td>4.5</td>
</tr>
<tr>
<td>18. Taiwan USA</td>
<td>9.4</td>
<td>6.5</td>
<td>71.0*</td>
<td>2.0</td>
<td>25.7</td>
<td>2.4</td>
</tr>
<tr>
<td>19. Taiwan USA</td>
<td>3.7</td>
<td>3.0</td>
<td>83.5</td>
<td>2.6</td>
<td>12.7</td>
<td>3.0</td>
</tr>
<tr>
<td>20. Taiwan USA</td>
<td>73.1</td>
<td>20.4*</td>
<td>2.4</td>
<td>4.1</td>
<td></td>
<td></td>
</tr>
<tr>
<td>21. Taiwan USA</td>
<td>63.7</td>
<td>27.7</td>
<td>4.5</td>
<td>4.1</td>
<td></td>
<td></td>
</tr>
<tr>
<td>22. Taiwan USA</td>
<td>25.7*</td>
<td>2.0</td>
<td>1.6</td>
<td>70.6</td>
<td></td>
<td></td>
</tr>
<tr>
<td>23. Taiwan USA</td>
<td>31.1</td>
<td>2.2</td>
<td>3.7</td>
<td>62.9</td>
<td></td>
<td></td>
</tr>
<tr>
<td>24. Taiwan USA</td>
<td>1.2</td>
<td>46.1*</td>
<td>0.4</td>
<td>51.0</td>
<td></td>
<td></td>
</tr>
<tr>
<td>25. Taiwan USA</td>
<td>3.0</td>
<td>37.1</td>
<td>0.7</td>
<td>58.4</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

*correct alternative

Table 1. Comparison of percentage for each item response selected by Taiwan and US samples
same item. Only 0.4 percent of the students in the United States and an equal percentage of Taiwan students selected alternative C. Test results show that for the majority of the 20 multiple-choice items, when a given alternative is selected by the highest or lowest proportion of examinees in one group, the same alternative is most likely to be selected by the highest or lowest percentage of subjects in another group.

VI. CONCLUSION

Researchers have been interested in translating tests and scales into another language for as long as they have existed. One reason is that it is considerably faster and cheaper to translate an instrument than it is to construct a new one for a different language group. A second reason is that translating instruments into a second language permits cross-national comparisons. The third and most important reason is that translated tests and scales enhance validity of the results in international assessment. However, the lack of careful attention to translation equivalence can result in research that builds on error. Unless the equivalence of the original and translated tests is evaluated, the validity of any interpretation of the results remains in question. Measurement items must be tested and revised until the original meaning is understood clearly and correctly by all subjects in a second culture.

Some researchers even suggest that every effort should be made at the item writing stage before a test is translated into another language. To write English that is likely to be translatable into a target language, Brislin, Lonner, and Thorndike [2] suggested a set of rules for researchers who plan to do cross-cultural studies. Some suggestions include the use of short, simple sentences, repetition of nouns rather than the use of pronouns, employment of the active rather than the passive voice, avoidance of metaphors, and avoidance of subjective mode. It is also recommended that the schedule and budget of a cross-cultural study should provide for the time and money necessary to deal with the issue of translation equivalence.

Other researchers such as Werner and Campbell [22] noted that for every sentence in the original language, there are a large number of possible appropriate translations in the target language. It is a “one to many relationship” between the original and the target versions. They suggested that two important things in translating a source language into a target language are (1) loyalty of meaning, and (2) equal familiarity and colloquialism in each language. Literal translation from English into Chinese without changing the original language structure may cause either meaning error or poor understanding of the subjects. Even simple sentences may be poorly translated as shown in the results of back-translation. Translations of the words and phrases should be flexible to preserve the equivalence of meaning.

One of the major translation techniques used in the present study is back translation. Back translation is a very useful technique while translating an instrument from the source to the target language. However, it can only be used as an initial check of translation equivalence. Therefore, a combination of translation procedures along with empirical methods are used in the study to verify the translation of the instrument. Also, problems identified in the original version of the test during the validation process are carefully examined, so similar problems won’t occur when the test is translated into another language for use in future studies. It should be noted that researchers worry that test results may be confounded by the ability differences between the two groups when both source and target language monolinguals are used. It is recommended that in future studies more efforts should be made to match examinees in the two groups on the ability. Furthermore, statistical analyses of subject responses such as examination of the means, standard deviations, correlation coefficients, and factor structures of the two language forms should be used in combination with judgmental methods in the future to ensure that the original and translated versions are equivalent.

REFERENCE


Received: Jun. 7, 2002 Revised: Oct. 4, 2002 Accepted: Oct. 18, 2002