Para citar este artículo, le recomendamos el siguiente formato:

Solano-Flores, G., Contreras-Niño, L. A. y Backhoff-Escudero, E. (2006). Traducción y adaptación de pruebas: Lecciones aprendidas y recomendaciones para países participantes en TIMSS, PISA y otras comparaciones internacionales. *Revista Electrónica de Investigación Educativa, 8* (2). Consultado el día de mes de año en: http://redie.uabc.mx/vol8no2/contenido-solano2.html

Revista Electrónica de Investigación Educativa

Vol. 8, No. 2, 2006

Traducción y adaptación de pruebas: Lecciones aprendidas y recomendaciones para países participantes en TIMSS, PISA y otras comparaciones internacionales[1]

Translation and Adaptation of Tests: Learned Lessons and Recommendations for Participant Countries in TIMSS, PISA and other International Comparisons

Guillermo Solano-Flores
Guillermo.Solano@colorado.edu
School of Education
University of Colorado, Boulder

249 UCB
Boulder, CO 80300-0249
United States of America

Luis Ángel Contreras-Niño
angel@uabc.mx
Universidad Autónoma de Baja California

A. P. 453
C. P. 22800
Ensenada, Baja California
México

Eduardo Backhoff-Escudero
backhoff@inee.edu.mx
Instituto Nacional para la Evaluación de la Educación

José María Velasco 101, Col. San José Insurgentes
C.P. 03900, México, D.F.
México

(Recibido: 21 de noviembre de 2005; aceptado para su publicación: 27 de julio de 2006)

Resumen

En este artículo presentamos un modelo conceptual y una metodología para la revisión de traducciones de pruebas en el contexto de comparaciones internacionales, como es el caso de TIMSS y PISA. También presentamos resultados de una investigación sobre la calidad de la traducción mexicana de TIMSS-1995 a la lengua española. Identificamos errores de traducción en un porcentaje considerable de los ítems, así como correlaciones relativamente altas entre la severidad de los errores de traducción y los valores *p* de los ítems. Estos hallazgos indican que nuestro sistema de codificación de errores es altamente sensible a los errores de traducción de pruebas. Los resultados ponen de manifiesto la necesidad de mejores procedimientos para traducir y revisar la traducción de pruebas en comparaciones internacionales. En nuestra opinión, para poder implementar apropiadamente los lineamientos para la traducción de pruebas en comparaciones internacionales, cada país participante debe tener procedimientos internos para la revisión rigurosa de sus propias traducciones. El artículo concluye con cuatro recomendaciones para países participantes en comparaciones internacionales. Dichas recomendaciones tienen que ver con: (a) las características del personal a cargo de traducir instrumentos, (b) la revisión durante del proceso de traducción de pruebas (no simplemente al final del mismo), (c) el tiempo mínimo necesario para que tengan lugar varias iteraciones de revisión de la traducción de las pruebas, y (d) la necesidad de documentar adecuadamente todo el proceso de traducción de pruebas.

Palabras clave: Pruebas de aprovechamiento, pruebas internacionales, traducción de pruebas, TIMSS, PISA.

Abstract

In this paper we present a conceptual model and a methodology for the review of translated tests in the context of such international comparisons as TIMSS and PISA. We also discuss results from a research on the quality of the Mexican, Spanish translation of TIMSS-1995. We identified translation errors in a considerably high percentage of items and observed relatively high correlations between the severity of translation error and the items' *p* values. These findings indicate that our approach for error coding is very sensitive to test translation error. These findings underscore the need for improved translation and translation review procedures in international comparisons. In our opinion, in order to properly implement guidelines for test translation in international comparisons, each participating country needs to have internal procedures that ensure a rigorous review of its own translations. As a conclusion, we offer four recommendations for countries participating in international comparisons. These recommendations address: (a) the characteristics of the individuals in charge of translating instruments, (b) the use of review

not simply at the end of the process but during the process of test translation, (c) the minimum time needed for various translation review iterations to take place, and (d) the need for properly documenting the entire process of test translation.

Key words: Educational tests, international tests, test translation, TIMSS, PISA.

Introducción

En las últimas dos décadas, la práctica de traducir y adaptar instrumentos de medición educativa a otras lenguas o para culturas diferentes se ha hecho más frecuente a consecuencia de una tendencia hacia la economía global. Los resultados de comparaciones internacionales como TIMSS (por sus siglas en inglés, Estudio Sobre las Tendencias en Ciencias y Matemáticas) y PISA (por sus siglas en inglés, Programa Internacional para la Evaluación de Estudiantes) influyen cada vez más profundamente en la opinión pública y en las políticas educativas de los países participantes, cuyo desempeño en relación con el de otros países es un indicador de progreso académico.

Una consecuencia del auge actual de las comparaciones internacionales ha sido el desarrollo de nuevos y más sofisticados procedimientos de traducción de pruebas que tienen el fin de asegurar que los ítems de una prueba sean equivalentes en múltiples idiomas, a pesar de que existan importantes diferencias culturales entre los países participantes. Un componente clave en las comparaciones internacionales es el uso de un conjunto de lineamientos que garanticen una consistencia mínima en los procedimientos empleados por los diversos países. Por ejemplo, las naciones participantes en TIMSS-1995 emplearon los lineamientos para la traducción de pruebas elaborados por Hambleton (1994) por encargo de la ITC (por sus siglas en inglés, Consejo Internacional de Evaluación) y, a partir de 2007, un nuevo conjunto de lineamientos entrará en vigor (ver Hambleton, 2005).

Otro componente clave en las comparaciones internacionales es el uso de un sistema de revisión de las traducciones de pruebas. Dicho sistema está basado en el empleo de un equipo de revisión de traducciones en el que no participan los traductores de los países participantes. Las traducciones de TIMSS-1995 fueron certificadas por una tercera instancia antes de se autorizara su aplicación por los países participantes (ver, por ejemplo, O'Connor y Malak, 2000).

