Metodología para evaluar la calidad de las traducciones de las pruebas internacionales: el caso mexicano de TIMSS-1995

Guillermo Solano-Flores* Eduardo Backhoff Escudero** Luis Ángel Contreras-Niño***

CUADERNO No. 19

Instituto Nacional para la Evaluación de la Educación

COLECCIÓN CUADERNOS DE INVESTIGACIÓN

ISSN 1665-9457

Metodología para evaluar la calidad de las traducciones de las pruebas internacionales: el caso mexicano de TIMSS-1995

Guillermo Solano-Flores* Eduardo Backhoff Escudero Luis Ángel Contreras-Niño*****

CUADERNO No. 19

Instituto Nacional para la Evaluación de la Educación

Colección Cuadernos de Investigación

ISSN 1665-9457

*University of Colorado, Boulder **Instituto Nacional para la Evaluación de la Educación ***Universidad Autónoma de Baja California

Este texto puede consultarse en: www.inee.edu.mx

MÉXICO, ABRIL, 2006

CONTENIDO

Introducción

La noción básica de validez establece que el puntaje obtenido por alguien examinado en una prueba no debe estar determinado por factores ajenos al constructo que el instrumento pretende medir (Messick, 1995). Un factor importante que influye en la validez de pruebas de desempeño académico es el lenguaje. El lenguaje es necesario para poder aplicar una prueba, aun cuando el constructo que se evalúa no sea la habilidad verbal, la lectura o la escritura (por ejemplo, la habilidad matemática o los conocimientos de ciencias naturales).

Van de Vijver y Poortinga (1997) postulan la existencia de tres tipos de sesgo: de constructo, de reactivo y de método. El sesgo de constructo ocurre cuando el constructo que se pretende medir no existe, o es significativamente diferente entre dos culturas. El sesgo de reactivo se refiere a la varianza irrelevante al constructo que se produce al nivel de reactivo. El sesgo de método ocurre por varianza irrelevante al constructo que se produce sistemáticamente al nivel del puntaje total en una prueba. Una traducción inadecuada puede ser, por lo tanto, una fuente de sesgo de método que puede afectar los puntajes obtenidos por los estudiantes con la prueba traducida.

En años recientes los especialistas en el uso y desarrollo de instrumentos de medida han prestado cada vez más atención a la manera en que el lenguaje afecta el desempeño de los examinados. Tres razones explican este fenómeno:

1. La manera en que se interpreta un reactivo es muy sensible a cualquier variación en la forma en que está redactado. Una sola palabra puede hacer una enorme diferencia en lo que los estudiantes creen que tienen que hacer para resolver un problema (e.g., Baxter, Shavelson, Goldman, y Pine, 1992).
2. Cada vez es más apremiante la necesidad de contar con estrategias efectivas para evaluar minorías lingüísticas. Ello es especialmente evidente en Estados Unidos. A pesar de los esfuerzos por incluir a los estudiantes que no hablan inglés como primera lengua, en los sistemas nacionales de evaluación los progresos en esa dirección han sido mínimos (Wise, Hauser, Mitchell y Feuer, 1999).
3. Como resultado de la globalización de la economía, se presenta ahora con más frecuencia la necesidad de traducir pruebas de aprendizaje a idiomas en los que no

fueron escritos originalmente o adaptarlos para poblaciones para los que no fueron creados (Hambleton, 1994).

Por lo anterior, la traducción de pruebas a idiomas para los cuales no fueron escritas originalmente es ahora un componente importante en el campo educativo de la evaluación internacional (Hambleton, 1994). Las comparaciones entre países han impuesto desafíos importantes para el desarrollo de instrumentos de medición en más de un idioma. Como resultado, las normas y procedimientos para la adaptación lingüística de pruebas siguen evolucionando (Grisay, 2002).

