La Teoría del Error de Traducción de Pruebas y las evaluaciones internacionales de TIMSS y PISA

Reportes de Investigación

No.36

Instituto Nacional para la Evaluación de la Educación

Resultados de Investigación

La Teoría del Error de Traducción de Pruebas y las evaluaciones internacionales de TIMSS y PISA

Primera edición 2011
ISBN: 978-607-7675-27-3

INSTITUTO NACIONAL PARA LA EVALUACIÓN DE LA EDUCACIÓN
José Ma. Velasco 101, 5° piso, Col. San José Insurgentes,
Delegación Benito Juárez, México, 03900, D.F.

Coordinación Editorial
Subdirectora de Difusión
María Norma Orduña Chávez

Corrección de estilo
César Rebolledo González
Norma Alfaro Aguilar
María Esther Saldívar Chávez

Diseño gráfico y Formación
Javier Franco Vázquez

Elaboración y diseño de gráficas y figuras
Pablo Josué Pulido Ramírez

Impreso en México
Distribución Gratuita

Junta Directiva

Presidente: Mtro. Alonso Lujambio Irazábal - Secretario de Educación Pública
Presidente Suplente: Lic. Francisco Ciscomani Freaner - Titular de la Unidad de Planeación y Evaluación de Políticas Educativas, SEP

Consejeros

Ing. José Enrique Villa Rivera - Director General del CONACYT
Mtro. Fernando González Sánchez - Subsecretario de Educación Básica, SEP
Lic. Carlos Montaño Fernández - Subsecretario de Egresos, SHCP
Dr. Enrique Cabrero Mendoza - Director General del CIDE, A.C.
Dr. José Pablo René Asomoza y Palacio - Director General del CINVESTAV, IPN
Dra. Sylvia Ortega Salazar - Rectora de la UPN
Dr. Efrén Parada Arias - Director General del Instituto Mexicano del Petróleo
Dr. Emilio Zebadúa González - Presidente de la Fundación para la Cultura del Maestro, A.C.
C. Eduardo Bohórquez López - Director General de Transparencia Mexicana
Profr. Juan Díaz de la Torre - Secretario General del Comité Ejecutivo del SNTE
Sr. Ramón Leopoldo García López - Presidente de la Federación Nacional de Asociaciones de Padres de Familia, A.C.
Sra. Consuelo Mendoza García - Presidenta de la Unión Nacional de Padres de Familia, A.C.
Lic. Gerardo Gutiérrez Candiani - Presidente de la Comisión de Educación del Sector Empresarial
C. Manuel Ulloa Herrero - Presidente de Observatorio Ciudadano de la Educación, A.C.
Secretario de la Junta Directiva - Lic. Raúl Sosa Bustamante
Coordinación de Órganos Desconcentrados y del Sector Paraestatal, SEP
Prosecretaria de la Junta Directiva - Dra. Annette Santos del Real
Directora General Adjunta, INEE
Comisario Público - Ing. Rafael Muñoz de Cote Sisniega
Comisario Público Propietario del Sector Educación y Cultura, SFP

Cuerpo Directivo

Dra. Margarita Zorrilla Fierro - Directora General
Dra. Annette Santos del Real - Directora General Adjunta
Dr. Héctor Virgilio Robles Vásquez - Director de Indicadores Educativos
Mtro. Andrés Eduardo Sánchez Moguel - Director de Pruebas y Medición
Mtra. Gabriela Asunción Barba Martínez - Directora de Evaluación de Escuelas
Lic. María Antonieta Díaz Gutiérrez - Directora de Proyectos Internacionales y Especiales
Mtra. Magdalena del Pilar González Martínez - Directora de Relaciones Nacionales y Logística
Ing. Enzo Molino Ravetto - Director de Informática
Lic. Javier de Jesús Noyola del Río - Director de Administración y Finanzas
Lic. César Javier Gómez Treviño - Director de Asuntos Jurídicos
Lic. Jorge Pedro Velasco Oliva - Titular del Órgano Interno de Control

Índice

Introducción

La globalización económica impone cada vez más la necesidad de adaptar pruebas y traducirlas a idiomas diferentes (Hambleton, 1994). Esta exigencia surge, entre otras razones, de los requerimientos planteados por estudios que comparan sistemas educativos de distintos países.

En este contexto es primordial la noción de validez de un instrumento evaluativo: el puntaje de una prueba no debe estar determinado por factores ajenos al constructo que el instrumento pretende medir (Messick, 1989). Este constructo puede referirse a conocimientos, competencias, habilidades, destrezas, actitudes o valores que se deseen evaluar en una población, como lo son estudiantes de una cierta edad o de un cierto grado escolar.

En el caso de la prueba PISA (*Programa Internacional de Evaluación de Estudiantes*, por sus siglas en inglés), los constructos por evaluar se refieren a lo que la OECD (*Organización para la Cooperación y el Desarrollo Económicos*, por sus siglas en inglés) denomina "competencias para la vida". En el caso de la prueba TIMSS (*Estudio de las Tendencias en Matemáticas y Ciencias*, por sus siglas en inglés), los constructos que se miden se refieren a habilidades y conocimientos disciplinarios comunes en los currículos de los diversos países participantes.

Preservar la validez de una prueba es un serio reto cuando se le traduce a otros idiomas, aún cuando el constructo a medir no involucra habilidades verbales o competencias de lectoescritura (Ercikan, 1998; Gierl, Rogers, y Klingner, 1999). Una traducción inadecuada puede ser una fuente de sesgo de método que afecta a los puntajes obtenidos por los estudiantes —a quienes se aplica la prueba traducida— y atenta, además, contra la validez de las interpretaciones derivadas de sus resultados. Ello se debe principalmente a las propiedades lingüísticas propias a las pruebas que las distinguen de otras formas de texto, tales como el uso de un lenguaje sintético; el uso, en reactivos[1] de opción múltiple, de una base gramaticalmente incompleta y las opciones de respuesta que la completan; así como una alta frecuencia de términos técnicos (Davis, 1991; Ferguson y Fairburn, 1985; Solano-Flores, 2006; Solano-Flores y Kidron, 2006).

Debido a estas características, la traducción de ítems de pruebas puede plantear retos más difíciles de resolver que otros tipos de texto. Desafortunadamente, a pesar de estas dificultades, en la práctica profesional actual, el tiempo y los recursos asignados para traducir una prueba y revisar su traducción es mucho menor al tiempo y los recursos asignados para desarrollarla en la lengua original (Valdés y Figueroa, 1994; Solano-Flores, Trumbull y Nelson-Barber, 2002).

[1] Los términos ítem y reactivo se utilizarán indistintamente a lo largo de este documento para hacer referencia al componente de una prueba que evalúa un constructo o un conjunto específico de constructos, y que puede considerarse como una unidad analítica de evaluación de conocimientos, habilidades o competencias. Estos términos son preferibles al término pregunta, ya que no todos los ítems de pruebas están planteados de manera interrogativa y pueden ser más complejos que una oración en forma de pregunta.

Los lineamientos para la traducción de pruebas internacionales elaborados por Hambleton (1994) constituyen un intento de importancia histórica para garantizar la equivalencia de versiones de una prueba en distintos idiomas, como TIMSS y PISA. Otras acciones incluyen el uso de juicios de revisores de las traducciones (Hambleton, 1994), el análisis del funcionamiento diferencial de ítems (Muñiz, Hambleton, y Xing, 2001) y el refinamiento de la traducción basado en administrar los ítems cuya traducción ha sido revisada en muestras de estudiantes en el lenguaje objetivo (Gierl, Rogers, y Klingner, 1999; Muñiz, Hambleton, y Xing, 2001).

Con el fin de garantizar tal equivalencia y la propia implementación de los lineamientos de traducción, se ha desarrollado una gran variedad de métodos y estrategias. Entre éstos destacan: el uso de traductores múltiples independientes y una tercera persona que integra las traducciones (Grisay, 2002); el uso de un sistema de revisión y certificación de traducciones antes de la aplicación de las pruebas (O'Connor y Malak, 2000); el reconocimiento de las limitaciones del procedimiento de traducción inversa (Grisay, 2003); y la utilización de más de un idioma fuente con el fin preservar el significado en la traducción (Harkness, 2003).

Por limitaciones económicas y de tiempo, no siempre es fácil que los países participantes en pruebas internacionales apliquen estos procedimientos de traducción recomendados por los organismos internacionales (Maxwell, 1996). Por ello, es de crucial importancia asegurar que los traductores sean certificados y altamente calificados. Entre otras características, los traductores deben, idealmente: tener un buen conocimiento del idioma fuente, ser hablantes nativos del idioma objetivo, tener experiencia en ambos idiomas y culturas, con estudiantes de las poblaciones objetivo, así como en el proceso de desarrollo de pruebas.

Aunque son necesarios los lineamientos para traducir, adaptar y revisar pruebas traducidas, estos documentos son prescriptivos, no analíticos. Tales documentos tienen limitaciones para guiar los tipos de razonamiento utilizados por los traductores o revisores de la traducción de pruebas. Por ello, no son una garantía de una traducción de alta calidad. En consecuencia, se requiere un marco conceptual y metodológico para evaluar la calidad de la traducción de pruebas y la aportación de evidencia empírica sobre el efecto de la traducción en las propiedades técnicas de sus reactivos, por ejemplo, su dificultad.

Implicación para Iberoamérica

Ante la necesidad de que los países iberoamericanos cuenten con un marco conceptual y metodológico para analizar y corregir posibles errores de traducción en los estudios internacionales de logro educativo, los autores de este cuaderno trabajan desde 2003 en el desarrollo de la Teoría del Error de Traducción de Pruebas (TETP), cuyos supuestos, metodología y resultados se han publicado en distintos medios (véase, Solano-Flores y Backhoff, 2003; Solano-Flores, Contreras-Niño y Backhoff, 2005; Solano-Flores, Backhoff y Contreras-Niño, 2009).

En este documento se presentan los principios conceptuales, metodológicos y prácticos de la TETP. Dicha teoría, así como los estudios empíricos que le dan apoyo, se han desarrollado principalmente a partir de estudios solicitados por el Instituto Nacional para la Evaluación de la Educación (INEE), con el fin de conocer la calidad de las traducciones mexicanas al español de dos pruebas internacionales de logro educativo. El primer estudio, realizado en 2003, analizó reactivos de la prueba de TIMSS (desarrollada por la Asociación Internacional para la Evaluación Educativa, IEA) utilizada

en la comparación internacional de 1995 y en un estudio nacional de 2000. El segundo trabajo se concluyó en 2009 y tuvo como propósito analizar la traducción de los reactivos de ciencias naturales de la prueba PISA-2006. Adicionalmente, se han revisado varios reactivos liberados de PISA-2003 y TIMSS-2003.

A medida que hemos acumulado experiencia revisando traducciones de pruebas, hemos refinado los procedimientos para implementar la TETP. Este cuaderno de investigación describe la versión actual de dichos procedimientos. En el primer apartado, se hace una breve revisión bibliográfica de los errores que se cometen inadvertidamente en el proceso de traducción de pruebas internacionales. En el segundo, se presentan los principios de la TETP y se dan ejemplos concretos de algunas dimensiones y categorías de error. En el tercero, se describe la metodología utilizada por la TETP para identificar posibles errores de traducción de las pruebas de logro, misma que enfatiza el uso de comités multidisciplinarios de revisión. En el cuarto, se muestran los resultados que arroja la revisión de errores de traducción de las pruebas TIMSS–1995 y PISA–2006. En el quinto y último apartado, se discute la utilidad de la TETP y se hacen algunas recomendaciones para que los países que participan en los estudios internacionales minimicen el error de traducción y con ello contribuyan a garantizar la validez de los resultados de dichos estudios. Este cuaderno se complementa con una sección de referencias bibliográficas y otra de anexos.

Antecedentes: el problema de la traducción de pruebas

En la comunidad de los profesionales de la medición educativa, es bien conocido que cuando un instrumento se traduce, es muy probable que se alteren los constructos en medición (Hambleton, 2005). Ello se debe a una falta de correspondencia perfecta entre las lenguas, pues codifican de manera diferente las ideas. A nivel léxico básico, por ejemplo, un concepto tan sencillo como "rectángulo" puede significar diferentes cosas en culturas y lenguas. En países de habla inglesa, *rectangle* es un tipo de figura geométrica que incluye al cuadrado y a cualquier otra figura geométrica con cuatro ángulos rectos. La traducción de *rectangle* como "rectángulo" puede no ser adecuada cuando un ítem involucra la distinción de un cuadrado como un caso especial de figura con cuatro lados de la misma longitud. Esta sutil diferencia puede no ser del conocimiento de quien traduce una prueba e incluso es difícil que la detecten pedagogos o maestros bilingües de matemáticas. Además de representar un problema de traducción de palabras, este ejemplo muestra que la palabra rectángulo implica distintas formas de organización del conocimiento. En un individuo de habla inglesa, *rectangle* evoca un conjunto de tipos de figuras geométricas del cual el cuadrado es una subcategoría; sin embargo, en un individuo de habla española, rectángulo evoca un conjunto de figuras geométricas del cual el cuadrado no forma parte (véase Solano-Flores, en prensa).

Tradicionalmente, se han empleado dos tipos de procedimientos básicos para controlar el efecto de la traducción en la calidad de una prueba. El primer tipo, al que se le puede llamar empírico, se basa en el análisis del funcionamiento diferencial de los ítems. Supóngase que una prueba ha sido traducida del inglés al español. Para examinar si un ítem específico no afecta de manera diferente a la población evaluada en inglés y a la población evaluada en español, se selecciona una muestra de estudiantes de cada población cuyo desempeño en la prueba total sea comparable. Si, a pesar de ser comparables en su desempeño en la prueba total, las dos muestras poblacionales difieren en su desempeño en este ítem, se puede suponer que la traducción ha producido un sesgo que favorece a una de las dos poblaciones.

Idealmente, antes de usar una prueba traducida, se debiera examinar el funcionamiento diferencial de cada uno de los ítems que lo constituyen y descartar o modificar aquéllos que tienen un sesgo considerable. Este procedimiento es costoso y consume tiempo, pues implica usar muestras poblacionales y analizar los datos antes de administrar la prueba en su forma final. Otra posibilidad consiste en identificar las razones específicas que causan el sesgo de un reactivo; por ejemplo, con base en la opinión de los estudiantes sobre sus interpretaciones de los ítems. En algunas ocasiones, una simple palabra puede producir en los estudiantes interpretaciones del reactivo en su totalidad muy diferentes de la interpretación que los autores tenían en mente cuando lo diseñaron. Como en el procedimiento anterior, éste es costoso y lento, y es improbable que se le utilice sistemáticamente debido a que los países participantes en comparaciones internacionales trabajan con calendarios de actividades muy restringidos que impiden el pilotaje de los ítems traducidos.

El segundo tipo de procedimiento empleado para asegurar que la calidad de una prueba no sea afectada negativamente por la traducción se basa en el juicio de expertos, quienes deciden, de acuerdo con su experiencia y sus conocimientos, si la traducción ha alterado el constructo medido. Un enfoque muy conocido, pero ahora desacreditado, es el de la traducción inversa. Supóngase que una prueba ha sido traducida del inglés al español. Para examinar si el constructo que mide uno de sus ítems es el mismo en las dos versiones, la traducción española se traduce *de regreso* al inglés. Entonces, dos o más personas examinan qué tanto difiere la versión *de regreso* en inglés de la versión original y efectúan modificaciones en la traducción en español cuando encuentran diferencias importantes. Aunque por muchos años se le aceptó como evidencia de rigor metodológico en la traducción de pruebas, ahora se sabe que este enfoque puede no llegar a detectar importantes alteraciones del sentido del texto de los ítems en la lengua original, pues no considera el contexto y el ambiente (Bulmer, 1998:161), y falla al identificar problemas textuales formales de sintaxis, morfología, ortografía o consistencia; además de resultar un método tardado y costoso (Maxizip s/f). Por estas y otras razones que cuestionan su validez como un paso efectivo para el control de la calidad de las traducciones (Bullinger, 2003 & Ozolins, 2008), los actuales estándares de traducción en Estados Unidos y Europa ya no incluyen a la traducción inversa como parte de su repertorio para el control del proceso de calidad (Ozolins, 2008; ASTM International, 2006; Halman, 2001).

