Metadata of the chapter that will be visualized online

Series Title		
Chapter Title	The Measurement of Translation Error in PISA-2006 Items: An Application of the Theory of Test Translation Error	
Chapter SubTitle		
Copyright Year	2012	
Copyright Holder	Springer Science + Business Media Dordrecht	

Corresponding Author	Family Name	Solano-Flores
	Particle	
	Given Name	**Guillermo**
	Suffix	
	Division	School of Education
	Organization	University of Colorado at Boulder
	Address	249 UCB, 80309, Boulder, CO, USA
	Email	Guillermo.Solano@Colorado.Edu
Author	Family Name	Contreras-Niño
	Particle	
	Given Name	**Luis Ángel**
	Suffix	
	Division	Instituto de Investigación y Desarrollo Educativo
	Organization	University of Baja California
	Address	Km. 103 Carretera Tijuana-Ensenada,, c.p. 22830, Ensenada, BC, Mexico
	Email	angel@uabc.edu.mx
Author	Family Name	Backhoff
	Particle	
	Given Name	**Eduardo**
	Suffix	
	Division	Instituto de Investigación y Desarrollo Educativo
	Organization	University of Baja California
	Address	Km. 103 Carretera Tijuana-Ensenada,, c.p. 22830, Ensenada, BC, Mexico
	Email	backhoff@uabc.edu.mx

Abstract

We examined the translation of PISA test items based on the theory of test translation error (TTTE), which has proven to allow detection of translation errors with unprecedented levels of detail. Translation error (TE) is defined as the lack of equivalence between the original and translated versions of items on multiple translation error dimensions (TEDs) that involve design, language, and content. According to the theory, TE results not only from poor translation, but also from factors that are beyond the translators' skills (e.g., languages encode meaning in different ways). We examined the Mexican, Spanish language translation of science and mathematics PISA 2006 items. A panel comprising teachers, translators, a linguist, a test developer, and a measurement specialist examined the translation of 193 text analytical units (55 pieces of introductory text and 138 items) and identified and coded the TEs identified on ten TEDs. For each item, TE was measured as the number of different TEDs on which the review panel identified TEs. To determine which TEDs are critical to student performance, we examined the correlation between TE and item difficulty (percentage of correct answers and mean proportional score, respectively for dichotomous and non-dichotomous items) considering different sets of TEDs. The highest correlations were observed

for the sets that included the dimensions, Grammar, Semantics, Register, Information, Construct, and Culture. We also observed different magnitudes of correlations for science and mathematics items and a stronger, statistically significant correlation for translated items whose translation the review panel identified more objectionable than for the rest of the items. These results confirm that language- and content-related TEs may threat the validity of translated items. They speak to the value using the TTTE as a formative evaluation tool that PISA countries can use to operationalize translation guidelines.

Solano-Flores, G., Contreras-Niño, L.A., & Backhoff, E. (2012). The measurement of translation error in PISA-2006 items: An application of the theory of test translation error. In Prenzel, M., Kobarg, M., Schöps, K., & Rönnebeck, S. (Eds.), *Research in the Context of the Programme for International Student Assessment*. Springer Verlag.

Chapter 8
The Measurement of Translation Error in PISA-2006 Items: An Application of the Theory of Test Translation Error

Guillermo Solano-Flores, Luis Ángel Contreras-Niño, and Eduardo Backhoff

[AU1] **Abstract** We examined the translation of PISA test items based on the theory of test translation error (TTTE), which has proven to allow detection of translation errors with unprecedented levels of detail. Translation error (TE) is defined as the lack of equivalence between the original and translated versions of items on multiple translation error dimensions (TEDs) that involve design, language, and content. According to the theory, TE results not only from poor translation, but also from factors that are beyond the translators' skills (e.g., languages encode meaning in different ways). We examined the Mexican, Spanish language translation of science and mathematics PISA 2006 items. A panel comprising teachers, translators, a linguist, a test developer, and a measurement specialist examined the translation of 193 text analytical units (55 pieces of introductory text and 138 items) and identified and coded the TEs identified on ten TEDs. For each item, TE was measured as the number of different TEDs on which the review panel identified TEs. To determine which TEDs are critical to student performance, we examined the correlation between TE and item difficulty (percentage of correct answers and mean proportional score, respectively for dichotomous and non-dichotomous items) considering different sets of TEDs. The highest correlations were observed for the sets that included the dimensions, Grammar, Semantics, Register, Information, Construct, and Culture. We also observed different magnitudes of correlations for science and mathematics items and a stronger, statistically significant correlation for translated items whose translation the review panel identified more objectionable than for the rest of the items.