Este documento es una contribución a la revisión de traducción de pruebas. En él se destaca el hecho de que, aunque necesarias, las guías y reglas para la traducción de pruebas no son suficientes para asegurar la validez de los instrumentos traducidos. Aquí describimos un procedimiento para revisar la traducción de pruebas que considera una variedad amplia de aspectos críticos en su traducción. En este trabajo presentamos un marco conceptual y metodológico para evaluar la calidad de la traducción de pruebas y aportar evidencia empírica sobre el efecto que tiene dicha traducción en la dificultad del reactivo.

Para probar este marco de referencia, se analiza la traducción mexicana de las pruebas del Tercer Estudio Internacional de Matemáticas y Ciencias (TIMSS, por sus siglas en inglés) realizado en 1995 y posteriormente, replicado en 2000 por las autoridades educativas mexicanas (véase Backhoff y Solano-Flores, 2003).

La mayor parte de los datos derivados de la aplicación del TIMSS-1995 en México fueron destruidos. Sin embargo, poco después de la creación de INEE, fue posible recuperar las copias en blanco de todos los cuadernillos de examen aplicados a los estudiantes. También fue posible recuperar información sobre los valores p de los ítems (proporción de estudiantes que respondieron correctamente cada reactivo) correspondientes a los años 1995 y 2000.

Por lo anterior, los objetivos de este trabajo son:

1. Desarrollar y poner a prueba un marco conceptual y metodológico para la revisión de la traducción de pruebas internacionales.
2. Evaluar la calidad de la traducción mexicana de las pruebas utilizadas en TIMSS-1995.

Antecedentes de la traducción de las pruebas

El documento sobre la traducción de pruebas de rendimiento académico utilizadas en estudios internacionales (Hambleton, 1994) constituyó la base para la traducción de las pruebas TIMSS-1995 que utilizaron los países participantes de este estudio. Cada país fue responsable de implementar correctamente estos lineamientos. Por ejemplo, los lineamientos recomendaron el uso de traducción múltiple, donde dos o más personas deben hacer traducciones independientes y una tercera persona las debe comparar (Grisay, 2002). Idealmente, un mismo par de traductores independientes debe traducir todos o la mayoría de los reactivos. Las discrepancias entre las traducciones se discuten y finalmente, con base en esa discusión, se decide cómo resolver las discrepancias. Sin embargo, debido a limitaciones económicas y de tiempo, no todos los países participantes aplican estos procedimientos (Maxwell, 1996). Ello hizo que la calidad de las traducciones de TIMSS-1995 dependiera más de la competencia misma de los traductores empleados en cada país, así como de los procedimientos de revisión de las pruebas (a cargo de una compañía de traducción ubicada en Canadá).

Los lineamientos establecían que los traductores de los reactivos debían tener cinco características: 1) buen conocimiento del idioma inglés, 2) conocimiento excelente del idioma objetivo, 3) experiencia en ambos idiomas y culturas, 4) experiencia con estudiantes de las poblaciones objetivo y 5) habilidades necesarias en el proceso de desarrollo de pruebas.

Los lineamientos también establecían que los traductores debían llevar a cabo seis acciones básicas en el proceso de traducción: 1) identificar y minimizar diferencias culturales, 2) encontrar palabras y frases equivalentes entre idiomas, 3) asegurar que el nivel de lectura de los reactivos en el idioma objetivo fuera el mismo que en la versión original en inglés, 4) asegurar que el sentido esencial de los reactivos no sufriera alteraciones, 5) asegurar que el nivel de dificultad de los reactivos no cambiara y 6) poner atención a posibles cambios en la apariencia de los reactivos como resultado de la traducción.

Aun cuando los países participantes aplicaran los lineamientos de TIMSS-1995, éstos tienen un amplio margen de interpretación. Por ejemplo, los criterios que definen a un buen traductor, serían muy ambiguos de acuerdo con la práctica y las normas de la *American Translators Association* (ATA) (2003).