Principios teóricos de la TETP

Como ya se comentó, la TETP se desarrolló con base en: 1) la revisión empírica de las traducciones de las versiones mexicanas de TIMMS-1995, TIMSS-2000, PISA-2003 y PISA-2006; 2) diversos documentos normativos sobre traducción de pruebas; 3) algunos elementos de la teoría sociolingüística; y 4) un enfoque multidisciplinario. Los propósitos centrales de la TETP son ayudar a operacionalizar las guías para la traducción y adaptación de pruebas, hacer posible el estudio y medición de las mismas y contribuir a sistematizar actividades relacionadas con su proceso de traducción, tales como la selección, entrenamiento, evaluación y certificación de traductores de pruebas.

A diferencia de la mayoría de los enfoques en el campo de la medición educativa que abordan asuntos lingüísticos, la TETP usa un enfoque probabilístico, no determinístico y tiene las siguientes características:

1. Considera al error de traducción de pruebas como inevitable, principalmente debido a que los lenguajes codifican el significado de manera diferente. Ello hace que la equivalencia perfecta de ítems entre idiomas sea casi imposible de lograr.

2. Incluye como fuentes de error de traducción a eventos que suelen tener lugar antes, durante y después del proceso de traducción de la prueba. Por ejemplo, antes de que se realice la traducción, la prueba puede contener errores conceptuales de origen; durante la traducción, pueden ocurrir alteraciones del constructo de un ítem; y después de la traducción, se puede presentar una producción gráfica incorrecta o una impresión defectuosa de la prueba.

3. Establece que los errores de traducción pueden ser clasificados en dimensiones y categorías. Hasta el momento se han identificado las siguientes diez dimensiones: Estilo, Formato, Convenciones, Gramática, Semántica, Registro, Información, Constructo, Cultura y Origen. Actualmente, estas dimensiones se han subdividido en 53 categorías de error para hacerlas más informativas y significativas. Sin embargo, tanto las dimensiones como las categorías de error de traducción no son universales y, en la práctica, pueden surgir nuevas dimensiones y categorías dependiendo de los propósitos, necesidades y recursos disponibles en cada proyecto específico de traducción de pruebas.

4. Concibe al error de traducción de pruebas como multidimensional. El mismo error puede afectar diferentes aspectos de un ítem. Por ejemplo, la inserción inadecuada de una coma puede violar algunas convenciones gramaticales y también puede cambiar el significado pretendido de una idea.

5. Permite la clasificación de errores de traducción en múltiples dimensiones de error. Por ejemplo, una traducción literal que altera el sentido de una oración y la forma en que se le puede entender puede ser clasificado en tres dimensiones: Gramática, Semántica y Constructo.

6. Identifica la existencia de una tensión entre las dimensiones de error de traducción. Muchas acciones de traducción favorecen a unas dimensiones, a costa de vulnerar a otras. Por ejemplo, preservar el significado al traducir un cierto ítem puede requerir alterar la estructura discursiva original.

7. Establece que, debido a la tensión entre dimensiones, el error de traducción de pruebas puede ser minimizado pero no eliminado. Preservar exactamente el mismo conjunto de demandas lingüísticas, de contenido y cognitivas en un ítem traducido es imposible. Una traducción de alta calidad minimiza el error de traducción al reducir la tensión entre las dimensiones de error involucradas, y se basa en ponderar el grado en que la traducción de un ítem es aceptable u objetable.

8. Establece que la severidad de un error de traducción no es absoluta. Está moldeada por factores contextuales y por las características de la población destinataria. Por ejemplo, la inserción de una coma puede ser un error gramatical severo o leve, y a la vez uno semántico severo o leve. Todo depende de la información que proporciona el ítem y la habilidad de los estudiantes para comprender lo que solicita.

9. Establece, debido a lo anterior, que no todo error de traducción necesariamente es fatal. Hasta cierto grado, los examinados pueden identificar y corregir cognitivamente algunos de éstos. Sin embargo, el efecto combinado de errores hace que un ítem imponga retos lingüísticos más altos.

10. La traducción de un ítem es aceptable u objetable en función de la frecuencia y severidad de los errores. Reducir la tensión entre las dimensiones de error de traducción requiere ponderar el grado en que la traducción de un ítem es aceptable u objetable.

11. La probabilidad de que un ítem traducido sea objetable (muy problemático) está determinada por el efecto combinado del número de errores de traducción y su severidad. Los errores de estilo, formato y convenciones tienden a ser leves; los de constructo, semántica, información operativa y registro tienden a ser más graves.

12. La medida en que la traducción de un ítem es aceptable u objetable, puede ser representada como un espacio probabilístico delimitado por la frecuencia y la severidad de los errores de traducción (ver figura 1).

Dimensiones y categorías de errores de traducción

De los supuestos antes mencionados sobresalen dos que por su importancia describiremos con mayor detalle. El primero se refiere a la clasificación del error en términos de dimensiones y categorías. El segundo se relaciona con el espacio probabilístico del error de traducción.

Como ya se mencionó, hasta el momento se han identificado diez dimensiones y 53 categorías de error de traducción, mismas que se presentan en la tabla 1. A continuación se discuten las diez dimensiones y se presentan ejemplos de algunas de sus categorías. Estas dimensiones de error no debieran ser interpretadas como exhaustivas o universales. Los ejemplos presentados ilustran errores de traducción con referencia al *curriculum* mexicano y al uso del idioma en México.

Dependiendo de la naturaleza de las pruebas traducidas, cada dimensión puede tener una relevancia mayor o menor, o se le debe especificar con mayor o menor detalle de acuerdo con el contenido de la prueba, lo que ésta pretende medir, así como los recursos disponibles para el proyecto de traducción. Por ejemplo, la correspondencia de la traducción de términos técnicos con el *curriculum* implementado tiene especial importancia en las pruebas TIMSS. En cambio, en las pruebas PISA tiene especial importancia la correspondencia de la traducción con el uso del idioma en el país. Estas diferencias resultan de los distintos énfasis de las pruebas. TIMSS examina conocimientos y habilidades escolares aprendidos como resultado de la instrucción formal, mientras que PISA examina lo que entiende por "habilidades para la vida".

Dimensión 1: *Estilo.* El estilo en el que está escrito el ítem en el idioma destinatario no es consistente con el estilo empleado en libros de texto y materiales impresos en el país. Por ejemplo: errores de puntuación, uso impropio de mayúsculas o minúsculas, e inconsistencias sujeto–verbo.

Tabla 1. Dimensiones y categorías de error de traducción de acuerdo con la TETP

Dimensión	Categoría
1. Estilo	DE1: Errores de puntuación
	DE2: Uso impropio de mayúsculas
	DE3: Uso impropio de minúsculas
	DE4: Errores de ortografía
2. Formato	DF1: Cambio en el tamaño, estilo o posición de tablas, gráficas o ilustraciones
	DF2: Cambio en el estilo, justificación o tamaño de fuentes o caracteres
	DF3: Márgenes más amplios o más reducidos
	DF4: Omisión de componentes gráficos
	DF5: Inserción de componentes gráficos
3. Convenciones	DC1: Inconsistencia gramatical entre la base del reactivo y las opciones en ítems de opción múltiple
	DC2: Uso inapropiado de puntos suspensivos para denotar continuidad entre la base y las opciones
	DC3: Cambio en el orden de las opciones; inconsistencia gramatical entre las opciones
	DC4: Uso inapropiado de mayúsculas al principio de las opciones
4. Información	CI1: Traducción inconsistente de un término no técnico que se repite varias veces en el original
	CI2: Un término no técnico importante aparece más o menos veces que en el original
	CI3: Cambio en la forma de escribir números
	CI4: Inserción de términos no técnicos, oraciones o explicaciones que no aparecen en el original
	CI5: Omisión de palabras importantes, términos no técnicos u oraciones o explicaciones que sí aparecen en el original
5. Gramática	LG1: Traducción literal (palabra por palabra)
	LG2: Estructura sintáctica no natural
	LG3: Inconsistencia sujeto–verbo
	LG4: Inconsistencia de singulares y plurales
	LG5: Uso inapropiado de preposiciones
	LG6: Uso inapropiado en la concordancia de tiempos
	LG7: Conflación de dos oraciones en una oración

Dimensión	Categoría
6. Semántica	LS1: Uso de cognados falsos
	LS2: Traducción impropia de expresiones idiomáticas
	LS3: Alteración en el sentido de una oración
	LS4: Una oración se puede interpretar de más de una manera
	LS5: No es claro lo que el ítem pide que se haga
	LS6: Cambio en el género de personajes
	LS7: Combinación de dos o más enunciados en uno
	LS8: Uso impreciso de términos y vocablos
	LS9: Uso de términos con significados múltiples
	LS10: Traducción inapropiada de un término
7. Constructo	CC1: Posible alteración de las demandas cognoscitivas del ítem
	CC2: Posible alteración de la forma en que se puede interpretar el ítem
	CC3: Uso impreciso de términos técnicos
	CC4: Traducción inconsistente de un término técnico que se repite varias veces en el original
	CC5: Inserción de términos técnicos
	CC6: Omisión de términos técnicos
	CC7: Sustitución de un término técnico por un término no técnico
	CC8: Sustitución de un término no técnico por un término técnico
8. Registro	LR1: Uso de palabras de baja frecuencia entre la población destinataria
	LR2: Traducción incorrecta de términos técnicos
	LR3: Traducción correcta de términos técnicos pero de una manera que no se enseña en el país
9. Cultura	CQ1: La información contextual y las situaciones que el ítem proporciona son poco comunes en la cultura nacional
	CQ2: Las unidades de medición son ajenas a la cultura nacional
	CQ3: El tipo de problema planteado por el ítem no tiene sentido en la cultura nacional
	CQ4: El ítem evalúa conocimientos que no se enseñan en el país
10. Origen	OO1: Hay más de una respuesta correcta
	OO2: Ninguna de las opciones es completamente correcta
	OO3: Inconsistencia entre las dos versiones fuente

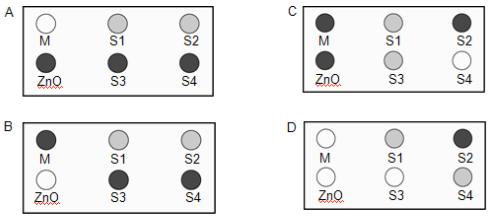
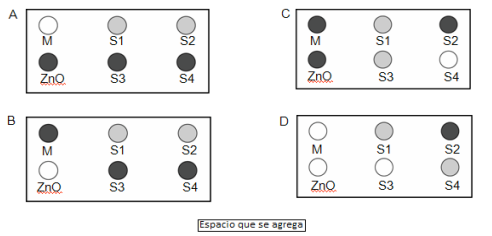
Ejemplo de error de Estilo (DE1. Errores de puntuación). Se abre el signo de interrogación en un lugar incorrecto.

Original	Traducción
Can these claims made in the article be tested through scientific investigation in the laboratory?	Las afirmaciones que se hacen en el artículo ¿se pueden comprobar mediante investigación científica en un laboratorio?

Ítem liberado de ciencias. Prueba PISA–2006.

Dimensión 2: *Formato.* El formato o composición visual del ítem traducido difiere del original en el idioma fuente. Por ejemplo: tamaño diferente de tablas, estilo diferente de fuentes de caracteres, márgenes más reducidos, e inserción u omisión de componentes gráficos.

Ejemplo de error de Formato (DF5. Inserción de componentes gráficos). Inserción de espacio después de los diagramas.

Original	Traducción
 Answer: ..... Explanation:	 Respuesta: ..... Explicación: .....

Ítem liberado de ciencias. Prueba PISA–2006.

Dimensión 3: *Convenciones.* La traducción del ítem no se realiza de conformidad con las prácticas convencionales de la escritura de ítems en el idioma destinatario o con los principios básicos de escritura técnica de ítems. Por ejemplo: inconsistencia gramatical entre la base y las opciones en ítems de opción múltiple, inconsistencia gramatical entre las opciones en ítems de opción múltiple, y extensión diferente de la respuesta correcta en ítems de opción múltiple.

Ejemplo de error de convenciones (DC5. Otros). Como no se tradujo la palabra *make*, se alteró la estructura de la opción D del ítem.

Original	Traducción
Why was the second sheet of plastic pressed down? *A) To stop the drops from drying out.* *B) To spread the drops out as far as possible.*	¿Por qué se hizo presión sobre el segundo pliego de plástico? A) Para impedir que las gotas se secaran. B) Para extender las gotas lo más posible. C) Para mantener las gotas dentro de los círculos que se marcaron. D) Para que las gotas tuvieran el mismo grosor.

Ítem liberado de ciencias. Prueba PISA–2006.

Dimensión 4: *Información.* La traducción cambia la cantidad, la calidad, o el contenido de información crítica para entender de qué se trata el ítem y lo que debe hacerse para responderlo. Por ejemplo: traducción inconsistente de un término que se repite varias veces en el original, y un término clave aparece más o menos veces que en el original.

Ejemplo de error de Información (CI4. Inserción de términos no técnicos, oraciones o explicaciones que no aparecen en el original). Se insertó la frase "Es posible construir un modelo".

Original	Traducción
The effect of acid rain on marble can be modeled by placing chips of marble in vinegar overnight.	Es posible construir un modelo del efecto de la lluvia ácida sobre el mármol, poniendo fragmentos de éste en vinagre toda la noche.

Ítem liberado de ciencias. Prueba PISA–2006.

Dimensión 5: *Gramática.* La traducción del ítem tiene errores gramaticales o la sintaxis es innecesariamente compleja o inusual para la población destinataria. Por ejemplo: traducción literal (palabra por palabra), estructura sintáctica no natural, y uso inapropiado de preposiciones.

Ejemplo de error de Gramática (LG2. Estructura sintáctica no natural). El término "agua (destilada) pura" no se usa en el idioma destinatario.

Original	Traducción
Students who did this experiment also placed marble chips in pure (distilled) water overnight.	Los estudiantes que hicieron este experimento también pusieron fragmentos de mármol en agua (destilada) pura toda la noche.

Ítem liberado de ciencias. Prueba PISA–2006.

Dimensión 6: *Semántica.* Las ideas y el significando transferidos al ítem traducido no son iguales a los del ítem en el idioma fuente. Por ejemplo: uso de cognados falsos y traducción impropia de expresiones idiomáticas.

Ejemplo de error de Semántica (LS10. Traducción inapropiada de un término). El verbo *model* (reproducirse o replicarse) se tradujo como "construir un modelo".

Original	Traducción
The effect of acid rain on marble can be ***modeled*** *by placing chips of marble in vinegar overnight.*	Es posible **construir un modelo** del efecto de la lluvia ácida sobre el mármol, poniendo fragmentos de éste en vinagre toda la noche.

Ítem liberado de ciencias. Prueba PISA–2006.

Dimensión 7: *Constructo.* La traducción altera el tipo de conocimiento o de habilidades necesarios para responder correctamente el ítem. Por ejemplo: traducción inexacta de términos técnicos; inserción u omisión de términos técnicos.

