# The Mexican Translation of TIMSS-1995: Lessons on Test Translation From a Post-Mortem Study

**Guillermo Solano-Flores**
*University of Colorado, Boulder*
**Luis Ángel Contreras-Niño**
*Universidad Autónoma de Baja California*
**Eduardo Backhoff**
*Universidad Autónoma de Baja California, Instituto Nacional para la Evaluación de la Educación*

Paper Presented the Annual Meeting of the National Council on Measurement in Education, Montreal, Canada, April 12-14, 2005.

The translation of tests into languages in which they were not originally written, and their adaptation for groups not originally targeted, has become an important component of education in the context of a global economy (Hambleton, 1994). International comparisons have posed important challenges for the development of measurement instruments in more than one language. As a result, norms and procedures for the linguistic adaptation of tests keep evolving (Grisay, 1998).

This paper is a contribution to test translation review. We address the fact that, while necessary, guidelines for test translation are not sufficient for ensuring the validity of translated tests. We describe an approach to test translation review that addresses a wider variety of translation issues, from production of translated tests, to curriculum representation, to social aspects of language use and language usage. More specifically, we offer a conceptual framework for the coding of translation errors, and provide some empirical evidence on the effect of translation errors on student performance.

We examined the Mexican translation of the Third International Mathematics and Science Study (TIMSS-1995) test. Mexican data from TIMSS-1995 are largely unknown because Mexico withdrew its participation after data had been collected but before the results were published. Examining this test addresses expectations of transparency that the Mexican public now has regarding education issues. While the test is ten years old, lessons learned from that experience will inform internal decisions about Mexico's participation in future international comparisons. This effort is funded by the National Institute for Educational Evaluation (INEE), whose creation in 2002 initiated in Mexico an era of public awareness of evaluation and sensitivity to accountability issues.

Most of the Mexican TIMSS-1995 data were destroyed. However, soon after the creation of INEE, we were able to recover blank copies of all the test booklets used with Population 1 (grade 3 and 4) and Population 2 (grade 7 and 8) students. We were also able to recover information on the items' p-values (the proportion of students who responded correctly to the items) for two years: 1995—in which data were collected as part of Mexico's participation in the international comparison—and 2000—in which the Mexican Ministry of Education's General Directorate of

Evaluation (DGE) administered the same test to a new sample of students (see Backhoff & Solano-Flores, 2003).

The lack of information at the student level (such as external measures of academic performance, demographic information, or information on linguistic abilities) limited the kinds of analyses that we could perform. Our analyses focus on the frequency of translation errors observed across populations, grades, and content areas, and on the effect of translation quality on student performance, as reflected by the correlations of item translation quality measures and item p-values.

## Conceptual Framework

Our conceptual framework for test translation review was developed based on (1) an initial examination of the Mexican TIMSS-1995 items (Solano-Flores & Backhoff, 2003); (2) comments from multidisciplinary translation review panels during a series of pilot translation review sessions (see below); (3) the test translation guidelines developed for TIMSS (Hambleton, 1994); (4) the set of criteria used by TIMSS as part of its translation verification procedures (Mullis, Kelly y Haley, 1996); (5) the criteria used by the American Translators Association (2003) to assess the quality of translations; (6) the set of norms and considerations currently used by PISA which address item cultural appropriateness (Grisay, 1998; Maxwell, 1996); (7) evidence from research on the use of structural linguistics for detecting unnecessary complexity in the syntax of sentences in test items (Solano-Flores, Trumbull, & Kwon, 2003); and (8) evidence on the sociolinguistic and epistemological factors that influence students' interpretations of science and mathematics items (Solano-Flores & Nelson-Barber, 2001; Solano-Flores & Trumbull, 2003).

We identify ten test translation error dimensions (Table 1). The obvious dimensions have to do with translation accuracy, grammatical correctness (Grammar, Semantics), and editorial features and the production of translated tests (Style, Format, Conventions). Other dimensions address translation accord with language use and usage by the target population in its social and instructional contexts (Register), content (Information, Construct), and curriculum representation (Curriculum). A tenth dimension (Origin) addresses the fact that flaws in the item in the source language may be carried over to the translation.