[AU2] G. Solano-Flores, Ph.D. (✉)
School of Education, University of Colorado at Boulder,
[AU3] 249 UCB, Boulder, CO 80309, USA
e-mail: Guillermo.Solano@Colorado.Edu

L.Á. Contreras-Niño, Ph.D. • E. Backhoff, Ph.D.
University of Baja California, Mexico, Km. 103 Carretera
Tijuana-Ensenada, c.p. 22830 Ensenada, BC, Mexico
e-mail: angel@uabc.edu.mx

M. Prenzel et al. (eds.), *Research on PISA: Research Outcomes of the PISA Research Conference 2009*, DOI 10.1007/978-94-007-4458-5_8,

These results confirm that language- and content-related TEs may threat the validity of translated items. They speak to the value using the TTTE as a formative evaluation tool that PISA countries can use to operationalize translation guidelines.

Increased awareness of the tremendous sensitivity of tests to language (e.g., Allalouf, 2003; Ercikan, 1998; Ercikan, Gierl, McCreith, Puham, & Koh, 2004; Gierl, Rogers, & Klingner, 1999; Grisay, 2007) in the context of international test comparisons has resulted in recent years in substantial improvements of test translation and adaptation procedures used by PISA (e.g., Grisay, 2003; Harkness, van de Vijver, & Mohler, 2003). As part of these improvements, revised sets of test translation guidelines (e.g., Halleux-Monseur, 2008; Hambleton, 1994; Hambleton, Merenda, & Spielberger, 2005; van de Vijver & Poortinga, 2005) have been made available for participating countries.

Unfortunately, whereas these revised procedures and guidelines are necessary, their implementation and interpretation may not be optimal without procedures that allow countries to perform detailed, systematic evaluations of their own translation work. Available evidence from research on the testing of linguistically diverse populations indicates that, in the absence of tools for systematically examining and discussing the linguistic features of items, reviewers may not be able to detect potential linguistic challenges of those items if they rely solely on their judgment (Solano-Flores & Gustafson, in press; Solano-Flores, Trumbull, & Kwon, 2003).

This need for conceptual tools in test translation led us to propose a theory of test translation error (TTTE; Solano-Flores, Backhoff, & Contreras-Niño, 2009) which defines translation error (TE) (Note 1) as the lack of equivalence between the original language version and the translated version of items. This lack of equivalence can be examined along multiple dimensions that have to do with the design or visual layout of the items (e.g., format, style), their linguistic features (e.g., grammar, syntax), and their content (e.g., information, construct).

The theory postulates that error in the translation of tests is inevitable. In addition to a poor translation job, TE is due to factors that are beyond the translators' skills. For example, languages encode meaning differently and have different sets of grammatical rules. In addition, TE is multidimensional—an error may involve multiple aspects of language (e.g., the lack of. a comma is a punctuation error but it also may be an error that alters the intended meaning of a sentence). Due to these reasons, and given the linguistic characteristics that are typical among test items (e.g., limited contextualization, high semantic load of terms, compact sentences), it is virtually impossible to preserve exactly the same meaning and linguistic complexity of items across languages.

The notion of test TE as something that cannot be eliminated but can be minimized should be easy to understand by professionals in the educational measurement community. As with measurement error, TE is due to multiple factors (and their interaction), many of which are beyond control. According to the TTTE, effective test translation can minimize, not eliminate, TE. Flawed translated items have many and/or serious TEs; acceptable translated items have few and/or mild TEs.

Conventional translation review procedures focus on determining whether translated items can be accepted (e.g., Grisay, deJong, Gebhardt, Berezner, & Halleux-Monseur, 2007; Mullis, Kelly, & Haley, 1996). They reflect researchers' and evaluators' tendency to emphasize confirming evidence over disconfirming evidence in hypothesis testing (see Church, 1991; Creswell & Miller, 2000). Unlike conventional translation review procedures, a TTTE-based approach focuses on looking for evidence that disconfirms the notion that the translation of test items is correct. We contend that this approach results in more rigorous translation review procedures.

We have used the TTTE to code errors in translated items and develop measures of TE in those items. Moreover, we have been able to link TE and student performance by correlating item difficulty with measures of TE (Backhoff, Contreras-Niño, & Solano-Flores, 2011; Solano-Flores, Backhoff, & Contreras-Niño, 2006). Our findings have shown consistently that translation review based on the TTTE allows detection of TEs with a level of detail not attained with conventional test translation review procedures (Solano-Flores, Contreras-Niño, & Backhoff, 2005; Solano-Flores, Contreras-Niño, & Backhoff, 2006).

In this chapter, we show how detection and measurement of TE can contribute to improved PISA translation procedures. More specifically, we show how coding and measuring TE based on the TTTE allows identification of serious errors in PISA translated items otherwise regarded as acceptable according to conventional translation verification procedures.