Este artículo está dirigido a los profesionales involucrados en la traducción de pruebas en comparaciones internacionales. En él mostramos evidencia de que, aunque necesarios, los lineamientos para la traducción de pruebas y para la certificación de calidad de traducción pueden ser insuficientes para asegurar una adecuada traducción de los instrumentos. La implicación fundamental de este trabajo es que, para poder beneficiarse de su participación en comparaciones internacionales, los países deben asegurarse de tener procedimientos internos de revisión, que permitan implementar los lineamientos de traducción de una manera

que sea sensible a aspectos sutiles pero muy importantes del uso del lenguaje en cada país.

El artículo está dividido en cuatro secciones. En la primera sección describimos un modelo conceptual para la revisión de pruebas. Dicho modelo incluye una variedad más amplia de aspectos de traducción que los que normalmente incluyen otros procedimientos. Estos aspectos se refieren a la producción de pruebas, la representación curricular y los aspectos sociales del uso del idioma. En la segunda sección describimos nuestro procedimiento de codificación de errores de traducción, basado en la facilitación de discusiones de comités interdisciplinarios de especialistas. En la tercera sección presentamos evidencia empírica de la sensibilidad de nuestro modelo de revisión. Esta evidencia proviene de nuestro análisis de la traducción mexicana de TIMSS-1995 y de un proyecto, actualmente en curso, en el que analizamos la calidad de la traducción mexicana de PISA-2003. Concluimos con un conjunto de recomendaciones para países que participan en comparaciones internacionales.

I. Marco conceptual para la revisión de traducción de pruebas

El desarrollo de nuestro marco conceptual se basó en el uso combinado de documentos normativos para la traducción de pruebas desarrolladas por el TIMSS (Hambleton, 1994; van de Vijver y Poortinga, 1996), procedimientos para la verificación de la calidad de la traducción de pruebas TIMSS (Mullis, Kelly y Haley, 1996), criterios usados por la American Translators Association (2003), normas y criterios que PISA emplea para determinar la adecuación cultural de los ítems de una prueba (Grisay, 2002; Maxwell, 1996), así como evidencia proveniente de nuestra propia investigación sobre el efecto de las características morfosintácticas de los ítems y los factores sociolingüísticos y epistemológicas, en las interpretaciones que hacen los estudiantes de los ítems en las áreas de ciencias naturales y matemáticas (Solano-Flores y Nelson-Barber, 2001; Solano-Flores y Trumbull, 2003; Solano-Flores, Trumbull y Kwon, 2003).

Nuestro marco conceptual está basado en cinco premisas: (a) inevitabilidad del error en la traducción de pruebas, (b) dimensiones del error de traducción, (c) relatividad de las dimensiones de error, (d) multidimensionalidad de los errores de traducción y (e) naturaleza probabilística de la aceptabilidad de la traducción de un ítem. A continuación explicamos cada una de estas premisas.

1.1 Inevitabilidad del error en la traducción de pruebas

En teoría, la equivalencia de constructo en dos idiomas es imposible de lograr, debido a que cada idioma es específico a una epistemología (Greenfield, 1997). Adicionalmente, la traducción implica que la versión de una prueba en el idioma original y la versión en el idioma objetivo han sido desarrolladas con procedimientos diferentes. La prueba en el idioma fuente ha sido desarrollada mediante un largo proceso iterativo que llega a incluir múltiples revisiones y el

piloteo con muestras de estudiantes. En cambio, el proceso de traducción de pruebas se lleva a cabo en un tiempo relativamente corto, lo que limita la oportunidad de poner a prueba la traducción con estudiantes piloto y refinar la redacción de los ítems (Solano-Flores, Trumbull y Nelson-Barber, 2002).

1.2. Dimensiones del error de traducción

Debido a la inevitabilidad del error, nuestro modelo se ocupa especialmente de la clasificación detallada de errores de traducción. Para los propósitos de nuestra revisión de la traducción de TIMSS-1995, misma que discutimos en la siguiente sección, desarrollamos un sistema de 10 dimensiones de errores de traducción (ver Tabla I). Las dimensiones obvias tienen que ver con exactitud de la traducción, la corrección gramatical (*gramática*, *semántica*) y las características editoriales y de producción de las pruebas traducidas (*estilo*, *formato*, *convenciones*). Otras dimensiones consideran el acuerdo de la traducción con el uso del idioma y con su usanza por la población destinataria, en sus contextos sociales e instruccionales (*registro*), el contenido (*información*, *constructo*) y la representación curricular (*currículum*). Una décima dimensión (*origen*) destaca el hecho de que las fallas del ítem en el lenguaje fuente pueden transferirse a la traducción.

Tabla I. Dimensiones de error en la traducción de pruebas.