Basic to our conceptual framework is the notion that an error can be classified along multiple error dimensions. For example, the improper insertion of a comma in a sentence in a translated item may violate some grammar conventions, but it also may change the intended meaning of an idea. Whereas, in general, Style, Format, and Convention errors tend to be milder than those belonging to other dimensions, there are exceptions to this rule. Suppose that a figure in the translated version of an item has font styles that are different from the font styles used in the original version. In the simplest case, the difference does not have any effect; in the most extreme case, the difference affects the attention that students pay to some components of the figure and even their interpretations of the item.

Also basic to our conceptual framework is its probabilistic view of the effects of translation flaws on item validity. First, rather than being absolute, the severity of a translation error is shaped by contextual factors and the characteristics of the target population. Thus, whether the improper insertion of a comma in the example above is a severe or mild grammar error or a severe or mild semantics error depends, in part, on the characteristics of the item (e.g., the kind

and amount of information it provides), the students' knowledge of the content being assessed, and the students' ability to figure out the intended meaning of the item.

Table 1. Test translation error dimensions.

| | |
|---|---|
| Style | The style with which the item is written in the target language is not in accord with the style used in textbooks and printed materials in the country. Examples: punctuation errors; inappropriate use of uppercase or lowercase letters; subject-verb inconsistency. |
| Format | The format or visual layout of the translated item differs from the original. Examples: different size of tables; different font style; narrower margins; insertion or omission of graphic components. |
| Conventions | The translation of the item is not in accord with accepted item-writing practice in the target language or country, or to basic principles of item writing. Examples: grammatical inconsistency between the stem and the options of multiple choice items; grammatical inconsistency among options in multiple choice items, different font size of the correct response in multiple choice items. |
| Grammar | The translation of the item has grammatical errors or the syntax is unnecessarily complex or unusual amongst the target population. Examples: literal (word-by-word) translation; unnatural syntactical structure; inappropriate use of prepositions. |
| Semantics | The ideas and meaning conveyed in the translated item are not the same as in the item in the source language. Examples: use of false cognates; inappropriate translation of idiomatic expressions. |
| Register | The translation of the item is not sensitive to the target population's word usage and social contexts. Examples: use of words of low frequency among the target population; technical terms translated correctly but in a way that is not common in the schools or textbooks in the country. |
| Information | The translation changes the amount, quality, or content of information critical to understanding what the item is about and what has to be done to respond to it. Examples: inconsistent translation of a term that appears several times in the original; a key term appears more or fewer times in the translation than in the original; numbers are written in ways that differ from the original. |
| Construct | The translation changes the knowledge or skills needed to respond to the item correctly. Examples: inaccurate translation of technical terms; insertion or omission of technical terms. |
| Curriculum | The item does not represent the curriculum of the target country. Examples: the knowledge or skill assessed by the item is not taught at the grade level targeted by the test; the way in which a problem is posed is not used in the curriculum of the target country. |
| Origin | The item in the source language has flaws that cannot be corrected in the translation and which pose limits to the effectiveness of its translation. Examples: there is more than one correct option; none of the options is entirely correct. |

Second, how likely a translated item is to be linguistically challenging is determined by the combined effect of the frequency and severity of translation errors. This notion can be represented as a probabilistic space, as shown in Figure 1. The light and shaded areas represent the likelihood of an item to be sound or challenging. Linguistically challenging items are likely to occur only when they have many mild translation errors or few, very serious translation errors, or both.

In our view, linguistically sound items are not necessarily error-free and flawed items are not necessarily linguistically challenging. From the most rigorous theoretical stand, no perfect translation exists and, in the case of tests, it is not possible to keep exactly the same meaning and to measure the same construct across languages (Greenfield, 1997). At the same time, *error* does not necessarily mean *fatal error*, because, to a certain extent, individuals are capable of identifying and dealing cognitively with certain translation errors or below a certain maximum number or errors.