Previous empirical evidence showing the sensitivity to TE of review procedures based on the theory comes from reviews of TIMSS items and relatively small samples of released PISA items (Solano-Flores et al., 2005, 2006). In this study, we reviewed a considerably larger sample of items and took into consideration the structure of many of the PISA items—assessment units consisting of one or several paragraphs with contextual information and one or more items related (see Bybee, McCrae, & Laurie, 2009).

8.1 Theoretical Framework

8.1.1 Definition of Translation Error

The theory of test translation error (TTTE) is not only about errors made in test translation, but also about errors in translated tests. According to our theory (Solano-Flores et al., 2009), *test translation* does not refer exclusively to the action of translating items but also to multiple aspects of the entire process through which translated versions of those items are created. *Translation error* does not result exclusively from poor translation job (e.g., inaccuracy of a chosen term, word-by-word translation, use of false cognates); it also results from factors that are beyond the translators' translation skills.

t1.1 **Table 8.1** Acceptability-objectionability of translated
t1.2 items according to the frequency and severity of test
t1.3 translation errors

t1.4		Mild errors	Severe errors
t1.5	Few errors	Acceptable	Questionable
t1.6	Many errors	Questionable	Objectionable

An example of these factors is the natural, well-known fact that no two languages in the world encode meaning in the same way (see Greenfield, 1997; Nettle & Romaine, 2000). While the translators' job is to ensure that meaning is preserved in their translations, in some cases this is accomplished at the cost of increasing the amount of text. Unlike other forms of text, this is not trivial matter in tests, which students usually need to respond to within certain time limits. Under these circumstances, a substantial increase in the amount of text in an item needed to express the same idea as in its original version may imply more reading time and a potential impact on the time students are left with to make sense of the item.

Another example of aspects beyond the translators' translation skills has to do with the formatting of translated items. Changes in font size and style, and alterations in the proportion of figures included in test items are not due to the translators' actions yet affect the equivalence between the original and the translated versions of an item.

A third example of aspects beyond the translators' translation skills is the extent to which the items reflect the culture of the target language country. While, technically, the translation of an item may not be flawed, the contextual information used in it may not be as familiar to the population tested in the target language as it is to the population tested in the source language.

8.1.2 Inevitability of Translation Error

As a result of the combination of multiple factors like these, strictly speaking, a translation cannot be expected to be perfect. Indeed, our findings from reviews of translated items show that the majority of translated items have TEs—although they are not necessarily fatally flawed (Backhoff et al., 2011; Solano-Flores et al., 2006; Solano-Flores et al., 2009).

8.1.3 Objectionability of Translated Items

To what extent a translated item is objectionable or acceptable depends on the relation between the frequency and severity of TEs. This relationship is represented in Table 8.1. Acceptable translated items have few mild TEs. Questionable translated items have many mild errors or few severe TEs; they may or may not affect student performance depending on the nature of the TEs, the characteristics of the item, and

the characteristics of the linguistic group tested. Objectionable translated items have many and severe TEs; they are very likely to alter the intended meaning of the original item and affect student performance.

8.1.4 Translation Error Dimensions

Our theory postulates the existence of test translation error dimensions (TEDs), grouped in three broad categories, Design, Language, and Content. Each TED comprises several types of TE, as shown in Table 8.2. While it parallels the systems of dimensions and types of TEs used in other investigations (e.g., Backhoff et al., 2011; Solano-Flores et al., 2009), the definitions of TEDs shown in Table 8.2 and the types of TE they comprise were respectively adapted and included with the intent to meet the needs of this particular translation review project.

The TEDs, Style, Format, and Conventions, are grouped in the category, Design. These TEDs have to do with the format, editorial features, and visual layout of translated tests. Convention errors are mainly observed in multiple-choice items. TEs belonging to the category, Design tend to be mild and are unlikely to impact student performance (Note 2).

The TEDs, Grammar, Semantics, and Register, are grouped in the category, Language. These TEDs have to do with the structural and functional aspects of the language used in the translation, the preservation of meaning across languages, and the characteristics of the language usage by the target population in social and instructional contexts.

The TEDs, Information, Construct, Culture, and Origin, are grouped in the category, Content. These TEDs have to do with the ways in which information is presented and how examinees are likely to understand and make sense of items. Unlike TEs belonging to the category Design, TEs belonging to the category, Content tend to alter the structural and functional aspects of language or the ways in which examinees make sense of items. Therefore, they tend to be severe and constitute a threat to the validity of a translated item. The TED, Origin addresses the fact that examining the linguistic equivalence of items allows detection of errors not detected throughout the entire process of test development of the item (Solano-Flores, Trumbull, & Nelson-Barber, 2002). Since Origin errors are not exclusive to the translated version of test items, they are included in the list of TEDs only for conceptual purposes, to allow documentation of any anomalies identified during the process of test translation review.