Finalmente, es importante señalar que en el caso de la traducción mexicana, no están documentados los procedimientos utilizados ni las personas que realizaron dicha traducción, por lo que se desconoce a ciencia cierta la forma en que se realizó la traducción nacional de TIMSS-1995. Lo único que se tiene es copia de los cuadernillos traducidos al español y la base de datos de los resultados obtenidos.

MARCO CONCEPTUAL

El marco conceptual que desarrollamos para la revisión de la traducción de pruebas se basó en las siguientes fuentes: 1) un análisis inicial de las características de los ítems correspondientes a la aplicación mexicana del TIMSS-1995 (Solano-Flores y Backhoff, 2003); 2) los comentarios surgidos en los paneles multidisciplinarios instrumentados para la revisión de la traducción, durante una serie de sesiones piloto para la revisión (ver más adelante); 3) las guías para la traducción de pruebas desarrolladas por TIMSS-1995 (Hambleton, 1994); 4) el conjunto de criterios empleados por el TIMSS-1995 como parte de sus procedimientos para verificar la traducción (Mullis, Kelly y Haley, 1996); 5) los criterios usados por la *American Translators Association* (2003) para evaluar la calidad de las traducciones; 6) el conjunto de normas y consideraciones actualmente empleadas por PISA para determinar la adecuación cultural del ítem (Grisay, 2002; Maxwell, 1996); 7) la evidencia aportada por la investigación sobre el uso de la lingüística estructural para detectar complejidad innecesaria en la sintaxis de los ítems de una prueba (Solano-Flores, Trumbull, y Kwon, 2003) y 8) cierta evidencia sobre la influencia de factores sociolingüísticos y epistemológicos en las interpretaciones que hacen los estudiantes sobre los ítems en las áreas de ciencias naturales y matemáticas (Solano-Flores y Nelson-Barber, 2001; Solano-Flores y Trumbull, 2003).

Identificamos diez dimensiones de error en la traducción de un test (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 y uso impropio de mayúsculas o minúsculas.
Formato	El formato o composición visual del reactivos 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 reactivos 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 reactivo 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 reactivo 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 necesarias para responder correctamente el reactivo. 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 enseña 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 reactivo 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.

Desde nuestro punto de vista, el proceso de traducción de una prueba no se limita al trabajo de los traductores, sino que incluye también acciones que tienen lugar durante el desarrollo de la prueba y después de que la prueba ha sido traducida. Así, *Currículum*, *Origen* y *Formato* tienen que ver, respectivamente, con la falta de alineación de los reactivos con el currículum, el proceso de desarrollo de la prueba antes de su traducción, y el proceso de producción gráfica e impresa una vez que la prueba ha sido traducida. Aunque estric-

tamente, no se refieren al trabajo de los traductores, estas tres dimensiones influyen en las características de una prueba traducida y, potencialmente, pueden afectar su validez y el desempeño del estudiante en esa prueba.

Una noción básica en nuestro marco conceptual es que un error puede ser clasificado en una o más dimensiones de error. Por ejemplo, la inserción inadecuada de una coma en una frase, en un ítem traducido, puede violar algunas convenciones gramaticales; pero también puede cambiar el significado pretendido de una idea.

En general, los errores de estilo, estructura, y convención tienden a ser más leves que aquéllos pertenecientes a otras dimensiones. Sin embargo, existen múltiples excepciones a esta regla. Supongamos que una figura en la versión traducida de un ítem tiene estilos de letra que son diferentes a los estilos empleados en la versión original. En el caso más simple, la diferencia no tiene efecto alguno; en el caso más extremo, la diferencia afecta la atención que los estudiantes prestan a algunos componentes de la figura e incluso su interpretación del ítem.

Una característica esencial de nuestro marco conceptual es que examina a los errores de traducción como fuentes de invalidez de pruebas desde una perspectiva probabilística. En primer lugar, más que ser absoluta, la severidad de un error de traducción está moldeada por factores contextuales y por las características de la población destinataria. Así por ejemplo, el hecho de que la inserción impropia de una coma, en el ejemplo mencionado anteriormente, sea un error de la gramática, severo o leve, o un error de semántica severo o leve depende, entre muchas otras cosas, de las características del ítem (por ejemplo, el tipo y cantidad de información que proporciona), el conocimiento de los estudiantes sobre el contenido que se está evaluando y la habilidad de los estudiantes para comprender el significado pretendido del ítem.