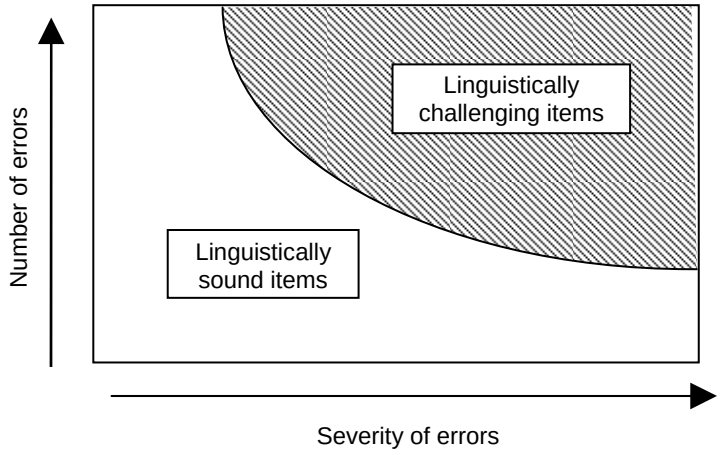


Figure 1. Probabilistic space of linguistically challenging items, defined by the frequency and the severity of translation errors.

## Methods

### Item Sample and Analyses

We examined the linguistic challenges of a sample of 319 items used in the TIMSS-95 test administered in Mexico in 1995 (as part of its participation in TIMSS) and in 2000. This sample excluded common items (i.e., items given to students from both populations). Of these 319 items, 88 were from the mathematics item pools and 81 were from the science item pools used with Population 1; 76 were from the mathematics item pools and 74 were from the science item pools used with Population 2.

We also performed a series of Pearson correlation analyses to determine how our measures of the translation quality of the items correlated with their p-values, calculated from the records of students from Population 1 (N=10,122 for grade 3 and N=10,194 for grade 4) and students from Population 2 (N=12,809 for grade 7 and N=11,843 for grade 8). The items included in these correlation analyses were actually a sub-sample of the original 319 items, because p-values were available only for 42 mathematics and 39 science Population 1 items and for 19 mathematics and 23 science Population 2 items.

### Participants and Procedures

We assembled two translation review panels composed of specialists with different backgrounds. One panel reviewed the Population 1 items. This panel was composed of one elementary school curriculum expert, two fourth-grade teachers, one linguist, one professional translator certified by the American Translators Association, and a psychologist and a psychometrician, both with extensive experience in the field of test development.

The other panel reviewed the Population 2 items. This panel was composed of one secondary school curriculum expert, one teacher for each of the four content areas taught in the Mexican

secondary school curriculum—mathematics, physics, chemistry, and biology—and the linguist, the professional translator, the psychologist, and the psychometrician who participated in the Population 1 review panel.

Whereas the linguist, the psychologist, the psychometrician, and the translator were fluent in English, the teachers and the curriculum experts in both panels had a limited English knowledge. This difference in the command of English was viewed as an advantage, as our intent was to obtain from the curriculum experts and the teachers an evaluation perspective of the items that was not influenced by their understanding of the items in the source language.

**Translation Review Procedure**

Prior to the formal translation review sessions, we facilitated a two-day pilot review session with each panel to refine both our test translation review conceptual framework and the test translation review procedure, and to train the members of the two panels in the use the coding form.

The field test translation review sessions lasted five days for each panel. Each of the items selected was reviewed according to the following procedure. Each reviewer was given a hard copy of the item in Spanish and was asked to read it and respond to it as if he or she were a student taking the test. This was done to ensure that each reviewer would become well acquainted with the item, aware of the kind of reasoning that the item elicited, and aware of the kind of knowledge needed to respond to it correctly. The original English version of the item was projected on a screen so that reviewers could compare the two language versions of the item. However, reviewers were instructed not to look at it until they had responded to it. Then they examined the item and independently recorded on a coding form the kinds of translation errors they identified and, when necessary, provided an explanation that supported their coding decisions (Figure 2).

| Item # | Type of Error | Code and Justification |
|---|---|---|
| | 1. Style | |
| | 2. Format | |
| | 3. Conventions | |
| | 4. Information | |

Figure 2. Translation of a portion of the form used to code translation errors.