Ejemplo de error de Constructo (7CC1. Posible alteración de las demandas cognoscitivas del ítem). Traducir *can be modeled* (puede reproducirse o replicarse) como "Es posible construir un modelo", hace más complejo el ítem pues cambia el contenido de información crítica para entenderlo.

Ejemplo de error de Constructo (7CC2. Posible alteración de la forma en que se puede interpretar el ítem). *Overnight* se tradujo como "toda la noche". ¿Se dejaron los fragmentos de mármol en el vinagre durante la noche o que la acción de colocar los fragmentos de mármol duró toda la noche? (ambigüedad introducida por uso del gerundio)

Original	Traducción
The effect of acid rain on marble ***can be modeled*** *by placing chips of marble in vinegar* ***overnight.***	**Es posible construir un modelo** del efecto de la lluvia ácida sobre el mármol, poniendo fragmentos de éste en vinagre **toda la noche.**

Ítem liberado de ciencias. Prueba PISA–2006.

Dimensión 8: *Registro.* La traducción del ítem no es sensible al uso común de palabras o a los diferentes contextos sociales en la población destinataria. Por ejemplo: uso de palabras de baja frecuencia entre la población destinataria, y; traducción correcta de términos técnicos, pero de una manera que no es común en las escuelas o en los libros de texto del país.

Ejemplo de error de Registro (LR2. Traducción incorrecta de términos técnicos). El témino *increased* (incrementado) fue traducido como "aumentado".

Original	Traducción
It is a fact that the average temperature of the Earth's atmosphere has increased.	Es un hecho que la temperatura promedio de la atmósfera terrestre ha aumentado.

Ítem liberado de ciencias. Prueba PISA–2006.

Dimensión 9: *Cultura.* El ítem no representa la cultura o currículo del país destinatario. Por ejemplo: el conocimiento o la habilidad que evalúa el ítem no se enseña en el país en el grado escolar de la prueba, o en grados anteriores, y; la manera de plantear un problema no se usa en el currículo del país destinatario.

Ejemplo de error de Cultura (CQ4. El ítem evalúa conocimientos que no se enseñan en el país). Lógica y teoría de conjuntos son temas que no son parte del currículum de primaria. Además, la transitividad no se enseña hasta la secundaria como pre-álgebra.

Original	Traducción
Henry is older than Bill, and Bill is older than Peter. Which statement must be true? *A. Henry is older than Peter.* *B. Henry is younger than Peter.* *C. Henry is the same age as Peter.* *D. We cannot tell who is oldest from the information.*	Enrique es mayor que Guillermo, y Guillermo es mayor que Pedro. ¿Cuál de las siguientes afirmaciones debe ser verdadera? A) Enrique es mayor que Pedro. B) Enrique es más joven que Pedro. C) Enrique es de la misma edad que Pedro. D) La información no es suficiente para saber quién es mayor.

Ítem liberado de matemáticas. Población 1 (nueve años). Prueba TIMSS–1995.

Dimensión 10: *Origen*. El ítem en el lenguaje fuente tiene fallas que no pueden corregirse en la traducción, lo que impone limitaciones para su adecuada traducción. Por ejemplo: hay más de una respuesta correcta; ninguna de las opciones es completamente correcta.

Ejemplo de error de Origen (OO3. Inconsistencia entre las dos versiones fuente). En la versión en inglés, Mary estaba convencida de la "seguridad de estas inoculaciones"; por ello "permitió" que sus hijos fueran inoculados. En la versión en francés, Mary estaba convencida "de que las inoculaciones no eran peligrosas"; por ello "hizo inocular a sus hijos".

Original en francés	Original en inglés
Mary Montagu fut si convaincue que ces inoculations étaient sans danger qu'elle fit inoculer son fils et sa fille.	Mary Montagu was so convinced of the safety of these inoculations that she allowed her son and daughter to be inoculated.

Ítem liberado de ciencias. Prueba PISA–2006.

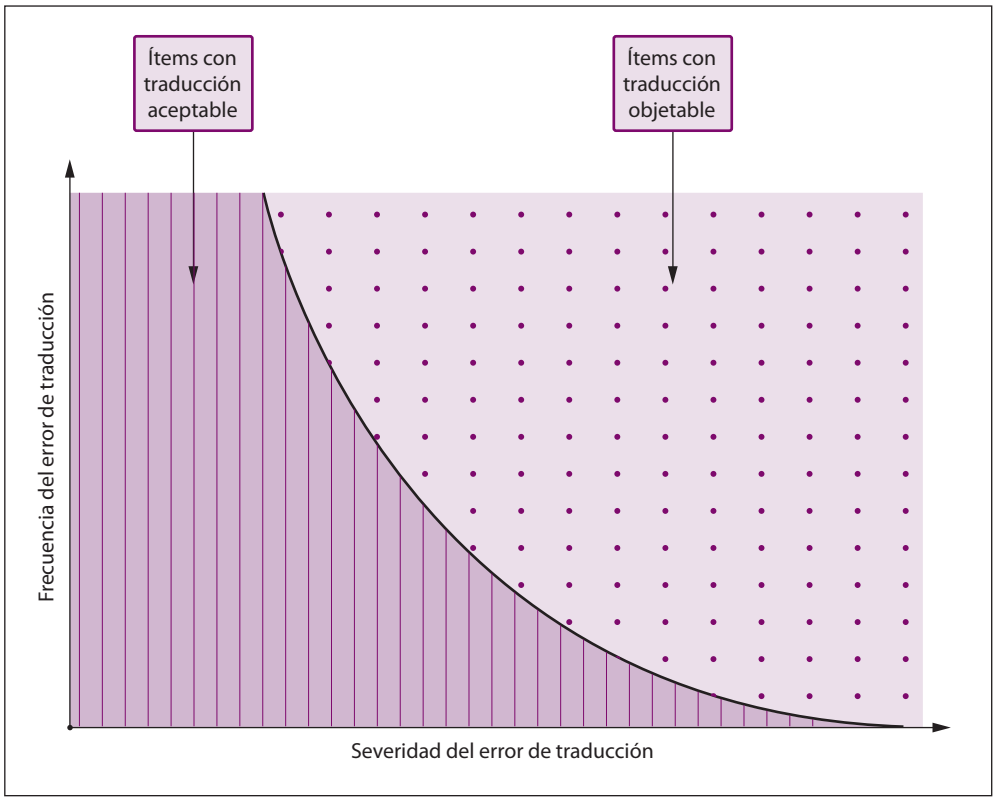
Espacio probabilístico del error de traducción

La TETP concibe al error de traducción desde un enfoque probabilístico. Lo anterior implica que los ítems aceptables no están totalmente libres de error y que no todos los ítems con errores son necesariamente objetables. La figura 1 muestra el grado en que es aceptable u objetable la traducción de un ítem como un espacio probabilístico. Tal espacio está delimitado por el número de errores de traducción (eje vertical) y por la severidad de esos errores (eje horizontal) (Solano-Flores, Contreras-Niño, y Backhoff, 2005). Así, un ítem traducido está ubicado en el área de aceptabilidad (área clara) porque tiene uno o varios errores leves. Por su parte, un ítem traducido está en el área objetable (área obscura) porque tiene uno o más errores de traducción severos, o porque tiene muchos errores leves.

Como puede observarse en la figura 1, el área donde se ubican los ítems con traducciones objetables hace contacto con el eje de severidad del error de traducción, lo cual significa, como ya se dijo, que la traducción de un ítem puede ser objetable por tener un solo error de traducción severo. Por ejemplo, cuando se altera el constructo que se mide, se modifica el significado de un término clave o se altera la información necesaria para responder el ítem. En cambio, puede observarse que el área de cuestionabilidad no hace contacto con el eje de frecuencia del error de traducción. Ello significa que no es posible determinar *a priori* un número máximo de errores leves que harían objetable la traducción de un ítem. Tal decisión resulta por necesidad una operación contextual

basada en juicios de expertos, como los del panel multidisciplinario propuesto más adelante en este trabajo.

Figura 1. Espacio probabilístico de ítems aceptables y objetables, definido por la frecuencia y severidad de los errores de traducción



Fuente: Solano-Flores, Backhoff, y Contreras-Niño, 2009.

Metodología para identificar errores de traducción

La revisión de ítems traducidos al español de México se ha llevado a cabo con siete acciones: 1) empleo de un equipo multidisciplinario de revisores, de acuerdo con el tipo de prueba a analizar; 2) familiarización de los revisores con cada uno de los ítems, 3) codificación individual de errores en cada ítem por los revisores, en las dimensiones y categorías de error que correspondan, 4) codificación colegiada de los errores identificados en cada ítem, hasta llegar a un consenso sobre los errores que presenta, 5) clasificación de los ítems como aceptables u objetables en función de la frecuencia y gravedad de los errores detectados, 6) preparación de la información recabada, a fin de consignar y analizar errores de traducción, y 7) análisis estadístico de los ítems con errores para conocer su impacto en las respuestas de las personas examinadas. A continuación se describe brevemente cada uno de estos elementos.

Empleo de un equipo multidisciplinario de revisores. Los errores de traducción correspondientes a las distintas dimensiones y categorías de error no pueden ser examinados con suficiente detalle si los revisores no son sensibles a distintas fallas que puede presentar el ítem. La composición de un comité de revisión de traducción debe ser sensible tanto a las características de la prueba como a sus propósitos y contenido, así como a las particularidades del lenguaje y la cultura fuente y destinataria.

La integración de un equipo multidisciplinario de revisores debe considerar idealmente los siguientes especialistas: 1) traductores profesionales certificados, en los idiomas fuente y destinatario, que puedan evaluar la precisión de la traducción desde una perspectiva técnica, 2) lingüistas que puedan examinar los aspectos estructurales, funcionales y semánticos de la traducción, 3) especialistas en currículo que puedan determinar si el contenido a evaluar y la terminología empleada en la traducción de los ítems corresponde al nivel de complejidad del currículo oficial, 4) maestros en ejercicio y con amplia experiencia en la disciplina y grados escolares correspondientes, conocedores del currículum, los libros de texto y el uso formal y coloquial del idioma en el contexto de la enseñanza, 5) especialistas en medición que puedan analizar el apego a las convenciones de escritura técnica de ítems y las demandas cognoscitivas generadas por las propiedades lingüísticas de los ítems en el idioma original y en el traducido, y 6) expertos en evaluación educativa con especialidad en sesgo cultural.

Familiarización de los revisores con los ítems traducidos. Con el objetivo de familiarizar a los revisores con los reactivos traducidos por revisar, se les solicita que los lean y los respondan como si fueran estudiantes. Con ello se asegura que los revisores examinen detenidamente cada reactivo y sean conscientes del tipo de demandas cognitivas necesarias para responderlo. Es indispensable que el revisor no conozca la respuesta correcta del ítem antes de contestarlo, de modo que se propicie la atención necesaria para responderlo como lo haría un estudiante.

Codificación individual de errores en un ítem. Una vez que responden al ítem traducido, los revisores proceden a identificar sus errores de traducción. Para ello, llenan el formato de evaluación mostrado en la tabla 2. En él, codifican, de manera independiente, los errores identificados en el ítem de conformidad con las dimensiones y categorías de error correspondientes, justifican sus juicios y anotan las observaciones que consideran pertinentes. Las versiones del ítem bajo análisis —en español, inglés y francés, en el caso de PISA— se proyectan simultáneamente en una pantalla, de tal manera que los revisores pueden comparar las versiones originales del ítem, así como su versión traducida al español.

Sin embargo, es importante hacer notar que los revisores codifican los errores de traducción de acuerdo con sus áreas de especialización, aunque se les pide indicar errores que detecten en cualquier dimensión. Por principio, el traductor se enfoca en las dimensiones de Estilo, Formato, Gramática, Semántica y Origen; el lingüista atiende las dimensiones de Gramática, Semántica y Registro; los maestros y expertos en currículo se orientan en las dimensiones de Información y Cultura; el especialista en medición y el psicómetra se concentran en las dimensiones de Estilo, Formato, Convenciones, Información, Constructo y Origen.

Para apoyar la codificación de los errores de traducción de los reactivos, los revisores utilizan la tabla de codificación que describe las diez dimensiones y 53 categorías de error de traducción utilizadas hasta el momento (ver tabla 1). Sin embargo, cabe señalar la posibilidad que existan nuevas categorías por agregar a este listado. Por tal razón, en cada dimensión existe la categoría *Otras* (que por economía de espacio no se presentó en la tabla 1), misma que se deberá utilizar cuando un revisor identifique un error que no haya sido codificado previamente.

Tabla 2. Sección de la hoja de codificación de errores de traducción

Ítem	Tipo de error	Codificación y justificación
(ingresar el número de ítem)	1. Estilo	
	2. Formato	
	3. Convenciones	
	4. Información	

PISA y otras pruebas internacionales utilizan familias de ítems, las cuales se componen de un texto introductorio y de varios reactivos asociados a esta introducción; usualmente, de dos a cuatro ítems. Por lo general, en los párrafos introductorios se da información contextual que es indispensable para responder los ítems asociados. A estos dos elementos (introducción e ítems) les denominamos unidades analíticas (UA), por lo que sus errores de traducción se analizan de manera independiente. Un ejemplo de familia de ítems se presenta en el Anexo 1.

Codificación colegiada de los errores identificados en un ítem. Cuando los especialistas terminan de analizar de manera individual los errores de traducción de un ítem, se procede a revisarlos grupalmente. Como ya se mencionó, las versiones del ítem que se va a analizar se proyectan en los idiomas fuente y destinatario. En esta situación, se discute si el reactivo correspondiente presenta errores en cada una de las diez dimensiones de error señaladas en la tabla de codificación. Los especialistas que hayan identificado algún error de traducción en el ítem deben de señalarlo y justificar su codificación al nivel de categoría. Los revisores que no están de acuerdo en que el ítem presente dicho error deben de justificar su postura y debatir el punto hasta alcanzar un consenso sobre la presencia o ausencia del error en la categoría en discusión. Una vez consensuados los errores de traducción en el ítem, los revisores modifican sus codificaciones originales en los formatos individuales, a fin de que éstas reflejen las decisiones tomadas por el comité. Este procedimiento se sigue con la totalidad de dimensiones y categorías de error, hasta terminar de analizar el ítem completamente. En el anexo 2 se presenta una sección de la discusión que se genera en el comité.

Al terminar la revisión colegiada de un reactivo, se procede a analizar grupalmente el siguiente ítem. Con el propósito de ejemplificar el caso de un reactivo al que se le han detectado errores de traducción, en el anexo 3 se presenta un ítem revisado que presenta errores en distintas dimensiones y categorías de error.

Antes de continuar con la descripción de la siguiente etapa, es importante enfatizar dos aspectos importantes del trabajo colegiado y su conducción:

1. El trabajo de grupo multidisciplinario es coordinado por el responsable del proyecto. La discusión de un solo ítem dura en promedio 15 minutos, aunque puede llegar a durar hasta 45. El enfoque multidisciplinario permite identificar errores de difícil detección cuando se trabaja de manera individual, lo cual suele suceder en el caso de los errores de Origen (versión original del ítem). Asimismo, es indispensable que los revisores tengan acceso a documentos curriculares (plan y programas de estudios, libros de texto y guías del maestro), así como a información liberada sobre el contenido de los ítems de la prueba en revisión. Estos documentos son importantes, pues permiten a los revisores entender el

significado de los reactivos en los idiomas fuentes y en el idioma destinatario, así como los conocimientos y habilidades que pretenden evaluar. En caso que los revisores tengan acceso a ítems no liberados de alguna prueba, deben firmar un contrato de confidencialidad en el cual se comprometen a no extraer ni divulgar dicha información.