En segundo lugar, la probabilidad de que un ítem traducido sea problemático desde una perspectiva lingüística está determinada por el efecto combinado de la frecuencia y severidad de los errores de traducción. Esta noción puede representarse como un espacio probabilístico, tal como se muestra en la figura 1. Las áreas claras y sombreadas representan, respectivamente, la probabilidad de que un ítem sea aceptable o cuestionable. Así, es probable que los ítems cuestionables ocurran cuando éstos tienen errores de traducción serios, o de poca importancia, pero numerosos.

En nuestra opinión, los ítems lingüísticamente aceptables no están totalmente libres de error y no todos los ítems con errores son necesariamente ítems cuestionables. Desde una postura teórica más rigurosa, no existe una traducción perfecta y, en el caso de las pruebas, no es posible transferir exactamente el mismo significado de un idioma a otro ni medir exactamente el mismo constructo en diferentes idiomas (Greenfield, 1997). A su vez, cometer un error no significa necesariamente una falta irremediable puesto que las personas tienen un papel activo al interpretar el contenido de los reactivos de una prueba. Dentro

de ciertos límites, un estudiante es capaz de reconocer ciertos errores de traducción o de manejar (corregir) cognitivamente un número máximo de errores.

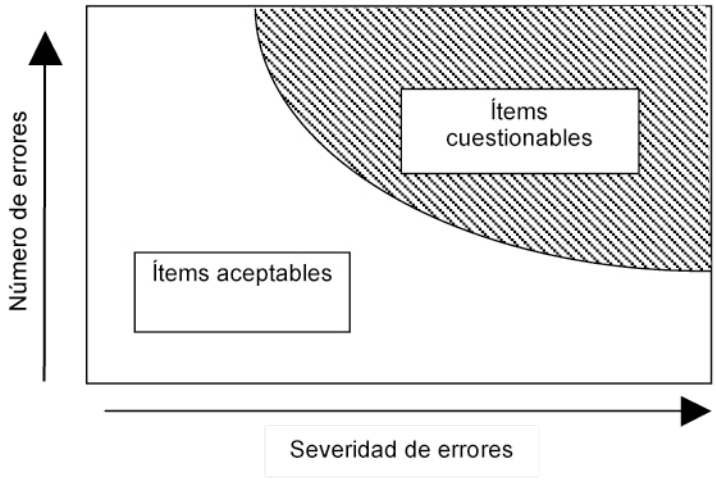

Figura 1. Espacio probabilístico de ítems cuestionables, definido por la frecuencia y severidad de errores de traducción

MÉTODO

Tanto el estudio de TIMSS-1995 de México, como la réplica de SEP-2000, contemplaron dos poblaciones de estudiantes:

- **Población uno:** estudiantes de tercer y cuarto grados de primaria. En el estudio de TIMSS-1995 y en el de la SEP-2000 se evaluaron a 20 mil 316 y 7 mil 335 estudiantes, respectivamente.
- **Población 2:** estudiantes de primero y segundo de secundaria. En el estudio de TIMSS-1995 y en el de la SEP-2000 se evaluaron a 24 mil 652 y 8 mil 318 estudiantes, respectivamente.

Para cada población se desarrolló una prueba particular con un diseño matricial de bloques incompletos.[1]

[1]Los diseños matriciales se utilizan cuando se desea evaluar una gran cantidad de contenidos, por lo que cada estudiante contesta sólo algunas partes de la prueba, que generalmente se agrupan en bloques. Se dice que los bloques son incompletos cuando no se hacen todas las permutaciones posibles entre ellos.