8.1.5 Translation Error Multidimensionality

The theory postulates that test TE is multidimensional. For example, the inappropriate use of commas in *the panda eats, shoots, and leaves* (when the intended meaning is, *the panda eats shoots and leaves*) (Note 3) is both a punctuation error (Style TED) and an error that affects the meaning of the sentence (Semantics TED).

Table 8.2 Translation error dimensions and types of translation errors (italics) considered in the analysis of translated PISA textual analytical units (TAUs)

Design dimensions
Style: The style used in the translation of the TAU is not used in printed materials in the country.
• *Punctuation* • *spelling* • *wrong use of uppercase letter* • *wrong use of lowercase letter*
Format: The visual layout of the translated TAU is different from the original.
• *Change of size, position, or style of an illustration, table, or graph* • *change of justification, font, or font size of text* • *change of margin width* • *omission of graphic component* • *insertion of graphic component*
Conventions: The translation of the TAU does not reflect item writing conventions used in the country.
• *Inconsistent syntactical structure of stem and options* • *wrong use of punctuation in the item's stem* • *change in order of options* • *inconsistent syntactical structure among options* • *wrong use of uppercase letters in options*
Language dimensions
Grammar: The translation of the TAU violates grammatical rules or uses grammatical structures that are not common in the country.
• *Literal translation* • *unnatural syntax of a sentence* • *subject-verb inconsistency* • *singular-plural inconsistency* • *wrong preposition* • *wrong tense* • *conflation of sentences*
Semantics: The translation of the TAU alters its original meaning.
• *Use of a false cognate* • *wrong translation or adaptation of an idiomatic expression* • *alteration of meaning* • *confusing translation of a sentence* • *multiple possible interpretations of a sentence* • *change of gender of a character* • *conflation of ideas* • *inaccurate terms* • *use of terms with multiple meanings* • *wrong translation of a word*
Register: The translation of the TAU does not reflect the terms, idiomatic expressions, and discursive forms used in the country.
• *Use of words of low frequency in the country* • *wrong translation of a technical term* • *translation of a technical term in a way not used in the country*
Content dimensions
Information: The translation of the TAU alters the amount, precision, or type of information provided.
• *Inconsistent translation of a non-technical term* • *change in number of times a technical term is used* • *insertion of technical term* • *insertion of a sentence or explanation* • *omission of a key word* • *omission of a technical term* • *omission of a sentence or explanations*
Construct: The type of skill or knowledge needed to understand and respond to the TAU is different from the skill or knowledge needed to understand and respond to the TAU in the source language.
• *Possible change of the item's cognitive demands* • *possible alteration of ways in which a task may be interpreted* • *wrong technical term* • *inconsistent translation of a technical term* • *undue insertion of a technical term* • *omission of a technical term* • *translation of a technical term as a non-technical term* • *translation of a non-technical term as a technical term*
Culture: The TAU does not reflect the characteristics of the culture or the curriculum in the target language.
• *Contextual information and situations that are uncommon in the country* • *measurement units not used in the country* • *problem posed not meaningful in the country's culture* • *knowledge assessed not taught in country*
Origin: The TAU carries over errors from the source language version.
• *Inconsistency in the content of the two source languages* • *conceptual errors in the design of the item* • *confusing directions* • *the answer to an item may give the clue for responding to another item within the same assessment unit*

8.1.6 Tension Among Translation Error Dimensions

Finally, the theory postulates that there is a tension between TEDs. Actions intended to avoid TE on a given TED may involve making errors on other TEDs. For example, the grammatical rules of the target language may prevent a noun from being repeated in the same sentence. In some languages, a marker needs to be used to refer to a noun in the rest of the sentence, once the noun appears in it. As a consequence, a key technical term that appears several times in the same sentence in the original version of the item appears only once in its translation. The grammatical rules of the target language need to be followed at the cost of altering the number of times that the key term appears in the sentence—which alters the amount of information provided by the item.

8.2 Methods

8.2.1 Sample of Assessment Units and Analytical Test Units

We examined 61 assessment units (one or several paragraphs with contextual information and one or more items related) from the Mexican, Spanish language version of PISA-2006. Of these 61 assessment units, 37 and 24 were respectively science and mathematics assessment units (Note 4). These 61 assessment units comprised a total of 193 text analytical units (TAUs), defined as either the introductory text or an item within an assessment unit. Of the 193 TAUs examined, 55 were introductory texts and 138 were items. Of these 138 items, 101 and 37 were respectively science and mathematics items.

8.2.2 Test Translation Review and Error Coding Procedures

In addition to the fact that most of the PISA 2006 items consisted of two forms of TAUs (an introductory text or an item), our coding procedure took into account that PISA items use two source languages, English and French (see Grisay et al., 2007).