2. El trabajo del coordinador es vital, toda vez que da orden a las actividades individuales y grupales de los especialistas, además de moderar y ponderar las discusiones para propiciar que alcancen los consensos sobre los errores de traducción de los ítems revisados. Éste requiere de la ayuda de un asistente que registre las decisiones del grupo en una base de datos. Asimismo, se auxilia del asistente para grabar en audio las discusiones generadas durante el proceso de revisión de los reactivos. Finalmente trabaja con personal diverso para que toda la información recabada sobre los ítems se transcriba y se capture en un programa *ad hoc* para su análisis posterior.

Clasificación de los ítems como aceptables u objetables. Una vez que fueron consensuados los errores de traducción de un reactivo, el comité formula un juicio sintético sobre el estatus de la traducción del ítem. Este juicio se realiza a partir de la frecuencia y severidad de los errores encontrados, a fin de clasificar al ítem en una de cuatro categorías: *pocos leves*, *pocos graves*, *muchos leves* y *muchos graves*. A partir de esta clasificación, el comité califica finalmente la traducción de un ítem como *aceptable* o como *objetable*, de acuerdo con la cantidad y severidad de los errores identificados.

Preparación de la información recabada. La información derivada de los estudios de revisión de traducciones de pruebas consiste de tres productos principales:

1. *Los protocolos de evaluación* en papel que utilizaron los jueces para llevar a cabo la evaluación individual de los ítems. Dichos formatos de evaluación consignan los errores de traducción que detectó cada juez, en cada reactivo evaluado, según las dimensiones y categorías de error contemplados en el protocolo.

2. *La base de datos* que consigna —los errores de traducción de cada ítem evaluado que fueron consensuados por el comité evaluador, según las categorías y dimensiones de error consideradas en el protocolo de evaluación.

3. *Las grabaciones* de las discusiones que tuvieron lugar entre los miembros del comité de expertos al evaluar cada ítem, hasta que lograron el consenso respecto a cada error de traducción y su ubicación adecuada en una categoría y dimensión de error determinadas.

Estos productos pueden ser considerados como los datos crudos que arroja cada estudio sobre el error de traducción de una prueba. En consecuencia, para facilitar su procesamiento y ser analizados de manera significativa, tanto cuantitativa como cualitativamente, se procede a darles el siguiente tratamiento:

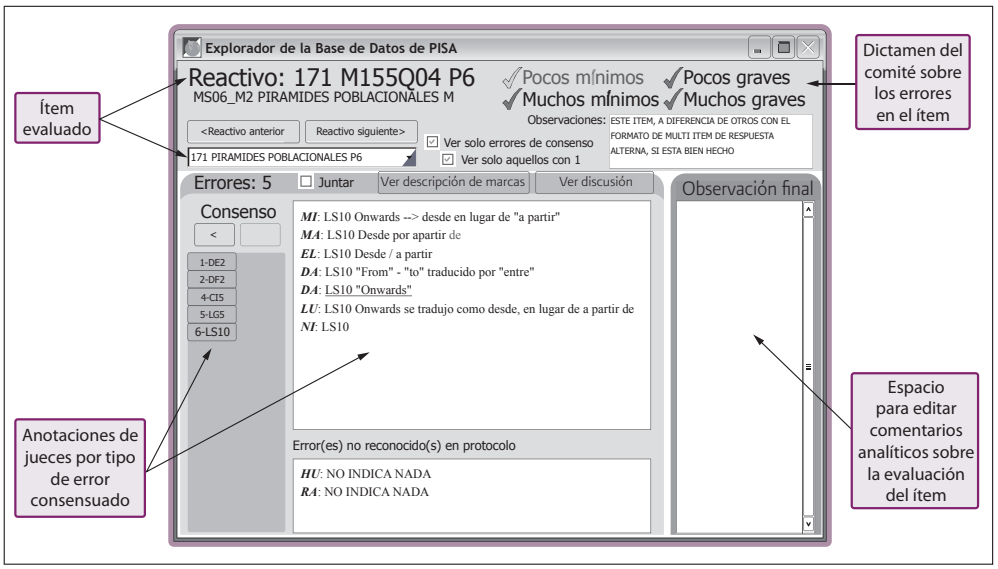
1. Los protocolos de evaluación en papel de los jueces, que consignan los errores de traducción detectados en cada reactivo evaluado, son transcritos a archivos digitalizados y estructurados en una base de datos.

2. La base de datos que consigna los errores de traducción que fueron consensuados en cada ítem evaluado, es depurada y validada.

3. Las grabaciones que contienen las discusiones suscitadas entre los miembros del comité al evaluar cada ítem, son transcritos a archivos digitalizados y estructurados en una base de datos.

Preparación de la información recabada. Con el propósito de facilitar el análisis de la información contenida en estas tres bases de datos, se elaboró una aplicación de cómputo que permite integrar la información disponible sobre todos y cada uno de los reactivos evaluados. Su propósito es contar con una herramienta que permita observar de manera simultánea toda la información sobre los errores de traducción recabados durante el trabajo de revisión, de modo que sea posible efectuar diversos análisis cualitativos y cuantitativos de los resultados.

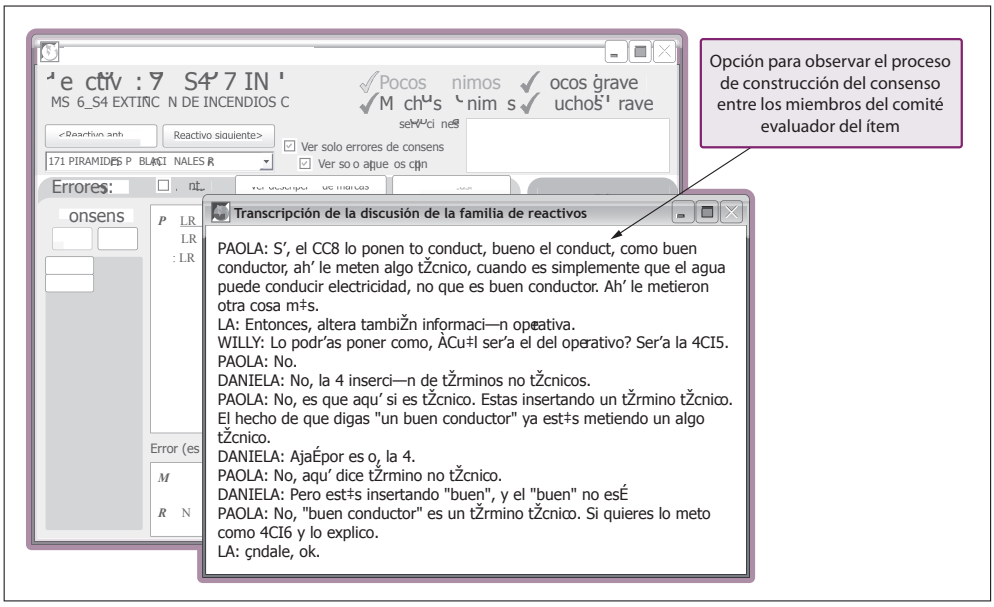
Figura 2. Pantalla donde se visualizan los errores de traducción de los ítems



La figura 2 ilustra los principales elementos del programa de cómputo que permiten visualizar la información de los errores de traducción de los ítems revisados. En términos generales, esta información es la siguiente: 1) identificación de cada ítem evaluado, 2) codificación de los errores de cada ítem por dimensión y categoría de error, 3) observaciones de cada revisor sobre cada uno de los ítems, 4) transcripción de las discusiones del comité evaluador, 5) clasificación de la severidad de los errores que presenta cada ítem, y 6) en su caso, observaciones particulares y grupales de la traducción de cada ítem.

En la figura 3 se presenta la pantalla que surge al activar el botón *Ver discusión*, la cual muestra una sección de una discusión suscitada entre los miembros del comité evaluador, en relación con sus observaciones individuales referidas al error de traducción correspondiente a la categoría LS10 (Traducción inapropiada de una palabra).

Figura 3. Pantalla que ilustra una discusión entre miembros del comité evaluador



De manera resumida, el programa de cómputo diseñado presenta los siguientes elementos y características:

- En la parte superior izquierda de la pantalla principal (ver figura 2) aparecen los elementos de identificación del ítem que fue posicionado y los elementos para navegar entre los demás reactivos evaluados. Así, es posible posicionarse en el reactivo anterior, en el siguiente o en cualquier otro que aparece en la lista de los ítems evaluados.

- En la parte superior derecha de la pantalla aparecen dos elementos que expresan una evaluación global del ítem considerado: a) un dictamen general consensuado sobre los errores de traducción del reactivo, mismo que está basado en una rúbrica holística que considera simultáneamente la cantidad y la gravedad de los errores identificados por el comité evaluador (pocos errores leves, muchos errores leves, pocos errores graves y muchos errores graves); y b) un comentario general que hace explícita y significativa la naturaleza del dictamen adoptado. Cabe señalar que la frecuencia y la gravedad de los errores encontrados en el ítem por el comité no son categorías mutuamente excluyentes; por lo que al formular el dictamen sobre un reactivo, los jueces pudieron juzgar que presenta simultáneamente pocos o muchos errores leves, a la vez que puede tener pocos o muchos errores graves, según sea el caso. (véase por ejemplo el caso de la figura 3).

- Como ya se señaló, en el recuadro central (ver figura 2) aparecen las anotaciones hechas por los miembros del comité evaluador cuando efectuaron la evaluación individual del ítem; y en la parte media, a la izquierda de la pantalla, se muestra el tipo de error señalado por los jueces (dimensión y categoría de error, en negritas), junto con los demás tipos

de error que fueron o no consensuados (según estén o no seleccionados los cuadros de cotejo ubicados encima del botón *Ver descripción de marcas*).

- El botón identificado como *Ver descripción de marcas*, colocado arriba del recuadro al centro de la pantalla (ver figura 2), presenta una descripción explícita que aclara el significado de los elementos de codificación empleados al consignar las observaciones de los jueces en el recuadro al centro de la pantalla, tales como: quién es el evaluador (lo que se identifica en el recuadro con las dos primeras iniciales de su nombre) y la observación que está subrayada (lo que significa que la codificación formulada por el revisor correspondiente fue consensuada por el comité).

- Como también se señaló, al apretar el botón *Ver discusión*, ubicado en la parte superior central de la pantalla, se muestra una ventana con la discusión que tuvo lugar entre los miembros del comité evaluador, a partir de sus observaciones individuales referidas a los errores de traducción que identificaron en cada reactivo, según la categoría y la dimensión de error consideradas (véase al respecto la figura 3).

- En la parte inferior central de la pantalla (ver figura 2) aparece un recuadro que identifica errores no reconocidos en el protocolo evaluativo que cumplimentó un juez o la ausencia de observaciones; es decir, el juez no formuló observación alguna, o anotó información relativa a algún error en el ítem para el cual no es posible identificar la categoría o dimensión de error al que se refiere, o hizo algún otro comentario general o particular que no fue posible entender y ubicar en la base de datos.

- En la parte superior izquierda de la pantalla (ver figura 2), debajo de los elementos de identificación del ítem que se evalúa, aparecen tres más, identificados con los nombres: *Errores*, *Juntar* y *Consenso*. El primero de ellos se refiere al número de categorías correspondientes a diferentes dimensiones de error, en las cuales se identificó al menos un error en el ítem. El segundo, a la opción que ofrece el programa para: a) agrupar y presentar en el recuadro central de la pantalla todas las observaciones, de todos los jueces, correspondientes a todas las categorías y dimensiones de error identificadas en el ítem a evaluar, a fin de observarlas todas juntas, o b) presentar en el recuadro central solo aquellas observaciones de los jueces correspondientes a una sola categoría y dimensión de error particulares (como es el caso en la figura 2).

- En la parte media, a la derecha de la pantalla de la figura 2, se presenta un recuadro de *Observación final*, el cual permite a los investigadores que analizan el conjunto de datos que fueron integrados por la aplicación poder efectuar observaciones, comentarios y juicios evaluativos de corte cualitativo, sobre los errores de traducción identificados en cada reactivo.

- Finalmente, cabe hacer notar la flexibilidad del programa de cómputo para seleccionar, articular y visualizar la información integrada a partir de las tres bases de datos originales. Ello permite observar de manera simultánea toda la información disponible sobre los errores de traducción, de modo que sea posible efectuar la comparación de los resultados y realizar distintos análisis cualitativos.

Análisis estadísticos. Una vez concentrada la información sobre los errores de traducción de los ítems de una prueba, se procede a analizarlos cuantitativamente, a fin de conocer su incidencia y su posible impacto en la ejecución de las personas evaluadas. Dado que se puede realizar una gran diversidad de análisis estadísticos con la información recabada, a continuación solo se mencionan algunos como ejemplos. Los errores de traducción se pueden analizar de acuerdo con: la frecuencia de errores, la cantidad de dimensiones que presentan al menos un error, la severidad del error, etcétera. En cuanto al impacto de los errores de traducción en la ejecución de los respondientes, se puede analizar la correlación que existe entre ciertas propiedades psicométricas de los reactivos, como es su dificultad y discriminación, con la incidencia de errores.

Resultados de dos estudios para la revisión de la traducción de pruebas: TIMSS–1995 y PISA–2006

En este apartado se describen dos estudios en los que se utilizó la TETP para revisar las traducciones de un par de pruebas internacionales importantes, utilizadas para comparar la calidad de los sistemas educativos de los países participantes: TIMSS y PISA. Dado que la TETP se fue desarrollando a partir de estos trabajos, en 2003 con el estudio de TIMSS–1995, y en 2008 con el de PISA–2006, presentaremos una síntesis de estos estudios en el orden como fueron realizados. Con ello se podrán apreciar algunos cambios en la teoría y metodología utilizadas que ilustran su evolución.

Estudio de TIMSS–1995

La información sobre la aplicación de TIMSS–1995 en México es prácticamente desconocida, debido a que las autoridades educativas del país decidieron retirar su participación antes de la publicación del informe internacional. Ello ocasionó que la información de este estudio no se analizara, ni se publicara en algún medio impreso o electrónico. Sin embargo, después de la creación del INEE (2002), este instituto financió dos estudios para conocer los resultados de México en este estudio internacional y examinar la calidad de la traducción mexicana de esta prueba.

La investigación sobre la traducción mexicana de TIMSS–1995 tuvo dos objetivos centrales: 1) analizar la calidad de la traducción al español de los reactivos de las pruebas de ciencias naturales y de matemáticas, a través del trabajo de un comité multidisciplinario de especialistas, y 2) analizar cómo influyen los errores de traducción de las pruebas internacionales en el desempeño de los estudiantes. Los resultados de esta investigación se dieron a conocer en el trabajo, *The Mexican translation of TIMSS-95: Test translation lessons from a post-mortem study* (Solano-Flores, Contreras-Niño y Backhoff-Escudero, 2005) y en el documento para el INEE: *Informe de resultados iniciales del análisis de la traducción de la prueba PISA-2006* (Contreras-Niño, Solano-Flores y Backhoff, 2008).

Muestra de ítems analizados. El INEE pudo recuperar los cuadernillos utilizados en dos poblaciones objeto de TIMSS–1995: estudiantes de 9 años (población 1), que cursaban el 3° y 4° grados de primaria, así como estudiantes de 13 años (población 2), inscritos en 1° y 2° grados de secundaria.

También fue posible recuperar de la Secretaría de Educación Pública (SEP) información sobre algunos valores *p* de los reactivos de la aplicación mexicana de 1995. Estos valores se definen como la "dificultad" de un reactivo, en términos de la proporción de aciertos obtenidos por la población de estudiantes que respondieron este ítem; entre más grande sea el valor de *p*, menor será la dificultad del reactivo.