Dimensión	Descripción y ejemplos
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. Ejemplos: errores de puntuación; uso impropio de mayúsculas o minúsculas; inconsistencias sujeto-verbo.
Formato	El formato o composición visual del ítem traducido difieren del original en el idioma fuente. Ejemplos: tamaño diferente de tablas; estilo diferente de fuentes de caracteres; márgenes más reducidos; inserción u omisión de componentes gráficos.
Convenciones	La traducción del ítem no se realiza de conformidad con las prácticas convencionales de la escritura de ítems, en el idioma o país destinatario o con los principios básicos de escritura técnica de ítems. Ejemplos: inconsistencia gramatical entre la base y las opciones en ítems de opción múltiple; inconsistencia gramatical entre las opciones en ítems de opción múltiple, extensión diferente de la respuesta correcta en ítems de opción múltiple.
Gramática	La traducción del ítem tiene errores gramaticales o la sintaxis es innecesariamente compleja o inusual para la población destinataria. Ejemplos: traducción literal (palabra por palabra); estructura sintáctica no natural; uso inapropiado de preposiciones
Semántica	Las ideas y el significando transferidos al ítem traducido no son iguales a los del ítem en el lenguaje fuente. Ejemplos: uso de cognados falsos; traducción impropia de expresiones idiomáticas.
Registro	La traducción del ítem no es sensible al uso común de palabras o a los diferentes contextos sociales en la población destinataria. Ejemplos: uso de palabras de baja frecuencia entre la población destinataria; traducción correcta de términos técnicos, pero de una manera que no es común en las escuelas o en los libros de texto del país.
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. Ejemplos: traducción inconsistente de un término que se repite varias veces en el original; un término clave aparece más o menos veces que en el original.
Constructo	La traducción altera el tipo de conocimiento o de habilidades necesarios para responder correctamente el ítem. Ejemplos: traducción inexacta de términos técnicos; inserción u omisión de términos técnicos.
Currículum	El ítem no representa el currículum del país destinatario. Ejemplos: el conocimiento o la habilidad evaluados por el ítem no se enseñan en el país antes o en el grado escolar de la prueba; la manera de plantear un problema no se usa en el currículum del país destinatario.
Origen	El ítem en el lenguaje fuente tiene fallas que no pueden corregirse en la traducción, lo que impone limitaciones para su adecuada traducción. Ejemplos: hay más de una respuesta correcta; ninguna de las opciones es completamente correcta.

Estas dimensiones de error no debieran ser interpretadas como exhaustivas o universales. 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 y los recursos disponibles para el proyecto de traducción. Por ejemplo, mientras que en las pruebas TIMSS la traducción de términos técnicos tiene que examinarse con base en su alineación con el currículum implementado

en grados escolares específicos, en las pruebas PISA es más pertinente examinarla con base en el uso del lenguaje a nivel nacional.

1.3. Relatividad de las dimensiones de error

Una traducción perfecta es prácticamente imposible de lograr, porque existe una tensión entre las dimensiones de error. Por ejemplo, usar un término clave en la traducción de un ítem más veces que en la versión original puede crear una diferencia en la cantidad de información crítica para entender un problema (dimensión de información); pero a la vez, esa inconsistencia puede ser necesaria para asegurar que exista una sintaxis natural según las convenciones en el uso del idioma de la traducción (dimensión de gramática).

1.4. Multidimensionalidad de los errores de traducción

Las propiedades lingüísticas de un ítem están interconectadas. Una consecuencia de esta interrelación es que una característica en la traducción de un ítem pueda pertenecer a más de una dimensión de error. Por ejemplo, la inserción inapropiada de una coma en una oración puede violar algunas convenciones en la redacción del idioma objetivo (dimensión de gramática), pero también puede alterar el significado de una oración (dimensión semántica).

La traducción de un ítem no puede ser totalmente adecuada ni totalmente inadecuada. Es más apropiado pensar en la calidad de la traducción de un ítem en términos de su aceptabilidad. Debido a la multidimensionalidad de errores de traducción, prácticamente ningún ítem está libre de tener errores de traducción. Sin embargo, *error* no debe interpretarse como *error fatal*. Nuestro modelo postula que las personas tienen la capacidad cognoscitiva de manejar los errores de traducción de manera que, dentro de ciertos límites, éstos no afectan a su desempeño en las pruebas. A la vez, algunos errores mínimos que de manera separada podrían no tener importancia, pueden, en conjunto, afectar al desempeño de los examinados.

1.5. Naturaleza probabilística de la aceptabilidad de la traducción de un ítem

De acuerdo con nuestro marco conceptual, el análisis de los errores de traducción puede hacerse con un enfoque probabilístico. Los ítems lingüísticamente aceptables no están totalmente libres de error y no todos los ítems con errores son necesariamente cuestionables. La Figura 1 muestra la aceptabilidad (o cuestionabilidad) de la traducción de un ítem como un espacio probabilístico. Tal espacio probabilístico está delimitado por el número de errores de traducción (eje vertical) y por la severidad (eje horizontal) de esos errores (Solano-Flores, Contreras-Niño, y Backhoff-Escudero, 2005). La traducción de un ítem está en el área de aceptabilidad (área blanca), porque tiene pocos errores o porque tiene varios errores que son poco severos. A su vez, la traducción está en el área de cuestionabilidad (área sombreada), porque tiene pocos errores de traducción muy severos o porque tiene muchos errores que son poco severos.

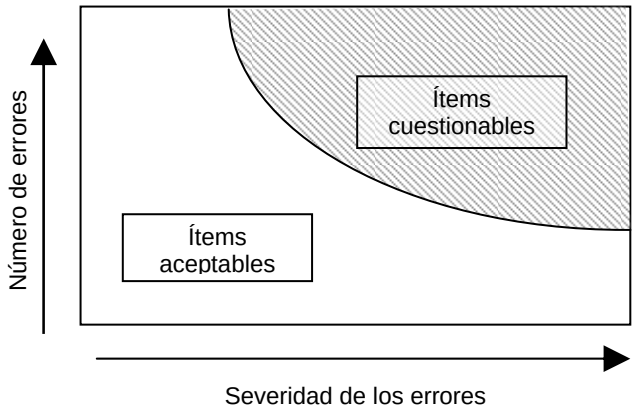

Figura 1. Espacio probabilístico de ítems aceptables y cuestionables, definido por la frecuencia y la severidad de los errores de traducción.

1.6. Procedimiento de revisión de traducción

El procedimiento de revisión de traducción de pruebas que hemos empleado al revisar traducciones de pruebas tiene dos componentes críticos: (a) el empleo de equipos multidisciplinarios de revisores y (b) la codificación de errores mediante comités.