We assembled a multidisciplinary translation review panel composed of three middle school teachers (Spanish, science, and mathematics); three high school teachers (Spanish, science, mathematics); one English-to-Spanish translator, and one French-to-Spanish translator (both certified by international translation professional organizations); one linguist; one test developer; and one psychometrician (measurement specialist).

The following procedure was used to review each TAU. First, the TAU in the target language (the translated item) was projected on a screen. Reviewers read the TAU and, in the case of items, responded to the item individually as if they were

students taking the test. This was done with the purpose of giving the reviewers the opportunity to become acquainted with the content of the item and to become aware of its cognitive and linguistic demands in the target language.

The reviewers then were asked to examine the TAU and individually record on a coding form all the types of TE they thought could affect the interpretation of the item. The reviewers were instructed to focus on a specific set of dimensions designated according to their professional background. However, they were allowed to record errors on all dimensions (Table 8.3).

Once the reviewers finished recording their comments, the original English and French versions of the TAU were projected on two additional screens. Then the reviewers were asked to compare the English and French versions with the TAU in the target language and to individually code any type of TE according to the list of types of errors listed above for each error dimension. They also wrote their comments on the TAU based on their experience reading and responding to it and on comparing the original and translated versions.

For each TED, the panel discussed each reviewer's coding. Project staff facilitated a discussion to ensure that the panel decided by consensus what errors should be recorded and on which TEDs they should be coded. In the case of items, the panel was asked to decide, based on the number and severity of the TEs, if the translated item should be classified as objectionable (i.e., an item with many and severe TEs which were likely to adversely affect student performance). The review coding decisions were captured on an electronic spreadsheet for further analysis.

8.2.3 Data Analysis

For the purpose of our analysis, we measured TE in each TAU as the number of different translation error dimensions (NDTED) on which TEs were observed in it. This coarse-grain measure has proven to be sensitive to important differences in translation quality among items (see Solano-Flores et al., 2005, 2006).

Also for the purpose of our analysis, we used the p-values of items as a measure of item difficulty. Item p-value was computed as the proportion of the item's highest possible score (see Adams, Berezner, & Jakubowski, 2010), which allowed to have proportional measures of difficulty for both dichotomous and partial-credit items. More specifically, for dichotomous items, difficulty was computed as the proportion of students who responded correctly; for partial-credit items, difficulty was computed as the mean score of the item divided by its maximum score.

To examine the impact of TE on student performance, we examined the Pearson correlations between NDTED and item p-value for different sets of TEDs, different content areas (science and mathematics), and items that were and were not identified as objectionable by the translation review panel. Impact on performance should be observed as a negative correlation.

Given the complex interaction of the students' knowledge of the content being assessed and the cognitive and linguistic demands of test items, it would be naive to

Table 8.3 Expertise provided and assigned focus on translation error dimensions by specialist: 1 = main role; 0 = adjuvant role

Expertise and Contribution	Style	Format	Conventions	Information	Grammar	Semantics	Construct	Register	Culture	Origin
Teacher: Formal and informal use of language in the school; consistency of TAUs with language used in the curriculum; content accuracy.	0	1	1	1	0	0	0	1	1	0
Translator: Language equivalence across languages; formal use of language in the translated TAUs.	1	1	0	1	1	1	0	0	0	0
Linguist: Formal (structural) and functional (sociolinguistic, cultural) aspects of the translated TAUs; use of language in context.	1	0	0	0	1	1	1	0	1	0
Test developer: Item writing; equivalence of TAUs; item complexity across languages; wording.	0	1	1	1	0	0	1	0	0	0
Psychometrician: Potential impact of language and design features on the interpretation of TAUs.	0	0	1	1	0	1	1	0	0	0

t4.1 **Table 8.4** Percentage of text
t4.2 analytical units (n = 193) with
t4.3 at least one error on each of
t4.4 the translation error
t4.5 dimensions

Dimension	Percent	
Style	48	t4.7
Format	53	t4.8
Conventions	3	t4.9
Information	53	t4.10
Grammar	53	t4.11
Semantics	78	t4.12
Construct	35	t4.13
Register	21	t4.14
Culture	5	t4.15
Origin	41	t4.16

252 expect to observe impressively high and statistically significant correlations. Rather,
253 we expected to observe patterns in those correlations that would indicate a systematic
254 impact of TE on student performance, especially for language- and content-related
255 TEDs and for items identified as objectionable by the translation review panel.