La tabla 3 muestra el total de ítems analizados, que ascendió a 319, de los cuales 169 fueron diseñados para estudiantes de 9 años, y 150 para la población de alumnos de 13 años. Del total, 164 reactivos eran de matemáticas y 155 de ciencias naturales.

Tabla 3. Reactivos analizados de la prueba TIMSS-1995

Población	Matemáticas	Ciencias naturales	Total
Estudiantes de 9 años (3° y 4° de primaria)	88	81	169
Estudiantes de 13 años (1° y 2° de secundaria)	76	74	150
Total	164	155	319

Para examinar el impacto de la calidad de la traducción en el desempeño de los estudiantes, se correlacionaron la cantidad de errores de traducción con los valores *p* de 42 ítems de matemáticas y 39 ítems de ciencias naturales, de la población de 9 años, y 19 ítems de matemáticas y 23 de ciencias naturales, de la población de 13 años.

Muestra de estudiantes. La tabla 4 describe la población total evaluada en el estudio de TIMSS–1995, que sumó 44 mil 968 estudiantes: 20 mil 316 de 9 años y 24 mil 652 de 13 años. Es importante decir que los grupos de alumnos participantes en este proyecto internacional fueron, en su momento, representativos de México.

Tabla 4. Población de estudiantes evaluados en el estudio TIMSS–1995

Población	Grado Escolar	Población
Estudiantes de 9 años	3o de primaria	10,122
	4o de primaria	10,194
Estudiantes de 13 años	1o de secundaria	12,809
	2o de secundaria	11,843
Total		44,968

Composición del comité revisor. Para evaluar la calidad de la traducción mexicana de los ítems de la prueba TIMSS–1995, se integró un comité de especialistas integrado por una traductora inglés – español certificada por la ATA (siglas en inglés de Asociación de Traductores de Norteamérica); seis profesores en servicio especialistas en docencia, dos de ellos del quinto grado de educación primaria, y cuatro más de secundaria, de las asignaturas de matemáticas, física, química y biología; un especialista en el *curriculum* de ciencias naturales de la educación secundaria; una especialista en el área de lingüística; un psicómetra especialista en medición educativa, y; un especialista en el área de desarrollo de pruebas.

Clasificación de ítems. Al finalizar el análisis de errores de traducción, el comité multidisciplinario acordó clasificar cada ítem como aceptable u objetable. Esta decisión fue tomada, como ya se explicó anteriormente, con base en el número y la severidad de los errores encontrados.

Resultados. Básicamente, en este trabajo se realizaron dos tipos de análisis: la incidencia de errores de traducción y su relación con la ejecución de los estudiantes. El estudio de la incidencia de errores tomó en cuenta el tipo ítem, las dimensiones de error, las poblaciones de estudiantes y los dominios curriculares de la prueba TIMSS–1995. Por su parte, la relación entre los errores de traducción y la ejecución de los estudiantes se realizó a través de un análisis de correlaciones utilizando los valores *p* de los ítems.

Una pregunta que se quiso responder tuvo que ver con las diferencias entre los ítems clasificados como aceptables y objetables, en términos de la distribución de errores. La tabla 5 muestra que cerca de 7.5% de los ítems analizados (24) fue clasificado como objetable y 92.5% como aceptable (295). Ambos tipos de ítems se diferencian ligeramente por la cantidad total de errores que presentan: así, mientras que la media de los primeros es de 8.8, la de los segundos es de 6.3; cifras similares se observan en la mediana; pero no en la moda, que en ambos casos es de 8 errores. En cuanto a la dispersión, la tabla muestra medidas muy similares: desviaciones estándar de 3.4 (aceptables) y 3.2 (objetables) y un rango de 17 errores en ambos tipos de ítems. Estos resultados nos indican que si bien no existe una diferencia importante en el número de errores, la gravedad o severidad de los errores es la propiedad que distingue a un reactivo considerado como aceptable u objetable.

Tabla 5. Estadística descriptiva de ítems identificados como aceptables y objetables de TIMSS–1995

Reactivos	k =	Media	D.E.	Mediana	Moda	Rango
Aceptables	295	6.25	3.37	6.0	8.0	0-17
Objetables	24	8.83	3.19	8.5	8.0	3-20

Fuente: Solano-Flores, Contreras-Niño y Backhoff, 2006

Una segunda pregunta que nos hicimos tuvo que ver con la distribución de los errores de traducción entre los ítems. Para ello, se presenta la tabla 6 que muestra el porcentaje de reactivos con errores en cada dimensión, de acuerdo con la población de estudiantes y el dominio curricular de la prueba (matemáticas y ciencias naturales). En esta tabla se puede observar que la gran mayoría de ítems tuvo algún tipo de error y que en promedio:

- Hay un mayor porcentaje de ítems con errores en las dimensiones de Semántica, Formato e información, mientras que hay un menor porcentaje de ítems con errores en las dimensiones de Cultura (currículo), Registro y Origen.
- Hay un mayor porcentaje de ítems con errores de traducción de ciencias naturales que de matemáticas.
- Hay un porcentaje equivalente de ítems con errores en los ítems dirigidos a ambas poblaciones de estudiantes (9 y 13 años de edad).
- El mayor porcentaje de ítems con errores se presenta en la sección de ciencias naturales de estudiantes de 13 años de edad.

- Las diferencias más grandes de errores se encuentran en la dimensión de Cultura (currículo), donde se observa una mayor proporción de ítems con errores en la población de 9 años de edad (13% en matemáticas y 17% en ciencias naturales), que en la de 13 años (4% en matemáticas y 3% en ciencias naturales).

Tabla 6. Porcentajes de ítems analizados de la prueba TIMSS–1995 que presentan errores de traducción por dimensión, población y dominio curricular.

Dimensión de error	Población: 9 años		Población: 13 años	
	Matemáticas	Ciencias naturales	Matemáticas	Ciencias naturales
Estilo	17.0	28.4	21.1	29.7
Formato	75.0	76.5	82.9	83.8
Convenciones	21.6	37.0	11.8	35.1
Gramática	36.4	39.5	36.8	47.3
Semántica	89.8	85.2	82.9	87.8
Registro	18.2	9.9	17.	23.0
Información	69.3	64.2	44.7	60.8
Constructo	23.9	27.2	15.8	33.8
Cultura (currículo)	12.5	17.3	3.9	2.7
Origen	17.0	14.8	17.1	21.6

Fuente: Solano-Flores, Contreras-Niño y Backhoff, 2006

Como caso especial, encontramos que muchos ítems de matemáticas presentan un error que consiste en traducir dos oraciones como una sola oración, en modo gramatical condicional.
Un ejemplo típico de lo anterior sería el siguiente:

Ítem original: Daniel compra cuatro libretas que cuestan $23.50 cada una.
¿Cuánto dinero necesita?

Ítem traducido: ¿Cuánto dinero necesita Daniel si compra cuatro libretas que cuestan $23.50 pesos cada una?

Como ya se había mencionado anteriormente, para conocer el impacto negativo o positivo que ejercen los errores de traducción en las puntuaciones de las pruebas analizadas, se realizó un análisis de correlación entre los valores *p* de los reactivos (porcentaje de estudiantes que respondieron correctamente el reactivo) y el número de errores de traducción de los mismos. Hay que señalar que una correlación negativa alta sugiere que los problemas de traducción interfieren con (o desfavorecen) el desempeño de los estudiantes en los reactivos; es decir, a mayor número de errores de traducción, menor será el número de estudiantes que respondan correctamente el reactivo. Por el contrario, una correlación positiva implica que las fallas en la traducción facilitan (o favorecen) los resultados de los estudiantes en los ítems. Lo anterior es cierto si las correlaciones son estadísticamente significativas.

Dicho lo anterior, las tablas 7 y 8 muestran, respectivamente, las correlaciones de los reactivos para las poblaciones de estudiantes de 9 y 13 años. En la primera de estas dos figuras se observan correlaciones negativas para los ítems de matemáticas y positivas para los ítems de ciencias naturales. Para la población de 13 años se observa un patrón opuesto: correlaciones negativas para los ítems de ciencias naturales y positivas para los de matemáticas.

Tabla 7. Correlaciones (r) entre los valores *p* de los ítems y el número de errores de traducción de una muestra de reactivos de TIMSS–1995, para estudiantes de 9 años de edad

Áreas de contenido	k=	3° de primaria		4° de primaria	
		1995	2000*	1995	2000*
Matemáticas	42	-0.23	-0.26	-0.25	-0.33
Ciencias naturales	39	0.03	0.02	0.11	0.08
Combinadas	81	-0.12	-0.13	-0.08	-0.14

Fuente: Solano-Flores, Contreras-Niño y Backhoff, 2006

* Réplica de la prueba TIMSS-1995 en el año 2000. En negritas se señalan las correlaciones que resultaron estadísticamente significativas (p<0.05).

Tabla 8. Correlaciones (r) entre los valores *p* de los ítems y el número de errores de traducción de algunos reactivos de TIMSS–1995, para estudiantes de 13 años de edad

Áreas de contenido	k=	1° de secundaria		2° de secundaria	
		1995	2000*	1995	2000*
Matemáticas	19	0.16	0.13	0.10	0.13
Ciencias naturales	23	-0.25	-0.27	-0.19	-0.25
Combinadas	42	0.05	0.00	0.05	0.04

Fuente: Solano-Flores, Contreras-Niño y Backhoff, 2006

* Réplica de la prueba TIMSS-1995 en el año 2000. En negritas se señalan las correlaciones que resultaron estadísticamente significativas (p<0.05).

Sin embargo, es importante señalar que de todas estas correlaciones solamente una fue estadísticamente significativa: la de matemáticas en estudiantes de cuarto de primaria, en el año 2000. Aunque la gran mayoría de las correlaciones no son significativas, los patrones de las correlaciones son consistentes en ambos grados y en los dos años de aplicación de la prueba (1995 y 2000). Asimismo, en ambas poblaciones, las correlaciones negativas son consistentemente más altas que las correlaciones positivas.

Aunque por el momento, no es posible dar una explicación plausible a estos resultados, la tendencia de ellos nos dice que en algunos casos los errores de traducción no solamente afectan negativamente la ejecución de los estudiantes, sino que en otros la pudieran favorecer.

Estudio de PISA–2006

Un segundo estudio realizado más recientemente tuvo como objetivo analizar las características de la traducción mexicana al español de los ítems de la prueba PISA-2006. Dicho estudio se realizó a petición del INEE, que solicitó previamente autorización de la OCDE para analizar los reactivos no liberados de las pruebas de ciencias naturales, matemáticas y lectura.

Con este propósito, un comité de especialistas evaluó, en verano de 2008, la calidad de la traducción al español de todos los reactivos de ciencias naturales y de algunos de matemáticas de la prueba PISA-2006. Posteriormente, con esta información se analizó la manera en que el error de traducción de estos ítems se relaciona con en el desempeño de los estudiantes mexicanos en lo que PISA denomina "competencias para la vida". Los resultados del estudio se dieron a conocer al INEE en un texto denominado *Informe de resultados iníciales del análisis de la traducción de la prueba PISA-2006* (Contreras-Niño, Solano-Flores y Backhoff, 2008).

Comité revisor. El comité revisor de la traducción de esta prueba quedó conformado por los siguientes participantes: un lingüista; dos traductores certificados (inglés–español y francés–español); un psicómetra; un evaluador educativo (con especialidad en sesgo cultural), y; seis docentes en ejercicio especialistas en español, matemáticas y ciencias naturales (dos de cada asignatura, uno de educación secundaria y otro de bachillerato). El comité evaluador trabajó por 8 días completos, equivalente a 70 horas efectivas de trabajo, quien pudo revisar 101 ítems de ciencias naturales (la totalidad) y 37 de matemáticas (77%).

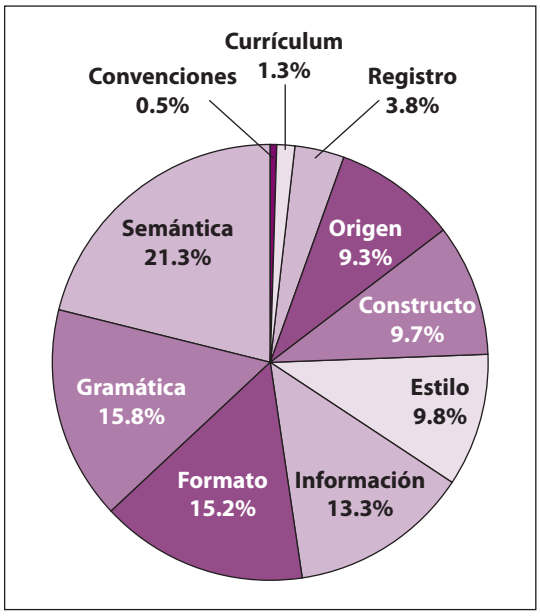
Tabla 9. Frecuencia y porcentaje de errores de traducción por dimensión, de 193 unidades analíticas evaluadas de PISA-2006

Dimensión	Número de errores	Porcentaje de errores
Estilo	107	9.8
Formato	165	15.2
Convenciones	5	0.5
Gramática	145	13.3
Semántica	172	15.8
Registro	232	21.3
Información	106	9.7
Constructo	41	3.8
Cultura (currículo)	101	1.3
Origen	100	9.3
Total	1088	100

Unidades analíticas. Como ya se había comentado en un apartado anterior, en PISA los ítems están organizados en familias; es decir, grupos de dos a cuatro reactivos que se relacionan con un párrafo introductorio común. La introducción presenta información contextual que el estudiante tiene que leer y entender para resolver los problemas planteados por los ítems. Por tal razón, la revisión de la traducción de PISA-2006 consideró como unidades analíticas tanto a las introducciones como a los ítems. En total, se revisó la traducción de 193 unidades analíticas.

Resultados. En las 193 unidades analíticas que se revisaron, de ciencias naturales y matemáticas, se identificaron un total de 1088 errores, distribuidos en diez dimensiones. Como se puede observar en la tabla 10 y en la figura 4, la mayoría de errores de traducción de PISA-2006 pertenecen a la dimensión de Semántica (21.3%), seguida a la de Gramática (15.8%), Formato (15.2%) e Información (13.3%).

Figura 4. Porcentaje de errores de traducción por dimensión de 193 unidades analíticas de PISA-2006



Por su parte, la menor proporción de errores detectados pertenecen a las dimensiones de Convenciones (0.5%), Currículum (1.3%) y Registro (3.8%). Es importante señalar que la dimensión de Convenciones se refiere principalmente a las normas por aplicar a los ítems de opción múltiple. Como en PISA este tipo de reactivo no es el más utilizado, es entendible la baja frecuencia de errores en esta dimensión.

Si consideramos solamente a las dimensiones que son las más críticas para la validez de los ítems —Constructo, Registro, Currículo y Origen—, encontramos que, entre estas cuatro, el número de errores suman 494, equivalente al 45% del total de errores. Esta cantidad de errores es importante de considerar, dado que pueden afectar seriamente las interpretaciones de los resultados de PISA, al menos para el caso de México. Sin embargo, debe recordarse que la presencia de uno o varios errores de traducción en un mismo ítem no necesariamente afecta al desempeño de los estudiantes, a menos que dichos errores sean muchos o muy graves.