Reactivos evaluados

Las pruebas de TIMSS-1995 estuvieron compuestas de reactivos de Matemáticas y Ciencias Naturales, tanto de selección como de respuesta construida. Como se muestra en la tabla II, el total de reactivos para la población uno fue de 199, mientras que para la población dos fue de 286.

Tabla II. Composición de las pruebas para las poblaciones uno y dos

Pruebas	Contenido temático	Número de reactivos (K)		
		Opción múltiple	Respuesta construida	Total
Población uno	Matemáticas	79	23	102
	Ciencias	74	23	97
Población dos	Matemáticas	125	26	151
	Ciencias	102	33	135
Total	**380**	**105**	**485**	

Para ambas pruebas, se examinó la traducción de una muestra de 319 ítems, es decir 66 por ciento del total de reactivos. De los ítems examinados de la población uno, 88 eran de Matemáticas y 81 eran de Ciencias Naturales; de los reactivos de la población dos, 76 eran ítems de Matemáticas y 74 de Ciencias Naturales.

Adicionalmente al análisis de la traducción de ítems, se correlacionaron las puntuaciones de la calidad de su traducción con sus correspondientes valores p, calculados a partir de los registros de estudiantes que se pudo rescatar. De hecho, los ítems incluidos en estos análisis de correlación fueron una submuestra de los 319 ítems originales, debido a que se contaban con los valores p. La tabla III muestra el total de reactivos utilizados en los análisis de este estudio.

Tabla III. Número y tipo de reactivos que se utilizaron en este estudio

Población	Contenido temático	Número de reactivos (K)		
		Aplicados en TIMSS-1995	Traducidos y evaluados	Utilizados en las correlaciones
Población uno	Matemáticas	102	88	42
	Ciencias	97	81	39
Población dos	Matemáticas	151	76	19
	Ciencias	135	74	23
Total	**485**	**319**	**123**	

PÁNELES DE JUECES

Para cada prueba se formó un panel de revisión de traducción integrados por los siguientes especialistas: docente, experto en currículum, lingüista, traductor certificado, psicólogo y psicómetra.

El panel uno (correspondiente a la prueba de primaria) estuvo integrado por un experto en el currículum de la educación primaria; dos maestros en servicio destacados del quinto grado; un lingüista; un traductor profesional certificado por la ATA; y un psicólogo y un psicómetra, ambos con amplia experiencia en el campo del desarrollo de pruebas.

El panel dos (correspondiente a la prueba de secundaria) estuvo integrado por un experto en las áreas de Matemáticas y Ciencias Naturales del currículum de la educación secundaria; un maestro en servicio destacado de cada una de las cuatro áreas de contenido que se exploran en el examen y que son parte del currículum de la educación secundaria mexicana (Matemáticas, Física, Química y Biología); y el lingüista, el traductor profesional, el psicólogo, y el psicómetra que participaron también en el panel uno.

Mientras que el lingüista, el psicólogo, el psicómetra y el traductor tenían dominio del idioma inglés, los maestros y los expertos en el currículum de ambos paneles poseían un conocimiento limitado de dicho idioma.[2]

[2]Más que desventaja, consideramos que esta diferencia en el dominio del inglés era una ventaja en nuestro estudio, ya que nuestro propósito era obtener de los expertos curriculares y maestros una perspectiva evaluativa de los ítems que no estuviera influida por su comprensión de los ítems en el idioma fuente.

Procedimiento para la evaluación de la traducción

Antes de la revisión de la traducción de las pruebas, se llevó a cabo una sesión de capacitación con el fin de entrenar a los especialistas en el uso del protocolo de codificación. Dicha sesión sirvió también para discutir y refinar nuestro marco conceptual de revisión de traducción de reactivos.