256 8.3 Results

257 *8.3.1 Frequency and Severity of Translation Errors*

258 We observed TEs on at least one dimension for almost all (96%) of the TAUs. Of the 138
259 TAUs which consisted of items, 26 were identified by the committee as objectionable.
260 Table 8.4 shows the percentage of TAUs identified as having at least one error on
261 each of the TEDs. As indicated above, many of these errors are not likely to bias test
262 results and many are even difficult to be noticed by individuals who have no experience
263 reviewing test translations. On the other hand, there are TEs that may potentially
264 threaten the validity of test items. Such is the case for errors on the Semantics,
265 Grammar, and Information dimensions, which were observed respectively in 78%,
266 53% and 53% of the TAUs.
267 On average, a TAU had errors on 3.9 different dimensions (s.d. = 1.834). As
268 Fig. 8.1 shows, the number of different dimensions in which error was observed had
269 a normal frequency distribution.

270 *8.3.2 Translation Error and Item Difficulty*

271 As Table 8.5 shows, Pearson correlation coefficients of −.059 and −.117 between
272 NDTED and p-value were observed respectively when all dimensions were consid-
273 ered and when the three language dimensions (Grammar, Semantics, and Register)
274 and three of the four content dimensions (Information, Construct, and Culture)

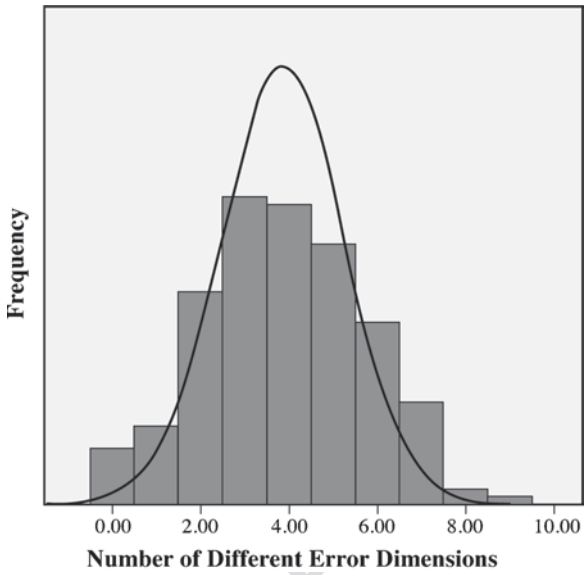

Fig. 8.1 Frequency distribution of text analytical units by number of different error dimensions

Table 8.5 Correlation between number of different dimensions on which error was observed and item p-value by set of dimensions considered, content area, and objectionability

Comparison	Correlation
By set of dimensions (138)	
All dimensions	−.059
Language and content dimensions[a]	−.117
By content area[a] (138)	
Science (n = 101)	−.115
Mathematics (n = 37)	−.213
By objectionability (language and content dimensions)[a](138)	
Non-objectionable items (n = 112)	−.084
Objectionable items (n = 26)	−.404**

Sample and subsample sizes in parentheses

[a] Includes the three language dimensions (Grammar, Semantics, and Register) and three of the four content dimensions (Information, Construct, and Culture)

**Significant at p = .01 (2-tailed)

were considered. (As mentioned above, since Origin errors are common to both the source and language versions of items, they were not included in the analyses). This difference supports findings from previous test translation reviews that design dimensions (Style, Format, and Conventions) are unlikely to affect student performance whereas language and content dimensions tend to have a greater impact on student performance and may potentially threaten the validity of translated items.

Correlation coefficients of −.115 and −.213 between NDTED and p-value were observed respectively for the science and mathematics items. These results are consistent with findings from other translation reviews, in which we (e.g., Solano-Flores, Backhoff, & Contreras-Niño, 2005) have observed higher correlations between NDTED and item difficulty for mathematics than science items.

Correlation coefficients of −.084 and −.404 (significant) were observed respectively for acceptable and objectionable items. This considerable difference indicates that the review procedure allows identification of items which have sets of errors that are likely to seriously impact student performance. This finding is important, considering that the number of items identified as objectionable (26) constitute about 19% of the 138 items examined.

8.4 Summary and Conclusions

The theory of test translation error (TTTE; Solano-Flores et al., 2009) postulates the existence of translation error dimensions (TEDs; e.g., Semantics, Construct, Grammar) and views translation error (TE) as multidimensional (a translation error can belong to several TEDs). It also postulates that a tension exists between TEDs (i.e., in translating a test item, avoiding error on one dimension may produce error on other dimensions). Accordingly, error-free test translation is impossible; effective test translation minimizes but does not eliminate error. The theory also postulates that while items usually have multiple TEs, most of them are mild and even unnoticeable. Objectionable translated items have many and severe TEs and are likely to pose serious linguistic challenges to examinees who are given the translated version of a test.