Tabla 10. Frecuencia de errores de traducción por categoría y dimensión de error de traducción

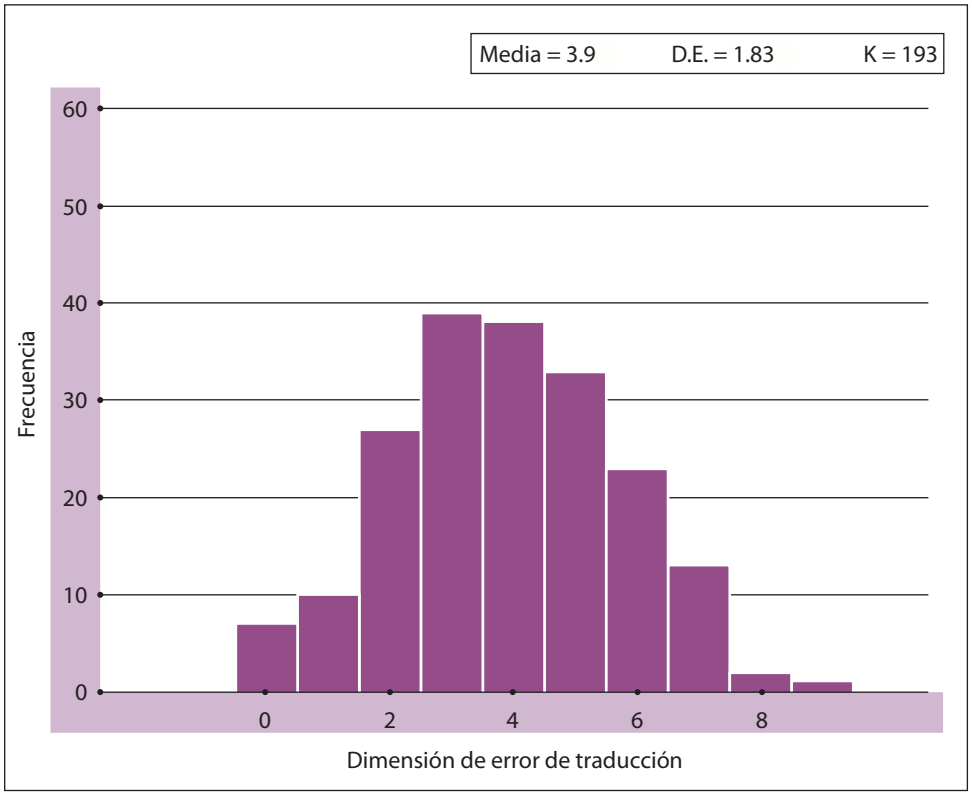
Dimensión	Categoría	Frecuencia de error	
		Categoría	Dimensión
Estilo	DE1: Errores de puntuación	87	107
	DE2: Uso impropio de mayúsculas	14	
	DE3: Uso impropio de minúsculas	1	
	DE4: Errores de ortografía	5	
	DE5: Otros	0	
Formato	DF1: Cambio en el tamaño, estilo o posición de tablas, gráficas o ilustraciones	15	165
	DF2: Cambio en el estilo, justificación o tamaño de fuentes o caracteres	40	
	DF3: Márgenes más amplios o más reducidos	5	
	DF4: Omisión de componentes gráficos	27	
	DF5: Inserción de componentes gráficos	64	
	DF6: Otros	14	
Convenciones	DC1: Inconsistencia gramatical entre la base del ítem y las opciones de respuesta	0	5
	DC2: Uso inapropiado de puntos suspensivos para denotar continuidad entre la base y las opciones	0	
	DC3: Cambio en el orden de las opciones; inconsistencia gramatical entre las opciones	1	
	DC4: Uso inapropiado de mayúsculas al principio de las opciones	0	
	DC5: Otros	4	
Información	CI1: Traducción inconsistente de un término no técnico que se repite varias veces en el original	33	145
	CI2: Un término no técnico importante aparece más o menos veces que en el original	1	
	CI3: Cambio en la forma de escribir números	1	
	CI4: Inserción de términos no técnicos, oraciones o explicaciones que no aparecen en el original	51	
	CI5: Omisión de palabras importantes, términos no técnicos u oraciones o explicaciones que sí aparecen en el original	53	
	CI6: Otros	6	
Gramática	LG1: Traducción literal (palabra por palabra)	51	172
	LG2: Estructura sintáctica no natural	54	
	LG3: Inconsistencia sujeto-verbo	3	
	LG4: Inconsistencia de singulares y plurales	2	
	LG5: Uso inapropiado de preposiciones	42	
	LG6: Uso inapropiado en la concordancia de tiempos	8	
	LG7: Conflación de dos oraciones en una oración	5	
	LG8: Otros	7	
Semántica	LS1: Uso de cognados falsos	1	232
	LS2: Traducción impropia de expresiones idiomáticas	2	
	LS3: Alteración en el sentido de una oración	16	
	LS4: Una oración se puede interpretar de más de una manera	12	
	LS5: No es claro lo que el ítem pide que se haga	0	
	LS6: Cambio en el género de personajes	2	
	LS7: Combinación de dos o más enunciados en uno	5	
	LS8: Uso impreciso de términos y vocablos	26	
	LS9: Uso de términos con significados múltiples	25	
	LS10: Traducción inapropiada de una palabra	143	
	LS11: Otros	0	

Dimensión	Categoría	Frecuencia de error	
		Categoría	Dimensión
Constructo	CC1: Posible alteración de las demandas cognoscitivas del ítem	25	106
	CC2: Posible alteración de la forma en que se puede interpretar el ítem	29	
	CC3: Uso impreciso de términos técnicos	2	
	CC4: Traducción inconsistente de un término técnico que se repite varias veces en el original	12	
	CC5: Inserción de términos técnicos	1	
	CC6: Omisión de términos técnicos	8	
	CC7: Sustitución de un término técnico por un término no técnico	22	
	CC8: Sustitución de un término no técnico por un término técnico	7	
	CC9: Otros	0	
Registro	LR1: Uso de palabras de baja frecuencia entre la población destinataria	1	41
	LR2: Traducción incorrecta de términos técnicos	40	
	LR3: Traducción correcta de términos técnicos pero de una manera que no se enseña en el país	0	
	LR4: Otros	0	
Cultura	CQ1: La información contextual y las situaciones que el ítem proporciona son poco comunes en la cultura nacional	5	14
	CQ2: Las unidades de medición son ajenas a la cultura nacional	0	
	CQ3: El tipo de problema planteado por el ítem no tiene sentido en la cultura nacional	2	
	CQ4: El ítem evalúa conocimientos que no se enseñan en el país	4	
	CQ5: Otros	3	
Origen	OO1: Hay más de una respuesta correcta	1	101
	OO2: Ninguna de las opciones es completamente correcta	0	
	OO3: Inconsistencia entre las dos versiones fuente	46	
	OO4: Otros	54	

La tabla 10 muestra en forma desagregada los errores de traducción por categoría de error (y por dimensión). Se puede apreciar que el mayor número de errores de traducción de los ítems de PISA-2006 pertenece a la dimensión de Semántica en la categoría de *Traducción inapropiada de una palabra*. Los 143 errores identificados en esta categoría equivalen a 13% del total de errores. Este dato nos indica que la versión mexicana de PISA contiene una cantidad importante de términos cuya traducción alteró el significado original de los reactivos (aunque no necesariamente el desempeño de los estudiantes en la prueba).

Sin considerar la categoría *Otros*, las diez categorías con mayor número de errores de traducción, después de la Traducción inapropiada de una palabra, fueron las siguientes: Errores de puntuación (87), de la dimensión Estilo; Inserción de componentes gráficos (64), de la dimensión Formato; Estructura sintáctica no natural (54), de la dimensión Gramática; Omisión de palabras importantes, términos no técnicos u oraciones o explicaciones que sí aparece en el original (53), de la dimensión Información; Inserción de términos no técnicos, oraciones o explicaciones que no aparecen en el original (51), de la dimensión Información; Traducción literal (palabra por palabra) (51), de la dimensión Gramática; Inconsistencia entre las dos versiones fuente (46), de la dimensión Origen; Uso inapropiado de preposiciones (42), de la dimensión Gramática; Cambio en el estilo, justificación o tamaño de fuentes o caracteres (40), de la dimensión Formato, y; Traducción incorrecta de términos técnicos (40), de la dimensión Registro.

Figura 5. Distribución de frecuencias de unidades analíticas con dimensiones de error



La experiencia en el análisis de la traducción de pruebas indica que los patrones de error de traducción se aprecian mejor en el nivel de dimensión que en el nivel de categoría. En publicaciones relacionadas con el análisis de error de traducción en pruebas internacionales, los investigadores y autores de este informe han empleado una medida gruesa del error de traducción: el número de diferentes dimensiones de error en los cuales se detecta error de traducción en un ítem. Este número no expresa en cuántas categorías se observa error para una dimensión determinada, ni cuantas veces. Sin embargo, a pesar de su naturaleza molar, esta medida permite apreciar el estado de la calidad de la traducción del conjunto de ítems analizados.

De acuerdo con tal experiencia, se calculó el número de dimensiones diferentes en las que se observó error para cada unidad analítica y se construyó la distribución de frecuencias que se presenta en la figura 5, la cual muestra una distribución normal con un promedio de 3.9 dimensiones de error de traducción y 1.8 de desviación estándar.

Ahora bien, en la tabla 11 se puede apreciar que de las 197 unidades analíticas evaluadas, más de 76% de ellas presentó algún error de traducción de Semántica, y poco más de 50% tuvo al menos algún error de Formato, Información o Gramática.

Tabla 11. Unidades analíticas de PISA-2006
con algún error de traducción, por dimensión

Dimensión	Número de errores	Porcentaje de errores
Estilo	92	46.7
Formato	103	53.4
Convenciones	5	2.6
Gramática	102	52.8
Semántica	103	53.4
Registro	150	77.7
Información	68	35.2
Constructo	40	20.7
Cultura (currículo)	10	5.2
Origen	79	40.9

Nota: el número total de unidades analíticas es de 193

Como dato complementario, aunque no derivado de la tabla anterior, se encontró que el 96.4% de las unidades analíticas presentó algún error de traducción. Es decir, solo siete de los 193 ítems y sus introducciones no presentaron errores de traducción al español. Por su importancia, hay que mencionar también la gran cantidad y diversidad de errores de traducción de la dimensión *Origen*, en especial aquellos identificados en la categoría Otros. Como se recordará, los errores de *Origen* son errores conceptuales o de diseño de los ítems en las versiones originales (en este caso, francés e inglés). Los resultados muestran que de los 101 errores en la dimensión *Origen*, 54 se clasificaron en la categoría *Otros*, es decir, un poco más de la mitad.

Con el propósito de ejemplificar lo anterior, la tabla 12 muestra 34 ítems de ciencias naturales y dos de matemáticas que presentan este tipo de error de traducción. Se puede apreciar que en 22 ítems el problema radica en que la base del reactivo es confusa. En la mayoría de estos casos, lo que el ítem le pide al estudiante atender, hacer o responder, se encuentra disperso en varios apartados, en algunas ocasiones re-fraseando parte de esta información, lo cual hace que el reactivo se vuelva innecesariamente complejo.

En la tabla 12 se puede identificar el siguiente agrupamiento de errores de *Origen* de la categoría *Otros*:

- Cuatro ítems, además del problema en que la base del reactivo es confusa y compleja, presentan otro problema de la misma categoría.
- En nueve ítems no se respetan las convenciones de escritura técnica de los reactivos.
- En seis ítems hay errores conceptuales.
- Dos ítems, que tuvieron errores de estructura sintáctica, fueron considerados como ítems con errores graves.

Tabla 12. Ejemplos de errores de *Origen* de la categoría Otros de los ítems de PISA-2006

Ítem	Error documentado
S508Q02	La base del reactivo se confunde con las instrucciones para responder
S426Q03	La base del ítem es confusa y compleja
S114	Falta de precisión en las variaciones de temperatura (segundo renglón del segundo párrafo) introduce ambigüedad; igual confunde el último párrafo
S521Q02	Inconsistencia gramatical entre la base y las opciones; también repite información innecesaria en las opciones
S495Q02	La base del ítem es confusa y compleja
S456Q01	La base del reactivo es confusa; además, no es lo mismo explicar que ayudar a explicar
S477Q02	Distractores no plausibles
S268Q06	El distractor D no es plausible
S519Q02	La base del ítem es confusa y compleja
S498Q02	La base del ítem es confusa y compleja
S524Q06	La base del ítem es confusa y compleja. Inconsistencia en punto final de enunciado en tablas
S510Q01	La base del ítem es confusa y compleja. No existe una lista como se indica en la base del reactivo
S326	Hubiera sido más riguroso decir que la leche es el primer alimento de los mamíferos tras su nacimiento La unidad de medida gramos debe estar colocada en grasas, proteínas y carbohidratos, no en los tipos de leche, pues ya se dijo que son 120 gramos de cada tipo de leche. Ello es incorrecto y confunde al examinado
S326Q03	La base del ítem es confusa y compleja
S408	Error conceptual al definir *Weed*. Falsa explicación de origen del nombre de la avena
S408Q01	Opción C no es plausible; es absurda
S408Q04	La base del ítem es confusa y compleja
S408Q05	La respuesta correcta (A) da pistas para el ítem 9, pues si las semillas germinan en la mezcla estiércol-tierra, la segunda opción en el ítem 9 es cierta
S437	El fuego es el resultado de la reacción química denominada combustión, no es un tipo de reacción química (falta de rigor conceptual)
S415Q01	La base del ítem es confusa y compleja
S415Q08	La base del ítem es confusa y compleja
S478Q02	La base del ítem es confusa y compleja
S478Q03	La base del ítem es confusa y compleja. El segundo argumento no es plausible
S447	Es confusa la descripción del procedimiento, en general. Por ejemplo, Diego pone una gota de cada sustancia dentro de un círculo; en vez de en cada círculo
S413Q04	La base del ítem es confusa y compleja
S458Q02	La base del ítem es confusa y compleja
S438Q01	La base del ítem es confusa y compleja
S438	No se tiene certeza del significado del signo "+" en la tabla
S438Q02	Se sospecha que se trata de un conocimiento metódico muy complejo para estudiantes de esta edad
S466Q01	La base del ítem es confusa y compleja
S466Q07	La base del ítem es confusa y compleja
S493Q01	La base del ítem es confusa y compleja
S493Q03	La base del ítem es confusa y compleja
S514Q04	La estructura del ítem es confusa y compleja, cuando se trata de un simple ítem de relación de columnas
M442Q02	La base del ítem es confusa y compleja
M273Q01	No se le indica al examinado que debe responder poniendo las letras o los dibujos directamente

El hecho de que los errores de origen se puedan detectar durante el análisis de la traducción, confirma la observación sobre el análisis de ítems, desde un punto de vista comparativo entre idiomas, el cual revela errores que no se detectan de otra manera. (Solano-Flores, Trumbull, y Nelson-Barber, 2002).

Por otro lado, a fin de tener una apreciación adicional de la calidad de la traducción de los ítems de PISA–2006, se hizo un conteo de la frecuencia con la cual los errores de traducción de las unidades analíticas fueron clasificados por los revisores como pocos leves, muchos leves, pocos graves, o muchos graves. Es importante decir que esta clasificación no necesariamente es excluyente, toda vez que un reactivo o introducción pueden tener errores leves (pocos o muchos) y/o errores graves (pocos o muchos). Los resultados de este conteo se presentan en la tabla 13, en donde se puede apreciar que los errores con la capacidad de atentar contra la validez de los reactivos caen en tres categorías: muchos errores leves (44 casos), pocos errores pero graves (35 casos), y/o muchos errores graves (9 casos). En una o más de estas tres condiciones se observaron 98 unidades analíticas de PISA–2006.

Tabla 13. Número de unidades analíticas de acuerdo con las categorías de error de traducción

Pocos leves	Muchos leves	Pocos graves	Muchos graves	Total
136	44	35	9	224

Nota: el total es diferente del número de unidades analíticas revisadas (193) porque las categorías no son excluyentes.