Las sesiones de campo para la revisión de la traducción de la prueba duraron cinco días para cada panel. Cada uno de los reactivos seleccionados fue revisado de conformidad con el procedimiento siguiente:

1) A cada revisor se le entregó una copia en papel del reactivo traducido al español y se le solicitó que lo respondiera como si fuera un estudiante. Esto se hizo con el propósito de asegurar que se familiarizarán con el ítem, fueran conscientes del tipo de razonamiento que suscita y razonara sobre el tipo de conocimiento necesario para responderlo.
2) La versión original del ítem en inglés se proyectó sobre una pantalla para que los revisores pudieran comparar las versiones del ítem en ambos idiomas.
3) Enseguida, los revisores examinaron el ítem y registraron sus observaciones en el protocolo evaluativo de manera independiente. Es decir, marcaron los tipos de errores de traducción que identificaron y, cuando fue necesario, proporcionaron una breve explicación que apoyara las decisiones que adoptaron al efectuar la codificación (ver figura 2).
4) Aunque los revisores eran libres de discutir e identificar errores en todas las dimensiones, se les pidió que se concentraran específicamente en aquellas dimensiones que correspondían a sus antecedentes profesionales. Así, el psicólogo y el psicómetra se concentraron 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 *Instrucción, Contenido* y *Currículum*.

Item #	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

5) Las decisiones que tomó cada revisor al codificar los errores de traducción de los reactivos evaluados fueron revisadas por los demás miembros del panel. Cuando fue necesario, las decisiones sobre codificación de errores fueron discutidas por el comité hasta que se logró un consenso. En algunos casos, la discusión permitió al comité identificar errores que no fueron detectados por ningún miembro cuando trabajaron de manera individual. Es en estos casos fue posible identificar errores en los cuales la calidad de la traducción estaba limitada por fallas del reactivo en la versión en inglés (errores de *Origen*). Además, el panel acordó etiquetar con una *bandera roja* aquellos reactivos que implicaban un reto lingüístico serio, cuando contenían algunos errores severos o muchos errores leves.
6) Durante las sesiones de revisión de la traducción del examen, los maestros y los expertos en el contenido tuvieron acceso a documentos curriculares clave como los libros de texto oficiales y las guías para el maestro. Además, se puso a disposición de los paneles información liberada por el TIMSS sobre el contenido de los ítems y sobre los conocimientos y habilidades que evaluaban, de tal manera que pudieran efectuarse interpretaciones apropiadas sobre el significado pretendido de los ítems tanto en el idioma fuente como en el idioma destinatario.
7) Los revisores actualizaron sus formatos individuales de revisión al final de la discusión en grupo de cada reactivo, para que éstos reflejaran las conclusiones a las que llegó el panel. La captura y el análisis de los resultados se basaron en el número de errores diferentes identificados por el panel en cada dimensión.

Resultados

Errores de traducción

En ambas pruebas se observaron errores de traducción en casi todos los reactivos de la muestra. Su inmensa mayoría tuvo errores que pertenecían a dos o más dimensiones de error de traducción, siendo el promedio de cuatro dimensiones por ítem (figura 3).