In this chapter, we report the results of our review of the Spanish language Mexican version of PISA-2006 science and mathematics text analytical units (TAUs). Consistent with results from our review of the Spanish Mexican translation of TIMSS-1995 (Solano-Flores et al., 2005) and the Spanish Mexican translation of PISA-2003 (Backhoff et al., 2011), our results show that translation reviews based on the TTTE are highly sensitive to TE.

The results also confirm previous findings that student performance tends to be resilient to TE on design-related TEDs and sensitive to TE on language- and content-related TEDs. Also, items whose translation was identified as objectionable by the review panel correlated higher with item difficulty than items whose translation was not identified as objectionable—a finding that speaks to the sensitivity of TTTE-based judgmental review procedures.

A limitation of our analyses of correlations of measures of TE and item difficulty stems from the fact that we did not account for the effect of TE observed in the introductory text of assessment units. Future research should explore models for examining this relationship.

Unlike other approaches created to examine translation quality, the TTE focuses on disconfirming (rather than confirming) evidence that the translation of test items

is correct. In addition, because they use multidisciplinary review panels which discuss the linguistic features of the items at length, TTTE-based coding procedures are sensitive to TE with a level of precision and detail not attained with conventional approaches.

Experienced test translators who have attended our workshops on the use of the TTTE and the methods described in this chapter (e.g., Backhoff, Solano-Flores, & Contreras-Niño, 2010; Solano-Flores et al., 2010) react initially with skepticism when we report our findings. They find it difficult to believe that items translated according to available translation guidelines have multiple TEs. It is not until they observe the discussions of the review panels examining specific translated items that they appreciate the level of sensitivity of the theory and our coding procedures to the nuances of language in translated items.

As with measurement error, TE cannot be entirely eliminated, but it can be minimized. As our results show, a theoretical perspective that assumes error inevitability in test translation is more sensitive to the complexities of language in translated PISA items and can contribute to the improvement of future PISA translation procedures. We hope that, in the future, PISA participating countries use our approach as a tool for operationalizing PISA translation procedures and formatively evaluating their own translation work.

[AU4]

Author's Note

Portions of this paper are originally from a paper presented at the PISA Research Conference, 14–16 September 2009, Kiel, Germany. The investigation reported in this paper was commissioned and funded by the National Institute for Educational Evaluation (INEE), Mexico, and conducted through a contract with the Autonomous University of Baja California (UABC), Mexico. The opinions expressed are not necessarily those of the funding agency. Contact author: Guillermo Solano-Flores, guillermo.solano@colorado.edu.

Notes

Note 1. While *translation error* (in singular) is used here to refer to lack of equivalence between the original language version and the translated version of an item, *translation errors* (in plural) or *a translation error* are used to refer to specific instances or types of translation error (e.g., the inaccurate translation of a term or an inappropriate use of punctuation).

Note 2. Of course, there are exceptions. For example, an alteration in the proportion of the length of the axes in a graph showing a functional relationship may make the line of the function look steeper in the translated item than in the original—which may affect how the examinee interprets the function.

Note 3. The example is based the story told by Lynne Truss (2004) at the beginning of her well-known book on punctuation, *Eats, shoots, and leaves*.

Note 4. One of the science assessment units and 17 of the mathematics assessment units consisted of a stand-alone item with no introductory text.

References

Adams, R., Berezner, A., & Jakubowski, M. (2010). *Analysis of PISA 2006 preferred items ranking using the percent-correct method* (OECD Education Working Papers, No. 46). OECD Publishing. Retrieved June 7, 2011, http://dx.doi.org/10.1787/5km4psmntkq5-en

Allalouf, A. (2003). Revising translated differential item functioning items as a tool for improving cross-lingual assessment. *Applied Measurement in Education, 16*(1), 55–73.

Backhoff, E., Contreras-Niño, L. A., & Solano-Flores, G. (2011). *The theory of test translation error and the TIMSS and PISA international test comparisons*. Mexico: National Institute for Educational Evaluation [Sp.].

Backhoff, E., Solano-Flores, G., & Contreras-Niño, L. A. (2010, February 18–19). *Analysis of the Mexican Spanish language translation of PISA-2006*. Presentation at the Ibero-American Seminar on the theory of test translation error in international comparisons. National Ministry of Education and National Institute for Educational Evaluation, Mexico City, Mexico.

Bybee, R., McCrae, B., & Laurie, R. (2009). PISA 2006: An assessment of scientific literacy. *Journal of Research in Science Teaching, 46*(8), 865–883.

Church, B. (1991). An examination of the effect that commitment to a hypothesis has on auditors' evaluations of confirming and disconfirming evidence. *Contemporary Accounting Research, 7*(2), 513–534.

Creswell, J. W., & Miller, D. L. (2000). Determining validity in qualitative inquiry. *Theory into Practice, 39*(3), 124–130.