The reviewers were instructed to look for errors in all of the dimensions, and were asked to focus on specific dimensions according to their professional backgrounds. Thus, the psychologist and the psychometrician focused on the Style, Format, Conventions, Origin, Information, and Construct dimensions; the translator focused on the Style, Format, Grammar, and Semantics dimensions; the linguist focused on the Grammar, Semantics, and Register dimensions; and the teachers and curriculum experts focused on the Instruction, Content, and Curriculum dimensions.

Each reviewer's error coding decisions were reviewed by the panel. When necessary, error coding decisions were discussed by the panel until consensus was reached. In some cases, the discussion allowed the panel to identify errors not detected by any of the reviewers individually.

Among these cases were those in which the quality of the translation was limited by flaws of the item in the English version—these were recorded as Origin dimension errors. Also, the panel agreed to flag some items as linguistically challenging when they identified many mild, and/or few but severe, translation errors. We refer below to those items as *red flag items*.

When necessary, during the test translation review sessions, the teachers and content experts used key curriculum documents such as the official textbooks and the official teacher guides. Also, information released by TIMSS on the content of the items and their targeted skills and knowledge was made available to the review panels, so that proper interpretations of the intended meaning of the items in both the source and the target languages could be made.

Reviewers were instructed to update their review forms at the end of the discussion for each item so that these forms reflected the conclusions arrived at by the panel. Data entering and analysis was based on the number of different errors identified by the panel in each dimension, as reflected by the review forms.

## Results

### Frequency of translation errors

We observed translation errors for nearly all of the items in the sample. The vast majority of the items had errors in two or more translation error dimensions. On average, items had errors in about four dimensions (Figure 3).

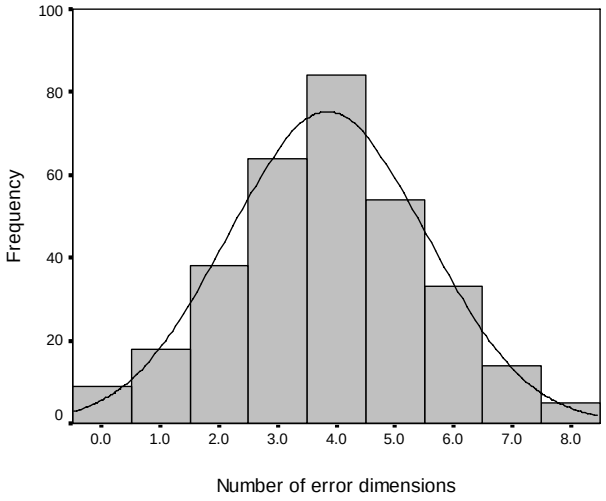


Figure 3. Frequency distribution of items along number of translation error dimensions. K=319; Mean=3.84; S.D.=1.68; Median=4.00; Mode=4.00.

In general, we observed similar patterns of frequency of errors across populations and content areas (Table 2). For example, the most frequent errors observed among Population 1 and Population 2 items and among mathematics and science items belonged to the Semantics (82.0%-89.8%) and Format (75.0%-83.8%) categories. However, important differences between populations or between content areas were observed for some dimensions. One of these dimensions is Register, whose frequencies were different across populations and across content areas within populations. Register errors were more frequent among mathematics items than among science items for Population 1 (18.2% vs. 9.9%); but more frequent among science items than among mathematics items for Population 2 (23.0% vs. 17.1%). Another dimension in which

this occurred is Curriculum. Errors in this dimension were much more frequent for Population 1 (12.5% for mathematics and 17.3% for science) than for Population 2 (3.9% for mathematics and 2.7% for science).