1.7. Equipos multidisciplinarios de revisores

Nuestro trabajo de revisión de traducciones de pruebas ha sido guiado por el principio de que las distintas dimensiones de error de traducción no pueden ser examinadas con suficiente detalle si los participantes en los comités de revisión no son sensibles a distintas formas de uso del idioma. La composición de un comité de revisión de traducción debiera reflejar, tanto las características de la prueba, como sus propósitos y contenido. Por ejemplo, un comité a cargo de revisar la traducción de una prueba de ciencia debiera incluir maestros con experiencia en la enseñanza a estudiantes de los grados escolares correspondientes, especialistas en currículum, lingüistas, traductores profesionales y especialistas en medición. Los maestros conocen el uso formal y coloquial del idioma en el contexto de la enseñanza (por ejemplo, en los libros de texto y en el aula); los especialistas en currículum pueden determinar si la terminología empleada en la traducción de los ítems corresponde al nivel de complejidad del currículum oficial; los lingüistas pueden examinar los aspectos estructurales, funcionales y semánticos de la traducción; los traductores pueden evaluar la precisión de la traducción desde una perspectiva técnica, y los especialistas en medición pueden analizar las demandas cognoscitivas que generan las propiedades lingüísticas de los ítems en la lengua original y en la traducción.

1.8. Procedimiento para la revisión de la traducción

La revisión de la traducción de los ítems se lleva a cabo mediante discusiones de grupo facilitadas por los investigadores del proyecto. Debido a que este proceso pone de manifiesto muchos errores que no pueden ser detectados con otros procedimientos, la discusión de un solo ítem puede llegar a durar hasta 45 minutos. Este tiempo no es muy diferente del que en promedio se debe dedicar a un ítem cuando se le desarrolla en su forma original (ver Solano-Flores, en prensa).

Antes de las sesiones formales de revisión de la traducción, se lleva a cabo una sesión piloto de uno o dos días con el comité de traducción, durante la cual se entrena a los miembros de los comités en el uso de un protocolo de codificación como el que se muestra en la Figura 2. Durante esta sesión piloto se perfecciona la definición de las dimensiones de error, con el fin de que esta definición corresponda con las características específicas de la prueba a revisar. El procedimiento de revisión de cada ítem es el siguiente:

Ítem #	Tipo de error	Codificación y justificación
	1. Estilo	
	2. Formato	
	3. Convenciones	
	4. Información	

Figura 2. Sección del protocolo utilizado para codificar los errores de traducción

1) A cada revisor se le entregó una copia en papel del ítem en español. Se le solicitó que lo leyera y los respondiera como si fuera un estudiante a quien se le aplicaba la prueba. Esto se hizo con el propósito de asegurar que cada revisor se familiarizara con el ítem, fuera consciente del tipo de razonamiento que suscitaba y razonara sobre el tipo de conocimiento necesario para responderlo correctamente.

2) La versión original del ítem en inglés se proyectó en una pantalla para que los revisores pudieran comparar las versiones del ítem en ambos idiomas. Sin embargo, previamente se les pidió que no la miraran hasta que hubieran respondido el ítem.

3) Los revisores examinan el ítem y registran de manera independiente sus observaciones en el formato evaluativo que se ilustra en la Figura 2. En dicho

formato, los revisores registran los tipos de errores de traducción que identifican y, cuando es necesario, justifican sus codificaciones.

4) Los revisores codifican errores de traducción según sus áreas de especialización. Así, el psicólogo y el especialista en medición se concentran en las dimensiones *estilo*, *estructura*, *convenciones*, *origen*, *información* y *constructo*; el traductor en las dimensiones *estilo*, *formato*, *gramática* y *semántica*; el lingüista en las dimensiones *gramática*, *semántica* y *registro*; y los maestros y expertos en currículum en las dimensiones *información*, *contenido* y *currículum*. Sin embargo, se pide a los revisores que indiquen cualquier error que detecten, aunque éste no corresponda a sus áreas de especialización, y que participen en la discusión de todos los errores.

5) Las codificaciones de los errores de traducción son discutidas por todos los miembros del comité hasta alcanzar un consenso. A menudo, esta discusión permite al comité identificar errores que no son detectados por ningún revisor al trabajar de manera individual. Este suele ser el caso de errores de origen (ver Tabla I); es decir, defectos del ítem en la versión original.

6) Durante las sesiones de revisión de la traducción, los maestros y expertos en el contenido tienen acceso a documentos curriculares esenciales como los libros de texto oficiales y las guías para el maestro. Además, se pone a disposición de los comités información liberada por el programa de evaluación (por ejemplo TIMSS o PISA) sobre el contenido de los ítems y sobre los conocimientos y habilidades que pretenden evaluar. Estos documentos permiten a los revisores efectuar interpretaciones apropiadas sobre el significado pretendido de los ítems, tanto en el idioma fuente como en el idioma destinatario.

7) Una vez que llegan a un consenso sobre la manera de codificar los errores de traducción de un ítem, los revisores modifican sus codificaciones individuales originales en los formatos individuales, a fin de que éstas reflejen las decisiones tomadas por el comité.

La evaluación de la calidad de la traducción de una prueba se basa en el análisis estadístico de los errores observados en todos los ítems. Hemos encontrado que los mejores indicadores de la calidad de traducción de una prueba son los que se basan en el número de dimensiones de error *diferentes,* identificadas en los ítems (Solano-Flores, Contreras-Niño, Backhoff-Escudero, 2005). Por ejemplo, un error de formato en un ítem se registra como uno, aun cuando haya varios errores de formato en el mismo ítem.