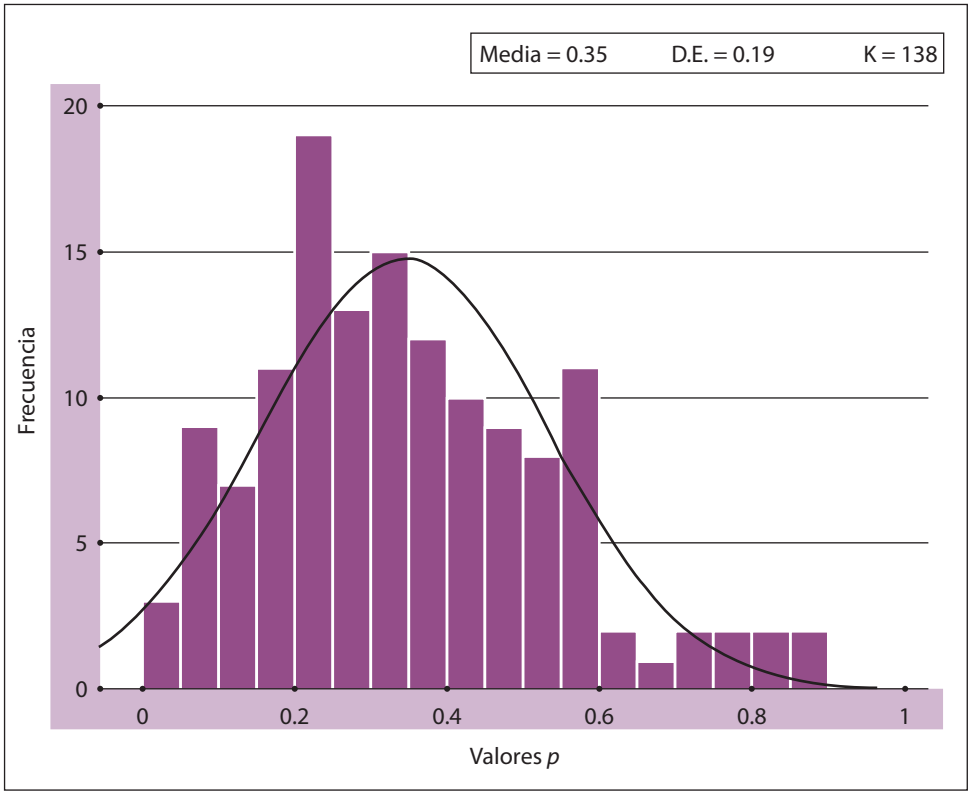
Finalmente, para conocer el impacto que pueden tener los errores de traducción en la ejecución de los estudiantes, se correlacionaron los errores de traducción de los 138 reactivos analizados de PISA–2006 con su índice de dificultad o valor *p* (proporción de estudiantes que respondieron correctamente el ítem).

La figura 6 muestra la frecuencia de valores *p* de los 138 reactivos analizados. Se puede apreciar que su media de dificultad es de 0.35, su desviación estándar de 0.19, y su distribución tiene un sesgo positivo; es decir, hay una cantidad mayor de reactivos de alta dificultad (o con valores *p* bajos).

Para conocer el efecto de los errores de traducción en el desempeño de los alumnos se correlacionaron el número de dimensiones de error y los valores *p* de los ítems, en las siguientes condiciones:

- Todos los ítems (sin desagregar) y 10 dimensiones de error.
- Todos los ítems (sin desagregar) y 6 dimensiones de error (las más importantes).
- Ítems desagregados por área de contenido (ciencias y matemáticas).
- Ítems desagregados en dos categorías (objetables y no objetables).

Figura 6. Distribución de valores *p* (aciertos)
en los 138 reactivos analizados de PISA-2006



La tabla 13 muestra las correlaciones entre los errores de traducción de los ítems y sus valores *p*, de acuerdo con diferentes condiciones de análisis. Se podrá observar que todas las correlaciones fueron negativas, lo que indica que hay una tendencia o relación inversa entre la dificultad del ítem y el número de dimensiones de error; es decir, los ítems se vuelven más difíciles conforme tienen un mayor número de dimensiones de error. Sin embargo, de las seis condiciones en que se analizó esta relación, solamente en una de ellas resultó ser estadísticamente significativa y considerablemente alta (de -0.40): la condición en que solo se correlacionaron los ítems objetables en las seis dimensiones de error de mayor importancia.

Tabla 13. Correlaciones entre los valores *p* y el número
de dimensiones de error de los ítems

Condiciones	K	r
Todos los ítems y diez dimensiones de error	138	-0.06
Todos los ítems y seis dimensiones de error	138	-0.12
Ítems de Ciencias y seis dimensiones de error	101	-0.12
Ítems de Matemáticas y seis dimensiones de error	37	-0.21
Ítems no-objetables y seis dimensiones de error	112	-0.08
Ítems objetables y seis dimensiones de error	26	**-0.40**

En negritas se señalan las correlaciones que resultaron estadísticamente significativas ($p<0.05$).

Los resultados anteriores aportan evidencia de validez del modelo que propone la TETP ya que, como se sugiere en el espacio probabilístico de los ítems aceptables y objetables (ver figura 1), el número de errores de traducción de los ítems no es el problema que potencialmente puede afectar más la ejecución de los estudiantes, sino la severidad de los errores, misma que logra marcar una diferencia entre los ítems clasificados aceptables y objetables.

Conclusiones y recomendaciones

Las comparaciones internacionales han mostrado que una traducción ligeramente inexacta de una palabra puede ser suficiente para afectar al funcionamiento diferencial de un ítem (Ercikan, 1998). Debido a la sensibilidad del lenguaje en pruebas de logro educativo, los estudios de PISA y TIMMS emplean ahora dos idiomas fuente como una estrategia orientada a asegurar que se conserve el significado en la traducción.

Procedimientos recientemente utilizados en estudios comparativos, como el uso del procedimiento de *traducción inversa* —donde la traducción de un idioma fuente a un destinatario tiene que ser traducido nuevamente del destinatario al fuente— ya no se acepta como prueba irrefutable de una traducción de pruebas. Esto debido a que se siguen buscando y mejorando procedimientos que garanticen la calidad de las traducciones a diversos idiomas de los instrumentos que se utilizan para comparar la ejecución de los estudiantes de diversos países.

No obstante el conocimiento del impacto que puede causar una traducción defectuosa en los resultados de las pruebas internacionales, el procedimiento para su revisión no ha recibido una atención suficiente. Por ejemplo, los aspectos culturales asociados al idioma no siempre se consideran de manera formal y, comúnmente, se supone que trabajar con traductores familiarizados con la cultura de la población destinataria es condición suficiente para atender los principales aspectos del uso del idioma.

Asimismo, no es posible identificar errores de traducción relacionados con los aspectos sociales de uso del idioma, si no se cuenta con un sistema detallado de codificación de errores y con un equipo multidisciplinario que lo ponga en práctica.

Lo anterior resulta evidente si tomamos en cuenta que el reporte sobre la calidad de los datos de TIMSS-1995 incluye un capítulo sobre los procedimientos empleados para revisar la calidad de las traducciones de la prueba. La versión mexicana de TIMSS-1995 pasó por este proceso de certificación basado en las normas utilizadas por la oficina del TIMSS, antes de ser aceptada para su uso con estudiantes mexicanos. Según dicho informe solo uno de los ítems utilizados por México había sido identificado como problemático (Mullis, Kelly, & Haley, 1996)

No obstante, en los estudios que hemos realizado con dos de las pruebas de logro educativo de mayor influencia en el mundo (TIMSS y PISA) se han observado errores de traducción en una gran parte de sus reactivos, algunos de los cuales son clasificados como errores graves. El error de traducción por sí mismo podría ser inocuo si éste no afectara la complejidad cognitiva de las pruebas o su contenido, pero hemos observado que los errores de traducción considerados como graves tienen un impacto negativo en la dificultad del reactivo, lo cual puede poner en riesgo la validez de la medición.

También hemos encontrado que el efecto de los errores de traducción sobre el desempeño de los estudiantes es más evidente en Matemáticas que en Ciencias Naturales, aunque no hemos hecho la comparación para el caso de lectura. Igualmente, es más evidente cuando se consideran solamente las seis dimensiones de error de traducción de contenido (Información, Gramática, Semántica, Constructo, Registro y Cultura), que cuando se consideran las diez dimensiones que estipula la TETP.

No se pone en duda de que los equipos técnicos de los países participantes en las evaluaciones internacionales, como es el caso de México, traten de seguir con fidelidad los lineamientos establecidos por los organismos internacionales que coordinan los estudios comparativos de educación (por ejemplo, Hambleton, 1996; 2005). Sin embargo, resulta obvio que estos lineamientos no cubren todas las necesidades de traducción que garanticen una trabajo de calidad.

Aunque los lineamientos para la traducción de pruebas son claros (pero generales), es necesario operacionalizar esos lineamientos e implementar procedimientos rigurosos de revisión de traducción. La metodología aquí descrita pudiera emplearse como parte del proceso de traducción.

Para terminar, queremos reiterar cuatro recomendaciones a los países iberoamericanos, con el propósito de mejorar sus procesos de traducción de pruebas (véase Solano-Flores, Contreras-Niño y Backhoff, 2006):

1. Asignar el personal calificado necesario para un proceso apropiado de traducción. Este personal va más allá de traductores certificados.

2. Asegurar que las actividades de revisión sean parte esencial de todas las etapas del proceso de traducción de pruebas, no simplemente la fase final.

3. Asignar suficiente tiempo para que haya varias iteraciones de traducción y revisión de la traducción de las pruebas.

4. Asegurar que los equipos de traducción de pruebas y de la revisión correspondiente documenten todas sus acciones y justifiquen las decisiones que se tomen.

Finalmente, debemos mencionar dos acciones que deberán realizarse en el futuro con el propósito de enriquecer y aprovechar la información que la TETP proporciona sobre las traducciones de pruebas:

1. Diseñar un protocolo en formato electrónico que permita capturar el contenido de las discusiones de los comités revisores, de manera que no siempre sea necesario efectuar una transcripción de los juicios emitidos por los especialistas del comité.

2. Realizar un estudio de validez cognitiva de los ítems traducidos de la prueba PISA-2006, para conocer la manera en que los errores de traducción impactan los procesos de pensamiento de los estudiantes al responder los ítems.

Bibliografía

ASTM International. (2006). F2575-06 Standard Guide for Quality Assurance in Translation. Consultado el 5 de mayo de 2011, en http://www.astm.org/Standards/F2575.htm

Bullinger, M. (2003). International comparability of health interview surveys: An overview of methods and approaches. In A. Nosikov and C. Gudex (Eds.), *EUROHIS: Developing Common Instruments for Health Surveys*. Amsterdam: IOS Press.

Contreras-Niño, L. A., Solano-Flores, G. y Backhoff, E. (2009). *Informe de resultados iniciales del análisis de la traducción de la prueba PISA-2006*. México: Universidad Autónoma de Baja California, Universidad de Colorado en Boulder e Instituto Nacional para la Evaluación de la Educación.

Davis, A. (1991). The language of testing. In. K. Durkin & B. Shire (Eds.), *Language in mathematical education: Research and practice* (pp. 40–47). Buckingham, UK: Open University Press.

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

Ferguson, A. M., & Fairburn, J. (1985). Language experience for problem solving in mathematics. *Reading Teacher*, 38, 504–507.

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

Grisay, A. (2002). Translation and cultural appropriateness of the test and survey material. En R. Adams y M. Wu (Eds.), *PISA 2000 Technical Report* (pp. 42-54). Paris: Organisation for Economic Co-operation and Development

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

Halman, L. (2001). The European Values Study: A Third Wave. Source Book of the 1999/2000 European Values Study Surveys. Tilburg: EVS, WORC, Tilburg University.

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. (2005). Issues, designs, and technical guidelines for adapting tests into multiple languages and cultures. En R. K. Hambleton, P. Merenda y C. D. Spielberger (Eds.), *Adapting educational and psychological tests for cross-cultural assessment* (pp. 3-38). Mahwah, NJ: Erlbaum.

Hambleton, R. K. (1996, March). Guidelines for adapting educational and psychological tests. Paper presented at the meeting of the National Council of Measurement in Education, New York, NY.

Harkness, J. (2003). Questionnaire translation. In J. Harkness, F. J. R. Van de Vijver and P. Mohler (Eds.), *Cross-Cultural Survey Methods*. New Jersey: John Wiley & Sons, Inc.

Maxizip, S/F. Back translation as a means of giving translators a voice. http://maxizip.com/2010/12/what-is-back-translation-definition-back-translation/

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

Muñiz, J., Hambleton, R. K., & Xing, D. (2001). Small sample studies to detect flaws in item translations. *International Journal of Testing*, 1(2), 115–135.

O'Connor, K. M. y Malak, B. (2000). Translation and cultural adaptation of the TIMSS instruments. En M. O. Martin, K. D. Gregory y S. E. Stemler (Eds.), *TIMSS 1999 Technical Report* (pp. 89-100). Chestnut Hill, MA: International Study Center, Boston College.

Ozolins, U. (2008). Issues of back translation methodology in medical translations. Ponencia presentada en FIT [International Federation of Translators] XVIII Congress, Shanghai.

Solano-Flores, G. (2006). Language, dialect, and register: Sociolinguistics and the estimation of measurement error in the testing of English-language learners. *Teachers College Record*, 108(11), 2354–2379.

Solano-Flores, G. (2008). Who is given tests in what language by whom, when, and where? The need for probabilistic views of language in the testing of English language learners. *Educational Researcher*, 37(4), 189-199.

Solano-Flores, G. (In press). Language issues in mathematics and the assessment of English language learners. In K. Tellez & J. Moschkovich (Eds.), *Latinos and Mathematics: Research on Learning and Teaching in Classrooms and Communities*.

Solano-Flores, G. y Backhoff, E. (2003). La traducción de pruebas en las comparaciones internacionales: un estudio preliminar (Informe técnico para el Instituto Nacional para la Evaluación de la Educación). México, D. F.: Instituto Nacional para la Evaluación de la Educación.

Solano-Flores, G., Backhoff, E. y Contreras-Niño, L. A. (2009). Theory of Test Translation Error. *Internacional Journal of Testing*. 9(2), 78-91.

Solano-Flores, G., Contreras-Niño, L. A. y Backhoff-Escudero, E. (2005, 12-14 de abril). The Mexican translation of TIMSS-95: Test translation lessons from a post-mortem study. Trabajo presentado en la Annual Meeting of the National Council on Measurement in Education. Montreal, Quebec, Canada.

Solano-Flores, G., Contreras-Niño, L. A., & Backhoff, E. (2006). Traducción y adaptación de pruebas: Lecciones aprendidas y recomendaciones para países participantes en TIMSS, PISA y otras comparaciones internacionales. REDIE: *Revista Mexicana de Investigación Educativa*, 8(2). [Sp.] http://redie.uabc.mx/vol8no2/contents-solano2.html

Solano-Flores, G., & Kidron, Y. (2006, April). Formal and judgmental approaches in the analysis of test item linguistic complexity: A comparative study. Paper presented at the meeting of the American Educational Research Association, San Francisco, California.

Solano-Flores, G., Trumbull, E. y 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.

Valdés, G., & Figueroa, R. A. (1994). *Bilingualism and testing: A special case of bias*. Norwood, NJ: Ablex.

Queremos manifestar nuestro agradecimiento a las siguientes personas por el apoyo moral, financiero e intelectual que nos brindaron para desarrollar, evaluar y difundir la Teoría del Error de Traducción de Pruebas (TETP), objeto de este trabajo. Dado que la TETP se ha desarrollado principalmente a partir de diversos y grandes proyectos, en nuestros agradecimientos nos referimos en lo particular a quienes nos han ayudado en cada uno de ellos.