Table 2. Translation errors by population, content area, and error dimension, in percentages of items.

| Error Dimension | Population 1 | | Population 2 | |
|---|---|---|---|---|
| | Mathematics (k=88) | Science (k=81) | Mathematics (k=76) | Science (k=74) |
| Style | 17.0 | 28.4 | 21.1 | 29.7 |
| Format | 75.0 | 76.5 | 82.0 | 83.8 |
| Conventions | 21.6 | 37.0 | 11.8 | 35.1 |
| Grammar | 36.4 | 39.5 | 36.8 | 47.3 |
| Semantics | 89.8 | 85.2 | 82.9 | 87.8 |
| Register | 18.2 | 9.9 | 17.1 | 23.0 |
| Information | 69.3 | 64.2 | 44.7 | 60.8 |
| Construct | 23.9 | 27.2 | 15.8 | 33.8 |
| Curriculum | 12.5 | 17.3 | 3.9 | 2.7 |
| Origin | 17.0 | 14.8 | 17.1 | 21.6 |

Some of these differences in error frequencies derive from influences of the discursive styles that are more common in one content area than in the other. For instance, in many mathematics items we observed an error that consists of collapsing two sentences (see Version 1, below) into one and changing the grammatical mode into a conditional form (Version 2):

Version 1: *Danny buys three pencils that cost 14 cents each. How much money does he need?*

Version 2: *If Danny buys three pencils that cost 14 cents each, how much money does he need?*

**Correlations Between Translation Quality and Student Performance**

We used the number of translation errors as a measure of translation quality. Measures based on weighing the severity of translation errors according to their dimensions were discarded because, as discussed above, error severity is shaped by a multitude of contextual characteristics of the item. In addition, weighted scores produced very skewed distributions.

One piece of evidence regarding the sensitivity of our test translation review conceptual framework was obtained by comparing the number of errors in red flag items (those identified by the panels as linguistically challenging because they identified many mild, and/or few but severe, translation errors) with the number of errors in non-flagged items. We found that red flag items have a higher mean number of errors (Table 3). Needless to say, the small number of red flag items prevented us from testing the statistical significance of this difference.

Table 3. Number of errors: Descriptive statistics for items which were and were not identified as red flag items.

| | Mean | S.D. | Median | S.E.M. | Mode | Range |
|---|---|---|---|---|---|---|
| Non-flagged items (n=295) | 6.25 | 3.37 | 6.00 | .19 | 8.00 | 0-17 |
| Red flag items (n=24) | 8.83 | 3.19 | 8.50 | .64 | 8.00 | 3-20 |

As mentioned before, the lack of disaggregated data at the student level and the lack of information on additional variables (e.g., TIMMS-1995 individual student scores, other measures of academic achievement, and language ability) that could be used as covariates limits the kinds of analyses that can be performed to assess test translation quality as a predictor of student performance. Thus, our analyses are limited to correlating our measures of translation quality with the items' p-values.

In principle, negative correlations should be observed when translation errors have an adverse effect on student performance. However, certain translation errors may attenuate the magnitude and the level of significance of these correlations. Inadequate translation is not always biased against the population tested in the target language (Solano-Flores, Trumbull, & Nelson-Barber, 2002). For example, using a key term more times than in the original language version may potentially bias an item in favor of the population tested in the target language. Also, while some important translation errors may impact the content validity of an item, they may not necessarily have observable effects on student performance.

These reasons and the limited statistical power of correlations involving Population 2 items (19 for mathematics and 23 for science) force us to focus on the magnitude and direction of these correlations rather than on their statistical significance and on the pattern of these correlations across populations and content areas.

We observed that item p-values and the number of translation errors correlated differently depending on population and content area (Table 4). For Population 1, negative correlations were observed for the mathematics items and positive correlations were observed for the science items. For Population 2, the opposite pattern is observed: negative correlations were observed for science items and positive correlations were observed for mathematics items. These patterns are consistent across grades and across the two years in which the test was administered. In both populations, negative correlations are consistently higher than positive correlations.

Table 4. Pearson correlations between item p-value and number of translation errors by population, content area, and year of administration.

Population 1

| | | Grade 3 | | Grade 4 | |
|---|---|---|---|---|---|
| Content Area | n | Year 1995 | Year 2000 | Year 1995 | Year 2000 |
| Mathematics | 42 | -.233 | -.245 | -.262 | -.333* |
| Science | 39 | .026 | .114 | .024 | .075 |
| Combined | 81 | -.119 | -.081 | -.128 | -.139 |

Population 2

| | | Grade 7 | | Grade 8 | |
|---|---|---|---|---|---|
| Content Area | n | Year 1995 | Year 2000 | Year 1995 | Year 2000 |
| Mathematics | 19 | .157 | .102 | .130 | .132 |
| Science | 23 | -.245 | -.187 | -.267 | -.252 |
| Combined | 42 | .048 | .052 | .003 | .039 |

*Significant ($p<.05$)

We cannot attempt to interpret these results in terms of the interaction between defective translation and the set of linguistic demands intrinsic to each content area. Indeed, this pattern of differences may be only a reflection of the different levels of skills among the translators who translated the science and mathematics items and the Population 1 and Population 2 items.