Recientemente hemos introducido una mejoría en el aspecto operativo del procedimiento descrito. Se ha empleado software especializado[2] para revisar la traducción mexicana de PISA-2003. Dicho software permite a los revisores ver la misma *pantalla de evaluación* (ver Figura 3), en la que se capturan en línea las codificaciones que los revisores acuerdan sobre la calidad de la traducción de los ítems. El programa permite a los revisores ver, tanto la traducción de cada ítem, como sus versiones en las lenguas originales (en el procedimiento empleado para las traducciones de PISA-2003, los ítems fueron traducidos de manera independiente de las dos lenguas fuente: inglés y francés). El uso de un programa de computadora nos ha permitido capturar los datos de codificación de cada ítem durante la discusión de los miembros del comité (Solano-Flores, Contreras-Niño, Backhoff-Escudero y Andrade, 2006). Esta modificación tiene el potencial de reducir el tiempo y los costos del proceso de revisión de traducción.

Figura 3. Ejemplo de "pantalla de evaluación" para la captura en línea de datos evaluativos de la calidad de traducción (El programa fue desarrollado por Edgar Andrade, del INEE)

II. Sensibilidad del modelo de revisión de traducción de pruebas

2.1. La traducción mexicana de TIMSS-1995

Los datos sobre la aplicación del TIMSS-1995 en México son prácticamente desconocidos, porque este país retiró su participación después de que los datos habían sido colectados, pero antes de que los resultados fueran publicados (Backhoff-Escudero y Solano-Flores, 2003).

La creación del Instituto Nacional para la Evaluación de la Educación (INEE) en 2002, inició en México una era de conciencia pública sobre la evaluación y una renovación de las expectativas de transparencia sobre asuntos educativos, especialmente en el contexto de rendición de cuentas.

Como parte de los esfuerzos por evaluar el trabajo previo del sistema educativo en materia de pruebas, el INEE financió una serie de estudios que contribuyeran a obtener el máximo beneficio posible de la participación de México en comparaciones internacionales. En este estudio examinamos la calidad de la traducción mexicana del TIMSS-1995 con la metodología descrita en la sección anterior.

2.2. Muestra de ítems

La mayor parte de los datos derivados de la aplicación del TIMSS-1995 en México fueron destruidos (hecho que en sí muestra la necesidad de desarrollar mecanismos apropiados para documentar el proceso de desarrollo y traducción de pruebas). Sin embargo, poco después de la creación del INEE fue posible recuperar las copias en blanco de todos los cuadernillos de examen aplicados a los estudiantes de la Población 1 (edad de 9 años; tercer y cuarto año de educación primaria) y a los estudiantes de la Población 2 (edad de 13 años; primer y segundo año de educación secundaria).

También fue posible recuperar información sobre los valores *p* de algunos ítems (la proporción de estudiantes que respondieron correctamente a los reactivos) correspondientes a la aplicación de 1995 (que tuvo lugar como parte de la participación de México en la comparación internacional) y a la aplicación realizada por la Dirección General de Evaluación de la Secretaría de Educación Pública en el año 2000. El corpus de ítems de nuestro estudio consistió en 88 ítems de matemáticas y 81 de ciencias naturales, aplicados a la Población 1, y 76 ítems de matemáticas y 74 de ciencias naturales, aplicados a la Población 2 (un total de 319 ítems).

Con el objeto de examinar (aunque de manera restringida) el posible impacto de la calidad de la traducción en el desempeño de los estudiantes, analizamos la correlación entre nuestras medidas de la calidad de traducción con los valores *p* de 42 ítems de matemáticas y 39 de ciencias naturales, aplicados la Población 1, y de 19 ítems de matemáticas y 23 de ciencias naturales, aplicados a la Población 2.

2.3. Participantes

Estudiantes. El análisis de correlación se basó en los valores *p* de los ítems calculados a partir de los registros de estudiantes de la Población 1 (N = 10,122 de tercer grado y N = 10,194, de cuarto grado de primaria) y de estudiantes de la Población 2 (N = 12,809 de primer grado y N = 11,843 de segundo grado de secundaria).

Comités de revisores y procedimiento de revisión. Se formaron dos paneles de revisión de traducción integrados por especialistas con diversa formación y experiencia profesional, de acuerdo con las configuraciones de especialistas descritas anteriormente. Los investigadores del proyecto facilitaron las sesiones de revisión, de acuerdo con los pasos 1 a 7 del procedimiento de revisión descrito también con anterioridad.

Mientras que el lingüista, el psicólogo, el especialista en medición y (obviamente) el traductor tenían dominio del idioma inglés, los maestros y los expertos en el currículum de ambos comités poseían un conocimiento limitado de dicho idioma. Un criterio de evaluación de la calidad de una traducción empleado por los traductores profesionales, es que todo material traducido debe verse como si hubiera sido escrito directamente en el idioma fuente. Consecuentemente, el conocimiento limitado del idioma inglés de los maestros y expertos en currículum permitió que sus juicios sobre los ítems en español no estuvieran influidos por su comprensión de los ítems en el idioma fuente.

Al finalizar la discusión de cada ítem, los comités de revisión acordaron clasificar al ítem como aceptable o cuestionable. Esta decisión fue tomada de manera un tanto subjetiva, aunque colegiada, con base en el impacto que, a juicio de los revisores, la calidad de la traducción pudiera tener en el desempeño del estudiante. Los revisores no tenían conocimiento del procedimiento empleado para cuantificar los errores de traducción de los ítems.

III. Resultados

La falta de información sobre los estudiantes (por ejemplo: medidas externas sobre su desempeño académico, información demográfica, o información sobre sus habilidades lingüísticas o de lecto-escritura) limitó la variedad de análisis que pudimos realizar. En consecuencia, nuestro estudio se limitó al análisis de la frecuencia de errores de traducción observada entre las poblaciones por grado y área de contenido, y a las correlaciones entre las medidas de la calidad de la traducción y los valores *p* de los reactivos.