Respecto al proyecto para la revisión de la traducción de los ítems de las pruebas TIMSS–1995 y PISA–2006, en primer lugar deseamos agradecer a Felipe Martínez Rizo que, siendo Director General del INEE, siempre mostró simpatía por el proyecto y confianza en sus autores, razón por la cual: 1) consiguió la autorización de la OCDE para poder analizar los reactivos no liberados de PISA–2006 y 2) consiguió el financiamiento para realizar dos estudios de la traducción de las pruebas internacionales antes mencionadas.

Reconocemos y agradecemos el apoyo moral que nos brindó Annette Santos del Real, Directora General Adjunta del INEE, quien siempre nos alentó a desarrollar las investigaciones aquí reportadas y quien nos brindó el apoyo para realizar el trabajo de análisis de los ítems de PISA-2006 en las instalaciones del INEE, en la primavera de 2008. Asimismo a María Antonieta Díaz Gutiérrez, Directora de Proyectos Internacionales y Especiales del INEE, quien nos facilitó las tres versiones de los reactivos de PISA-2006: inglés, español y francés, sin los cuales no hubiera sido posible realizar una parte importante de este trabajo.

A Margarita Zorrilla Fierro, actual Directora General del INEE, a José Fernando González Sánchez, Subsecretario de Educación Básica de la SEP, y a Graciela Cordero Arroyo, Directora del Instituto de Investigación y Desarrollo Educativo, de la UABC, deseamos manifestarles nuestro agradecimiento por habernos brindado su apoyo financiero, político y logístico para realizar en la ciudad de México, en febrero de 2010, el Taller Iberoamericano sobre la Teoría del Error de Traducción de Pruebas Internacionales, en donde diseminamos entre colegas latinoamericanos los principios y la metodología de la teoría.

En cuanto al proyecto para la revisión de la traducción de los ítems de la prueba TIMSS–1995 deseamos hacer patente nuestro agradecimiento a las siguientes personas:

Avelina Martínez, traductora inglés – español,
Guadalupe López Bonilla, lingüista,
Jeanette Cortes Flores, profesora de secundaria,
Mayra Terán Licona, profesora de secundaria,
María Martha Alvarado Murillo, profesora de secundaria,
Ramón Ibarra López, profesor de secundaria,
María Engracia Morales Molinar, profesora de secundaria,
Jaime Alberto Chávez Salas, profesor de primaria,
Sandra Edith Guerra Trejo, profesora de primaria,
Mónica López Ortega, capturista de bases de datos,
Sofía Contreras Roldán, capturista de bases de datos,
Flor Magaña Oviedo, operadora de video grabación, y
Angélica Ceseña Ojeda, gestión administrativa.

Igualmente, agradecemos al sinnúmero de aportaciones técnicas e intelectuales a todas las personas que nos han ayudado a realizar el análisis de los reactivos PISA–2006:

Georgina Barraza Carbajal, lingüista,
María de Jesús Chaparro Velasco, traductora francés – español,
Paola Núñez Acevedo, traductora inglés – español,
Esperanza Minerva Guevara, profesora de ciencias naturales,
Edgar Andrade Muñoz, analista de sistemas,
José Luis Ramírez Cuevas, programador,
Kiyoko Nishikawa Aceves, capturista de bases de datos,
Sobrina de Willy, capturista de bases de datos,
Nelly Del Carmen Pulido Páez, transcriptora,
Anahí Ornelas Ley, transcriptora, y
María Angélica Ceseña Ojeda, gestión administrativa.

Finalmente, respecto al proyecto para la revisión de la traducción de los ítems liberados de la prueba PISA–2003, queremos reconocer las contribuciones de:

Paola María Núñez Acevedo, traductora inglés – español,
María de Jesús Chaparro Velasco, traductora francés – español,
Gabriela Preciado Almonte, lingüista,
Ana Miramontes Bush, profesora de bachillerato,
Luís Álvarez Aldaco, profesor de bachillerato,
Virginia Vargas Bautista, profesora de secundaria,
Carlos Morales Saldaña, profesor secundaria,
Edgar Andrade Muñoz, analista de sistemas,
Flor Magaña Oviedo, operadora de video grabación, y
Angélica Ceseña Ojeda, gestión administrativa.

Anticipadamente, pedimos una disculpa a quienes nos ayudaron en estos trabajos y que, por razones ajenas a nuestra voluntad, no mencionamos.

Anexos

ANEXO 1. Ejemplo de familia de ítems de PISA-2006

INVERNADERO

Lee el texto a continuación y responde las preguntas que aparecen después.

EL EFECTO INVERNADERO: ¿REALIDAD O FICCIÓN?

Unidad analítica: Párrafo introductorio

Los seres vivos necesitan energía para sobrevivir. La energía que mantiene la vida en la Tierra viene del Sol, que irradia esta energía al espacio debido a su alta temperatura. Una pequeñísima porción de esta energía llega a la Tierra.

La atmósfera de la Tierra actúa como una cobija protectora sobre la superficie de nuestro planeta, impidiendo los cambios de temperatura que existirían en un mundo sin aire.

La mayor parte de la energía irradiada que llega del Sol pasa por la atmósfera de la Tierra. La Tierra absorbe parte de esta energía y parte la refleja de regreso desde su superficie. Parte de esta energía reflejada la absorbe la atmósfera.

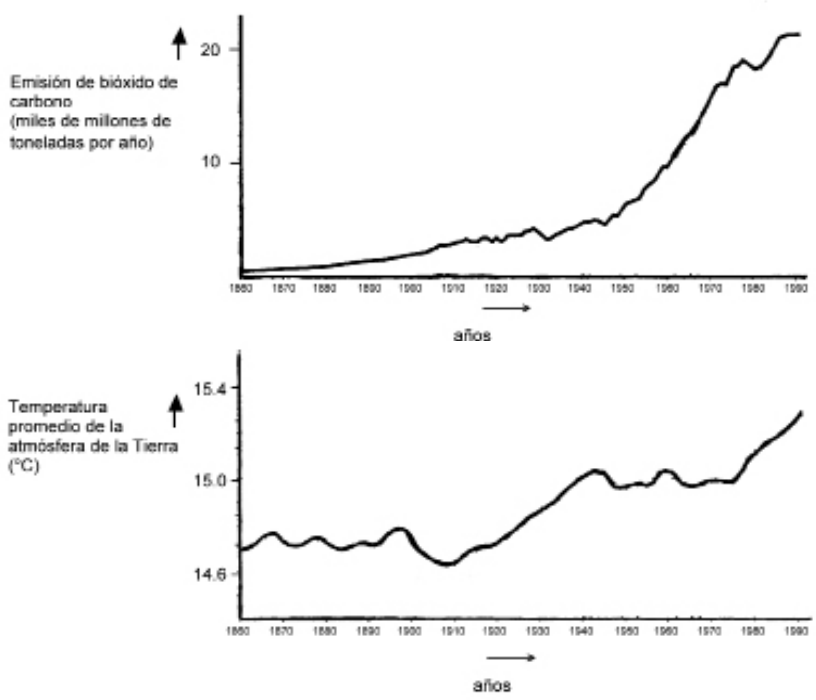
Como resultado de esto, la temperatura promedio de la superficie de la Tierra es más alta de lo que sería si no hubiera atmósfera. La atmósfera de la Tierra tiene el mismo efecto que un invernadero, de ahí el término efecto invernadero.

Se dice que el efecto invernadero se volvió más pronunciado durante el siglo XX.

Es un hecho que la temperatura promedio de la atmósfera terrestre ha aumentado. En los periódicos y revistas se afirma a menudo que la fuente principal del incremento de la temperatura durante el siglo XX es el aumento de las emisiones de bióxido de carbono.

Un estudiante llamado Andrés está interesado en comprender la posible relación entre la temperatura promedio de la atmósfera terrestre y la emisión de bióxido de carbono en el planeta.

En una biblioteca, se encontró con las dos gráficas que ves abajo.



Unidad analítica: Párrafo introductorio

De estas dos gráficas Andrés concluyó que es cierto que el aumento de la temperatura promedio de la atmósfera terrestre se debe al incremento en la emisión de bióxido de carbono.

Pregunta 1: INVERNADERO S114Q03- 01 02 11 12 99

¿Qué elementos de las gráficas apoyan la conclusión de Andrés?

Pregunta 2: INVERNADERO S114Q04- 0 1 2 9

Otra estudiante, Juana, no está de acuerdo con la conclusión de Andrés. Al comparar las dos gráficas Juana dice que algunas partes no apoyan su conclusión.
Proporciona un ejemplo de partes de las gráficas que no apoyen la conclusión de Andrés. Explica tu respuesta.

Pregunta 3: INVERNADERO S114Q05- 01 02 03 11 12 99

Andrés insiste en su conclusión de que el aumento de la temperatura promedio de la atmósfera terrestre se debe al incremento en la emisión de bióxido de carbono. Pero Juana piensa que su conclusión es prematura. Ella dice: "Antes de aceptar esta conclusión debes estar seguro de que los otros factores que podrían influir sobre el efecto invernadero son constantes".
Menciona uno de los factores a los que se refiere Juana.

Unidades analíticas:
Item 1
Item 2
Item 3

ANEXO 2: Sección de una discusión entre los miembros del comité al analizar los errores de traducción de uno de los ítems

A continuación se transcribe una porción de la conversación colegiada de los miembros del grupo multidisciplinario, donde se pueden apreciar los argumentos que dan las distintas personas sobre los errores que detectan en este reactivo. Por ejemplo, Paola señala que el reactivo analizado presenta problemas del tipo DF6 (de la dimensión *Formato* y de la categoría *Otros*) porque el error detectado no se trata de un componente gráfico, sino de un espacio omitido. Gina refuerza esta opinión comentando que el espacio (omitido) no es un componente gráfico. Consensado este punto, la conversación sigue con la discusión del resto de los errores.

Paola: Es DF6, porque no son componentes gráficos, sino que se omite el espacio.
Gina: Por eso; espacio no es componente grafico.
Paola: Espacio es, cuando es espacio entre palabra y palabra; eso es un componente gráfico, porque es un caracter. Cuando es espacio entre el texto y la tabla, ahí es DF6. Son como intros, ¿no?
Gina: Entonces es DF6.
Willy: Que más señores, ¿nada del 1, 3, 4, 5?
Voz: Yo creo que el "ride" es muy literal; fui a montar en bicicleta, fui a andar en bicicleta, algo así.
Willy: ¿Hace daño?
Voz: No hace daño, pero no me gusta.
Luís: Maneja bicicleta.
Minerva: Andan en bicicleta; el término coloquial es andan en bicicleta.
Willy: Eso sería una cuestión ...
Voz: Literal.
Willy: ¿Literal?, entonces aquí le ponemos LG1. ¿Más señores? ¿del 5?, ¿del 6?
Paola: LS10. Otra vez por el (palabra en inglés). Puse un LS10, por "tabla" por "cuadro". Otra vez un LS10, por eso que dice "la siguiente tabla muestra". No, "en el siguiente cuadro se muestra". Y ahí, por ahí, un (palabra en ingles) lo pusieron como las bicicletas, en lugar de sus bicicletas.
Willy: Sí. Ese seria 6, pero también lo de las dos ruedas; yo diría que también es traducción literal. Porque en realidad sería una cosa, como que avanzan las bicicletas por cada vuelta que dan.
Voz: (inaudible).
Willy: Bueno sin traducción. LS10.
Marichú: Es que en francés decía "recorrida por sus bicicletas"; yo creo que de ahí tomaron el "sus".
Willy: ¿A poco no menciona las ruedas?
Marichu: Sí; pero al final está perfecto, porque es a cada vuelta completa de rueda.
Willy: Bueno, de cualquier manera en español está correcto. ¿Qué más del 7?, ¿8?, ¿9?, ¿10?. Nos vamos.

ANEXO 3. Ejemplo de un reactivo de PISA-2006 revisado, con algunos errores de traducción (cultivos genéticamente modificados)

En bloque de color se señalan los errores de traducción y en los recuadros del lado izquierdo se describen brevemente.

CULTIVOS GENÉTICAMENTE MODIFICADOS

Traducción incorrecta de Wildlife conservation groups

Inserción de palabras (gené-tica-mente modificado) y omisión de palabra (GM)

Inserción de palabra (muy)

Omisión de coma (,) después de potente

Preposición inadecuada (en)

Inserción de doble espacio

DEBE PROHIBIRSE EL MAÍZ GM

Grupos ecologistas exigen que se prohíba el nuevo maíz genéticamente modificado (GM).

Este maíz genéticamente modificado está diseñado para resistir a un nuevo herbicida muy potente que mata a las plantas de maíz convencionales. Este nuevo herbicida matará a la mayor parte de la maleza que crece en los maizales.

Los ecologistas afirman que debido a que esta maleza es alimento para animales pequeños, especialmente insectos, el uso del nuevo herbicida con el maíz GM será dañino para el ambiente. Los que apoyan el uso del maíz GM afirman que un estudio científico ha demostrado que esto no sucederá.

A continuación se presentan detalles del estudio científico mencionado en el artículo anterior:

Empleo impreciso de un término (una mitad)

Estructura sintáctica no natural

Omisión de coma (,) después de nuevo

Omisión de palabra (mitad)

• Se plantó maíz en 200 campos de cultivo en todo el país.

• Cada campo de cultivo se dividió en dos. En una de las partes se cultivó maíz genéticamente modificado (GM), tratado con el potente herbicida nuevo y en la otra el maíz convencional tratado con un herbicida convencional.

• El número de insectos encontrados en el maíz GM, tratado con el nuevo herbicida, era aproximadamente el mismo que el número de insectos en el maíz convencional, tratado con el herbicida convencional.

Pregunta 4: CULTIVOS GENÉTICAMENTE MODIFICADOS S508Q02

En el estudio científico mencionado en el artículo, ¿cuáles factores fueron variados inten-
cionalmente? Encierra en un círculo "Sí" o "No" en cada uno de los siguientes factores.

¿Fue este factor intencionalmente variado en el estudio?	¿Sí o No?
El número de insectos en el ambiente	Sí / No
El tipo de herbicida que se usó	Sí / No

Traducción literal

Falta coma (,) después de No

Error de puntuación (:)

Estructura sintáctica no natural

Traducción inapropiada (los tipos)

Traducción inapropiada (utilizado)

Error de origen: doble consigna en la base, que dificulta innecesariamente la respuesta

Pregunta 5: CULTIVOS GENÉTICAMENTE MODIFICADOS S508Q03

Se plantó maíz en 200 campos de cultivo en todo el país. ¿Por qué los científicos usaron más de un lugar?

Expresión no natural (utilizaron)

A Para que muchos agricultores pudieran poner a prueba el nuevo maíz GM.
B Para ver cuánto maíz GM podían cultivar.
C Para cubrir la mayor cantidad de tierra posible con el cultivo GM.
D Para incluir distintas condiciones de crecimiento para el maíz.

Traducción inadecuada (cultivo)

La Teoría del Error de Traducción de Pruebas
y las evaluaciones internacionales de TIMSS y PISA
Se terminó de imprimir en septiembre de 2011 en
los talleres de Impresora y Encuadernadora Progreso, S. A. de C.V.
Av. San Lorenzo # 244, Col Paraje San Juan
Del. Iztapalapa, C.P. 09830, México D.F.
Para su formación se emplearon los tipos Myriad Pro
y Presidencia Fina y Fuerte a 9 y 11 puntos.
Se imprimieron mil ejemplares