However, since we do not have information about the process with which the test was translated either, all we can conclude is that that our approach was sensitive enough to translation error to reveal systematic differences across content area and population in the correlations between translation error and p-values.

## Summary and Conclusions

Our knowledge of the delicate aspects of test translation has increased in the last ten years. For example, research in international comparisons has shown that even a slightly inaccurate translation of a word may influence an item's differential functioning (Ercikan, 1998). The procedures used in the translation of tests in international comparisons have evolved accordingly. For example, PISA and TIMMS now use two source languages to ensure that meaning is preserved in translation (Grisay, 1998). Also, the procedure of back translation—in which a translated test is translated back to the original language and the back-translated and original versions of the test are compared to identify and correct for any substantial content differences—is no longer accepted as irrefutable proof of equivalence across languages.

However, in spite of that progress, test translation review has not received the same kind of attention. For example, the cultural aspects associated with language are not always addressed formally. In current test translation practices, it is usually assumed that using translators who are familiar with the culture of the target population is sufficient to address aspects of language usage (e.g., Behling & Law, 2000). Yet, as we have discussed here, translation errors related to the social aspects of language use cannot be identified unless a detailed coding system is in place.

We have reported a study on the quality of the translation of the Mexican version of TIMSS-1995. We presented a conceptual framework for test translation review and described how we used that conceptual framework with an interdisciplinary panel of reviewers to identify and code translation errors in a sample of items from that test. Our conceptual framework is very sensitive to any kind of error, from minor production and printing errors to semantics errors and errors of curriculum representation. It also incorporates error dimensions that address language use and language usage in social and instructional contexts.

We observed translation errors in the vast majority of the items. On average, items had translation errors in about four error dimensions. The most frequently observed errors belonged to the Format and the Semantics dimensions.

An examination of the correlations between the number of errors identified in the items and their p-values (the proportion of students who responded to them correctly) suggests that translation errors' effects on student performance depend on content area and grade. For Population 1 (Grades 3 and 4), these correlations were consistently negative for the mathematics items and consistently positive for the science items. The opposite effect was observed for Population 2 (Grades 7 and 8); these correlations were consistently negative for the science items and consistently positive for the mathematics items.

Negative correlations were consistently higher than positive correlations, which supports the notion that translation errors tend to bias the items against the target students. We believe this differential effect of translation quality across content area and grade is due to the fact that different content areas (and the items that intend to assess knowledge in those areas) pose different linguistic demands at different grade (age) levels.

The finding that almost all of the items examined were found to have translation errors may be shocking on first impression. However, these high frequencies can be deceiving if they are not interpreted properly. For one thing, we need to emphasize that *error* does not mean *fatal error*. Our deficit scoring procedure—which is designed to detect errors rather than simply decide if items are acceptable—is based on the assumption that perfect test translation is virtually impossible. It is the additive effect of many and/or critical errors which make an item linguistically challenging for the target language.

Still, we believe that the frequencies of errors observed are tremendously high. We identify two reasons for that. The first, of course, is that the quality of the test translation was indeed poor. While we have not been able to find information that documents the process used to develop the Mexican translation of TIMSS-1995 or the professional qualifications of the individuals who translated the test, it was clear to the test translation review panels that the translation of some items was objectionable. Case in point is the set of 24 items (over 7% of the total of items examined) identified by the panels as red flag items, and which somehow made their way to the final version of the test.

Second, our conceptual framework appears to be more sensitive to translation errors than any other approach to test review. Because it addresses aspects of language not usually examined in detail by other approaches (such as semantics, curriculum representation, and language use and usage in the social and instructional contexts), it might also reveal many translation errors in other language versions of TIMSS-1995.