Frecuencia de errores de traducción. En la mayoría de los ítems se observaron errores de traducción pertenecientes a dos o más dimensiones. En promedio, los ítems tuvieron errores en aproximadamente cuatro dimensiones (media=3.84; S.D. = 1.68; mediana = 4.00; moda = 4.00). Alrededor de 7.5% de los ítems analizados fueron identificados como inaceptables. Éstos tendieron a tener los puntajes de error más altos (Tabla II).

Tabla II. Número de errores: estadística descriptiva de ítems identificados como aceptables y como cuestionables.

	Media	Desviación estándar	Mediana	Error estándar	Moda	Rango
Ítems aceptables (n = 295)	6.25	3.37	6.00	.19	8.00	0-17
Ítems cuestionables (n = 24)	8.83	3.19	8.50	.64	8.00	3-20

En general, observamos patrones similares de frecuencia de errores en ambas poblaciones y en las dos áreas de contenido (Tabla III). Por ejemplo, los errores más frecuentes observados en los ítems de la Población 1 y de la Población 2, y en los ítems de matemáticas y ciencias naturales, provienen de las categorías *semántica* (82.0%-89.8%) y *formato* (75.0%-83.8%). Sin embargo, se observaron diferencias importantes entre las poblaciones o entre las áreas de contenido, en ciertas dimensiones. Una de ellas fue *registro*, cuyas frecuencias fueron diferentes entre poblaciones y entre áreas de contenido dentro de cada población. Los errores de *registro* fueron más frecuentes en los ítems de matemáticas (18.2%), que en los ítems de ciencias naturales (9.9%) para la Población 1; pero fueron más frecuentes en los ítems de ciencias naturales (23.0%), que en los ítems de matemáticas (17.1%) para la Población 2. Otro caso fue *currículum*. Los errores en esta dimensión fueron mucho más frecuentes en los reactivos empleados con la Población 1 (12.5% y 17.3% para matemáticas y ciencias naturales, respectivamente), que con aquéllos utilizados con la Población 2 (3.9% y 2.7% para matemáticas y ciencias naturales, respectivamente).

Tabla III. Errores de traducción por población, área de contenido y dimensión de error, en porcentajes de ítems.

Dimensión de error	Población 1 (9 años)		Población 2 (13 años)	
	Matemáticas (88 ítems)	Ciencias naturales (81 ítems)	Matemáticas (76 ítems)	Ciencias naturales (74 ítems)
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	17.1	23.0
Información	69.3	64.2	44.7	60.8
Constructo	23.9	27.2	15.8	33.8
Currículum	12.5	17.3	3.9	2.7
Origen	17.0	14.8	17.1	21.6

Algunas diferencias en las frecuencias de error provienen de influencias en los estilos discursivos que son más comunes en un área de contenido que en otra. Así, en muchos ítems de matemáticas observamos un error que consiste en

traducir dos oraciones (por ejemplo, “Daniel compra cuatro libretas que cuestan $23.50 cada una. ¿Cuánto dinero necesita?”), como una sola oración en modo gramatical condicional (“¿Cuánto dinero necesita Daniel si compra cuatro libretas que cuestan $23.50 pesos cada una?”).

Correlación entre calidad de traducción y valores *p.* Al interpretar las correlaciones de los puntajes de error con los valores *p* de los reactivos, hay que tener en cuenta que no es lo mismo una correlación que una relación causa-efecto. Una correlación negativa alta simplemente sugiere la posibilidad de un efecto importante de la calidad de la traducción en el desempeño de los estudiantes. También es importante tener en cuenta que la traducción inadecuada no siempre está sesgada en contra de la población examinada en el idioma destinatario (ver Solano-Flores, Trumbull y Nelson-Barber, 2002; van de Vijver y Poortinga, 2005). Por ejemplo, usar más veces un término clave que en la versión del idioma original puede sesgar potencialmente a un ítem en favor de la población examinada en el idioma destinatario.

Los valores *p* de los ítems y el número de errores de traducción correlacionaron de manera diferente dependiendo de la población y el área de contenido (Tabla IV). Para la Población 1 se observaron correlaciones negativas para los ítems de matemáticas y positivas para los ítems de ciencias naturales. Para la Población 2 se observó un patrón opuesto: correlaciones negativas para los ítems de ciencias naturales y positivas para los ítems de matemáticas. 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. En ambas poblaciones, las correlaciones negativas son consistentemente más altas que las correlaciones positivas.

Tabla IV. Correlaciones de Pearson (dos colas) entre el valor *p* del ítem y el número de errores de traducción por población, área de contenido y año de aplicación.

Población 1 (9 años)					
		3 º de primaria		4 º de primaria	
Área de contenido	n	Año 1995	Año 2000	Año 1995	Año 2000
Matemáticas	42	-0.233	-0.262	-0.245	-0.333*
Ciencias naturales	39	0.026	0.024	0.114	0.075
Combinadas	81	-0.119	-0.128	-0.081	-0.139
Población 2 (13 años)					
		1º de secundaria		2º de secundaria	
Área de contenido	n	Año 1995	Año 2000	Año 1995	Año 2000
Matemáticas	19	0.157	0.130	0.102	0.132
Ciencias naturales	23	-0.245	-0.267	-0.187	-0.252
Combinadas	42	0.048	0.003	0.052	0.039

*Correlación significativa (p<.05)

Debido a la carencia de datos complementarios, no es posible hacer una interpretación de estos resultados en términos de la interacción entre una traducción defectuosa y las demandas lingüísticas intrínsecas de cada área de

contenido. Las diferencias en las direcciones de las correlaciones pueden ser un mero efecto de los distintos niveles de calificación de los traductores que tradujeron los ítems de las dos áreas de contenido o los ítems aplicados a las dos poblaciones. Pero esta hipótesis tampoco puede ser examinada debido a la carencia de información sobre el proceso de traducción de los ítems.