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

Ercikan, K., Gierl, M. J., McCreith, T., Puham, G., & Koh, K. (2004). Comparability of bilingual versions of assessments: Sources of incomparability of English and French versions of Canada's national achievement tests. *Applied Measurement in Education, 17*(3), 301–321.

Gierl, M. J., Rogers, W. T., & Klinger, D. (1999, April). *Using statistical and judgmental reviews to identify and interpret translation DIF*. Paper presented at the annual meeting of the National Council on Measurement in Education, Montreal, QC.

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. (2003). Translation procedures in OECD/PISA 2000 international assessment. *Language Testing, 20*(2), 225–240.

Grisay, A. (2007). *The challenge of adapting PISA materials into non Indo-European languages: Some evidence from a brief exploration of language issues in Chinese and Arabic*. OECD Core A Consortium.

Grisay, A., de Jong, J. H., Gebhardt, E., Berezner, A., & Halleux-Monseur, B. (2007). Translation equivalence across PISA countries. *Journal of Applied Measurement, 8*(3), 249–266.

Halleux-Monseur, B. (2008). *Translation, adaptation and verification of test material in OECD international surveys*. Paris: Directorate for Education, Institutional Management in Higher Education Governing Board, OECD.

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

Hambleton, R. K., Merenda, P. F., & Spielberger, C. D. (Eds.). (2005). *Adapting educational and psychological tests for cross-cultural assessment*. Mahwah, NJ: Lawrence Erlbaum Associates, Publishers.

Harkness, J., van de Vijver, F. J. R., & Mohler, P. (Eds.). (2003). *Cross-cultural survey methods*. Hoboken, NJ: Wiley.

Mullis, I. V. S., Kelly, D. L., & Haley, K. (1996). Translation Verification Procedures. In M. O. Martin & I. V. S. Mullis (Eds.), *Third international mathematics and science study: Quality assurance in data collection*. Chestnut Hill, MA: Boston College.

Nettle, D., & Romaine, S. (2000). *Vanishing voice: The extinction of the world's languages*. New York: Oxford University Press.

Solano-Flores, G., Backhoff, E., & Contreras-Niño, L. A. (2006). *Methodology for evaluating the quality of test translations in international test comparisons: The case of Mexico, TIMSS-1995. [Sp.]*. Mexico: National Institute for Educational Evaluation (INEE).

Solano-Flores, G., Backhoff, E., & Contreras-Niño, L. A. (2009). Theory of test translation error. *International Journal of Testing, 9*, 78–91.

Solano-Flores, G., Backhoff, E., & Contreras-Niño, L. A. (2010, February 18–19). *Test translation review sessions: A demonstration*. Presentation at the Ibero-American Seminar on the theory of test translation error in international comparisons. National Ministry of Education and National Institute for Educational Evaluation, Mexico City, Mexico.

Solano-Flores, G., Contreras-Niño, L. A., & Backhoff, E. (2005, April 12–14). *The Mexican translation of TIMSS-95: Test translation lessons from a post-mortem study*. Paper presented at the 2005 annual meeting of the National Council on Measurement in Education. Montreal, Quebec, Canada.

Solano-Flores, G., Contreras-Niño, L. A., & Backhoff, E. (2006). Test translation and adaptation: Lessons learned and recommendations for countries participating in TIMSS, PISA, and other international comparisons. *REDIE: Electronic Journal of Educational Research, 8*(2). [Sp.] http://redie.uabc.mx/vol8no2/contents-solano2.html

[AU5] Solano-Flores, G., & Gustafson, M. (In Press). Assessment of English language learners: A critical, probabilistic, systemic view. In M. Simon, K. Ercikan, & M. Rousseau (Eds.), *Improving large scale assessment in education: Theory, issues, and practice*. Taylor & Francis, Routledge.

Solano-Flores, G., Trumbull, E., & Kwon, M. (2003, April 21–25). *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.

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.

Truss, L. (2004). *Eats, shoots & leaves*. New York: Gotham Books.

van de Vijver, F. J. R., & Poortinga, Y. H. (2005). Conceptual and methodological issues in adapting tests. In R. K. Hambleton, P. F. Merenda, & C. D. Spielberger (Eds.), *Adapting educational and psychological tests for cross-cultural assessment*. Mahwah, NJ: Lawrence Erlbaum Associates, Publishers.

Please see the Author's responses in the Word document sent along with the proof

Author Queries

Chapter No.: 8 0001528456

Queries	Details Required	Author's Response
AU1	Please provide keywords	
AU2	Please confirm the corresponding author.	
AU3	Please provide the name of the department in the affiliations.	
AU4	Please confirm the placement of Author's Note.	
AU5	Please update Reference Solano–Flores and Gustafson (in press).	