We have to say, in addition, that the high frequency of errors observed not only speaks to the quality of the Mexican translation of the test but also to the need for revising existing paradigms for test translation and test translation review. The Mexican version of TIMSS-1995 went through a process of certification based on standards used by TIMSS before it was approved for use with Mexican students. A report on the quality of TIMSS-1995 data includes a chapter on the procedures used to review the quality of the test translations (Mullis, Kelly, and Haley, 1996) developed by the participating countries; it also provides the numbers of items identified as problematic for each country. Amazingly, according to that report—which probably went to press at the same time Mexico was withdrawing its participation—none of the items used by Mexico had been identified as problematic.

The tremendous inconsistency between that report and our results underscores the need for developing more robust methodologies for test translation and test translation review. This study has been a step in that direction.

A final word on costs: we need to stress the fact that facilitating test translation review panels and using our conceptual framework was not expensive. Cost is not (and should not be) an obstacle for enhanced test translation procedures.

## References

American Translators Association: Framework for standard error marking and explanation. http://www.atanet.org retrieved October 10, 2003.

Backhoff, E., & Solano-Flores, G. (2003). Third International Study of Mathematics and Science (TIMSS): 1995 and 2000. (Sp.) Mexico: National Institute for Educational Evaluation.

Behling, O. & Law, K.S. (2000). Translating questionnaires and other research instruments: Problems and solutions. Thousand Oaks, CA: Sage.

Ercikan, K. (1998). Translation effects in international assessment. *International Journal of Educational Research, 29*, 543-553.

Greenfield, P.M. (1997). You can't take it with you: Why ability assessments don't cross cultures. American Psychologist, 52 (10) 1115-1124.

Grisay, A. (1998). Translation and cultural appropriateness of the test and survey material.

Hambleton, R.K. (1994). Guidelines for adapting educational and psychological tests: A progress report. *European Journal of Psychological Assessment, 10*(3), 229-244.

Maxwell, B. (1996). Translation and cultural adaptation of the survey instruments. En Martin, M.O. y Kelly, D.L. (eds). *Third International Mathematics and Science Studdy (TIMSS Technical Report, Volume 1: Design and Development*. Chestnut HIll, MA: Boston College.

Mullis, I.V.S., Kelly, D.L., and Haley, K. (1996). "Translation Verification Procedures" in M.O. Martin and I.V.S. Mullis *Third International Mathematics and Science Study: Quality Assurance in Data Collection.* Chestnut Hill, MA: Boston College.

Solano-Flores, G., & Backhoff, E. (2003). Test translation in international comparisons: A preliminary study. (Sp.) Mexico: Mexican Department of Education, National Institute for Educational Evaluation.

Solano-Flores, G., & Nelson-Barber, S. (2001). On the cultural validity of science assessments. *Journal of Research in Science Teaching, 38(*5*)*, 553-573.

Solano-Flores, G., & Trumbull, E. (2003). Examining language in context: The need for new research and practice paradigms in the testing of English-language learners. *Educational Researcher, 32*(2), 3-13.

Solano-Flores, G., Trumbull, E., & Nelson-Barber, S. (2002). Concurrent development of dual language assessments: An alternative to translating tests for linguistic minorities. *International Journal of Testing, 2*(2), 107-129.

Solano-Flores, G., Trumbull, E., & Kwon, M. (2003). The metrics of linguistic complexity and the metrics of student performance in the testing of English language learners. Symposium paper presented at the 2003 Annual Meeting of the American Evaluation Research Association. Chicago, IL, April 21-25.

## Authors' Note

This investigation was funded by the Instituto Nacional para la Evaluación de la Educación (INEE) through a research contract with the Universidad Autónoma de Baja California (UABC). We are grateful to our colleagues at UABC and to the educators who took part in this project for their enthusiastic participation as translation reviewers. We are especially grateful to Dr. Larry Suter for his interest in this project.

Authors' e-mail addresses:

Guillermo Solano-Flores: Guillermo.Solano@Colorado.edu

Luis Ángel Contreras-Niño: angel@uabc.mx

Eduardo Backhoff: backhoff@ineemexico.org