Todo lo que podemos concluir, a partir de los resultados obtenidos, es que nuestro procedimiento fue suficientemente sensible a los errores de traducción para poner al descubierto diferencias sistemáticas entre los ítems de distintas áreas de contenido y distintas poblaciones, respecto a las correlaciones entre las medidas de error y los valores *p* de los ítems.

IV. Comentarios finales

Nuestro conocimiento sobre los aspectos delicados que involucra la traducción de pruebas ha aumentado en los últimos 10 años. Por ejemplo, la investigación realizada en el campo de las comparaciones internacionales ha mostrado que, incluso una traducción ligeramente inexacta de una palabra puede ser suficiente para afectar al funcionamiento diferencial de un ítem (Ercikan, 1998). Los procedimientos utilizados en la traducción de pruebas en las comparaciones internacionales han evolucionado de conformidad con ese conocimiento. Por ejemplo, PISA y TIMMS emplean ahora dos idiomas fuente como una estrategia orientada a asegurar que se conserve el significado en la traducción (Grisay, 2002). Además, el uso del procedimiento de *traducción reversa* (basado en volver a traducir al ítem a la lengua original) ya no se acepta como prueba irrefutable de equivalencia entre idiomas.

Sin embargo, a pesar del progreso logrado, la revisión de la traducción de pruebas no ha recibido el mismo tipo de atención; por ejemplo, los aspectos culturales asociados al idioma no siempre se consideran de manera formal. Además, en la práctica usual de traducción de pruebas es común suponer que trabajar con traductores que estén familiarizados con la cultura de la población a quien se destina la traducción es condición suficiente para atender a los principales aspectos del uso del idioma (ver Behling y Law, 2000). Sin embargo, como ya lo hemos discutido aquí, no es posible identificar errores de traducción relacionados con los aspectos sociales de uso del idioma, si no se cuenta con un detallado sistema de codificación.

En este trabajo hemos presentado un marco conceptual para la revisión de la traducción de pruebas, el cual establece dimensiones de error que consideran, entre otras, el uso del idioma en contextos sociales e instruccionales. También hemos descrito cómo se pueden facilitar sesiones de revisión de traducción en las que comités multidisciplinarios de revisores discuten y codifican los errores de traducción.

Los resultados de nuestro análisis de los ítems de la prueba TIMSS-1995 indican que el enfoque es muy sensible a cualquier tipo de error, desde errores menores de producción e impresión hasta errores de semántica y de representación curricular. En promedio, los ítems tuvieron errores de traducción que pertenecen aproximadamente a cuatro dimensiones de error. Los errores más frecuentes se observan en las dimensiones *formato* y *semántica*.

El hallazgo de que casi todos los ítems examinados tuvieron errores de traducción puede resultar impactante en una primera impresión. Sin embargo, estas frecuencias altas pueden ser engañosas si ellos no se interpretan apropiadamente. En primer lugar, necesitamos recordar que *error* no significa *error fatal*. Nuestro procedimiento de calificación está basado en un modelo deficitario; es decir, está diseñado para detectar fallas, más que simplemente decidir si los ítems son o no aceptables. Está basado en el supuesto de que la traducción perfecta de una prueba es virtualmente imposible y en el supuesto de que es el efecto aditivo de muchos errores o la presencia de errores críticos, lo que hace que la traducción de un ítem sea cuestionable.

No obstante, pensamos que las frecuencias de errores observadas son muy altas. Una razón posible de ello es obvia: la calidad de la traducción de la prueba fue deficiente, como lo muestran los 24 ítems identificados por dichos paneles como casos severos de traducción inadecuada y que, por alguna razón inexplicable, pasaron por todos los filtros evaluativos. Otra razón es que nuestro marco conceptual es más detallado y considera aspectos de traducción como registro y semántica desde una perspectiva lingüística más formal.

Es necesario señalar que la alta frecuencia de errores detectados no sólo habla de la calidad de la traducción mexicana de la prueba, sino también de la necesidad de revisar los paradigmas existentes sobre la traducción de pruebas y sobre la revisión de la traducción de pruebas. La versión mexicana de TIMSS-1995 pasó por un proceso de certificación basado en normas utilizadas por la oficina del TIMSS, antes de que fuera aceptada para su uso con estudiantes mexicanos. Un reporte sobre la calidad de los datos del TIMSS-1995 incluye un capítulo sobre los procedimientos empleados para revisar la calidad de las traducciones de la prueba (Mullis, Kelly y Haley, 1996). Además, proporciona el número de ítems identificados como problemáticos en cada país. Según dicho informe –quizá publicado al mismo tiempo que México estaba retirando su participación–, ninguno de los ítems utilizados por México había sido identificado como problemático.

Sería injusto atribuir enteramente la inconsistencia entre ese informe y los resultados de nuestro estudio a las limitaciones de los lineamientos para la traducción de pruebas que en ese momento se emplearon para coordinar el trabajo de traducción de los países participantes. En nuestra opinión, dado su carácter general, no es razonable esperar que dichos documentos normativos produzcan por sí mismos traducciones válidas de pruebas, si los países participantes no tienen procedimientos internos rigurosos de revisión que aseguren la implementación adecuada de esos lineamientos. El que muchos datos

de la aplicación de TIMSS-1995 en México hayan sido destruidos y que no exista la documentación del proceso de traducción de la prueba, son indicios de que indican que los países deben desarrollar mecanismos que garanticen una participación efectiva en comparaciones internacionales.

Para concluir, ofrecemos cuatro recomendaciones para países interesados en asegurar que su participación en comparaciones internacionales cuente con el respaldo de traducciones válidas de pruebas.

1) *Los países deben asignar el personal calificado necesario para un proceso apropiado de traducción*. La traducción de pruebas debe efectuarse por docentes y profesionales de distintas especialidades que trabajan en equipo, discutiendo con detalle cada aspecto de la traducción. Adicionalmente, el proceso de traducción debe ser coordinado por profesionales que tengan un buen conocimiento del proceso de desarrollo de pruebas y por profesionales en el campo de la lingüística.

2) *Los países deben asegurarse de que las actividades de revisión sean parte esencial de todas las etapas del proceso de traducción de pruebas, no simplemente la fase final*. Traducir adecuadamente una prueba debe ser entendida como un proceso tan minucioso como el de desarrollar una prueba en el idioma original. En dicho proceso, las versiones de los ítems se deben perfeccionar a partir de la discusión colegiada de los miembros de los comités de revisión. Esas versiones de los ítems también se deben pilotear con muestras de estudiantes de la población objetivo para asegurarse de que su interpretación sea la adecuada.

3) *Los países deben asegurarse de asignar suficiente tiempo para que haya varias iteraciones de revisión de la traducción de sus pruebas*. Un problema bien conocido sobre la validez de pruebas traducidas es el poco tiempo asignado al proceso de traducción. Un tiempo restringido para la traducción de pruebas limita la posibilidad de discusión entre traductores y revisores y, consecuentemente, la posibilidad de identificar y resolver problemas de traducción.

4) *Los países deben asegurarse de que los equipos a cargo de la traducción de pruebas y de la revisión de la traducción de éstas, documenten todas sus acciones y justifiquen todas sus decisiones*. Documentar el proceso de traducción forma parte de los estándares de la práctica profesional en las áreas de medición y la evaluación (ver American Educational Research Association, American Psychological Association-National Council on Measurement in Education, 1999) Además, contribuye a la formación de recursos humanos en dichas áreas, pues hace posible que las personas involucradas en proyectos de comparaciones internacionales aprendan de la experiencia de otros colegas.

No debiera subestimarse la importancia de estas recomendaciones. Se refieren a aspectos identificados consistentemente como críticos en la literatura sobre el desarrollo, la traducción y la adaptación de pruebas. Para países con una historia reciente de evaluación con pruebas estandarizadas y con escasez de profesionales con entrenamiento formal en el área de medición, estos aspectos son especialmente problemáticos. Ocuparse de ellos es crucial para que puedan derivar en un beneficio sustancial de su participación en comparaciones internacionales.

Referencias

American Educational Research Association, American Psychological Association-National Council on Measurement in Education. (1999). *Standards for Educational and Psychological Testing*. Washington, DC: Autor.

American Translators Association. (2003). *Framework for standard error marking and explanation*. Consultado el 10 de octubre de 2003 en: http://www.atanet.org

Backhoff, E. y Solano-Flores, G. (2003). *Tercer estudio internacional de matemáticas y ciencias naturales (TIMSS): resultados de México en 1995 y 2000*. (Col. Cuadernos de Investigación No. 4). México: Instituto Nacional para la Evaluación de la Educación.

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

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

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

Grisay, A. (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). París: Organisation for Economic Co-operation and Development

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.

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: Boston College.

Mullis, I.V.S., Kelly, D.L., y Haley, K. (1996). Translation verification procedures. En M.O. Martin e I.V.S. Mullis (Eds.), *Third International Mathematics and Science Study: Quality assurance in data collection*. Chestnut Hill, MA: Boston College. Consultado el 10 de octubre de 2003 en: http://timss.bc.edu/timss1995i/TIMSSPDF/QACHP1.PDF

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.

Solano-Flores, G. (en prensa). Successive test development. En C.R. Reynolds, R.W. Kamphaus y C. DiStefano (Eds.), *Encyclopedia of psychological and educational testing: Clinical and psychoeducational applications.* New York: Oxford University Press.

Solano-Flores, G. y Backhoff-Escudero, 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., 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, Québec, Canadá.

Solano-Flores, G., Contreras-Niño, L. A., Backhoff-Escudero, E. y Andrade, E. (2006). Development and evaluation of software for test translation review sessions. Poster presented at the 5th Conference of the International Test Commission: Psychological and Educational Test Adaptation Across Languages and Cultures: Building Bridges Among People. Brussels, Belgium, July 6-8.

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

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

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

Solano-Flores, G., Trumbull, E., y Kwon, M. (2003, 21-25 de abril). *The metrics of linguistic complexity and the metrics of student performance in the testing of English language learners*. Symposio presentado en 2003 Annual Meeting de la American Evaluation Research Association. Chicago, IL.

Vijver, F. van de y Hambleton, R. J. (1996). Translating tests: Some practical guidelines. *European Psychologist 1*, 89-99.

Vijver, F. van de y Poortinga, Y. H. (2005). Conceptual and methodological issues in adapting tests. En R. K. Hambleton, P. Merenda y C. D. Spielberger (Eds.), *Adapting educational and psychological tests for cross-cultural assessment* (pp. 39-63). Mahwah, NJ: Erlbaum.

[1] Algunas porciones e ideas de este artículo provienen de una presentación de los autores en el congreso anual del *National Council on Measurement in Education* en Montreal, Québec, Canadá, en abril de 2005, y de dos presentaciones en el congreso de la International Test Commission en Bruselas, Bélgica, en julio de 2006. Esta investigación fue financiada por el Instituto Nacional para la Evaluación de la Educación (INEE), a través de un convenio de colaboración con la Universidad Autónoma de Baja California. Expresamos nuestro agradecimiento a Felipe Martínez Rizo, Andre Rupp, Norma Larrazolo, Patricia Barón y Arturo Bouzas, por sus comentarios y observaciones a una versión anterior de este trabajo.

[2] El software fue desarrollado por Edgar Andrade, del Instituto Nacional para la Evaluación de la Educación (INEE), México.