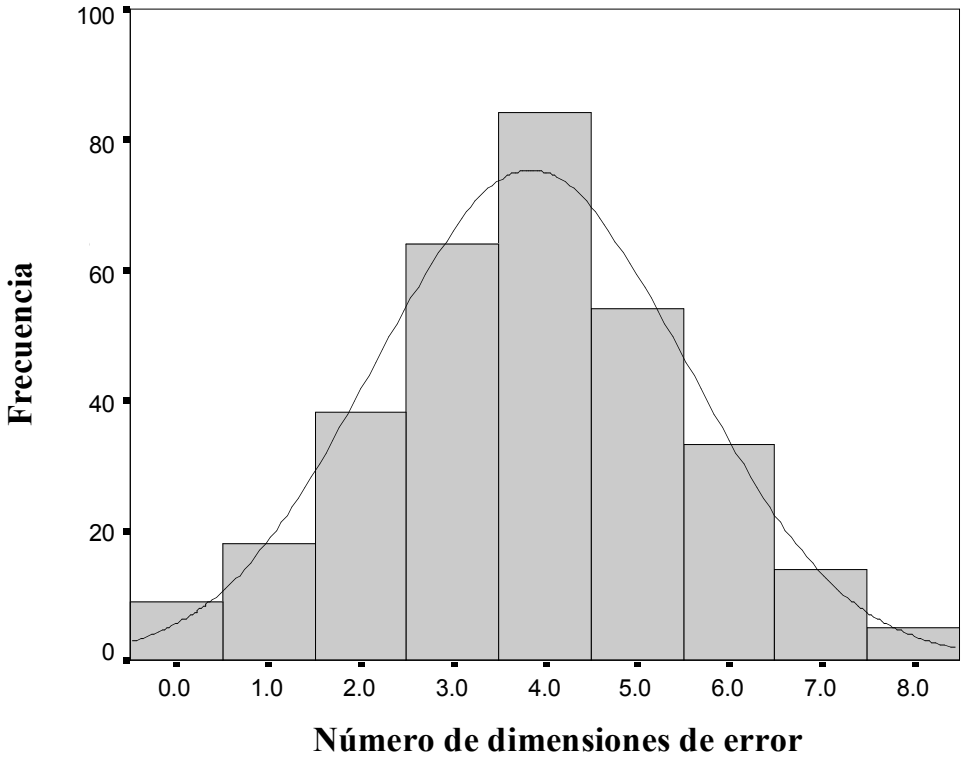

Figura 3. Distribución de frecuencia de ítems según el número de dimensiones de error de traducción observados. K=319; Media=3.84; D.E. =1.68; Mediana=4.00; Moda=4.00

En general, observamos patrones similares de frecuencia de errores en ambas pruebas y en las dos áreas de contenido (tabla IV). Por ejemplo, los errores más frecuentes observados en los reactivos de matemáticas y ciencias naturales pertenecen a las categorías *Semántica* (82.0-89.8 por ciento) y *Formato* (75.0-83.8 por ciento). Sin embargo se observaron diferencias importantes en las dimensiones de error entre los reactivos de las dos pruebas y áreas temáticas. Así, los errores de registro fueron más frecuentes en los reactivos de matemáticas (18.2 por ciento) que en los ítems de ciencias naturales (9.9 por ciento) para la

población uno; pero fueron más frecuentes en los reactivos de ciencias naturales (23.0 por ciento) que en los ítems de matemáticas (17.1 por ciento) para la población 2. Asimismo, los errores en esta dimensión *Currículum* fueron mucho más frecuentes en los reactivos empleados con la población uno (12.5 y 17.3 por ciento, respectivamente, para Matemáticas y Ciencias Naturales) que en aquéllos de la Población 2 (3.9 y 2.7 por ciento, respectivamente, para matemáticas y ciencias naturales).

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

Dimensión de error	Población uno		Población dos	
	Matemáticas (k=88)	Ciencias naturales (k=81)	Matemáticas (k=76)	Ciencias naturales (k=74)
1. Estilo	17.0	28.4	21.1	29.7
2. Formato	75.0	76.5	82.9	83.8
3. Convenciones	21.6	37.0	11.8	35.1
4. Gramática	36.4	39.5	36.8	47.3
5. Semántica	89.8	85.2	82.9	87.8
6. Registro	18.2	9.9	17.1	23.0
7. Información	69.3	64.2	44.7	60.8
8. Constructo	23.9	27.2	15.8	33.8
9. Currículum	12.5	17.3	3.9	2.7
10. Origen	17.0	14.8	17.1	21.6

Nota: K = número de reactivos

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 reactivos de matemáticas observamos un error que consiste en traducir dos oraciones (por ejemplo, *Daniel compra tres lápices que cuestan dos pesos cada uno ¿Cuánto dinero necesita?*) como una sola oración en la que el modo gramatical es condicional (*¿Cuánto dinero necesita Daniel si compra tres lápices que cuestan dos pesos cada uno?*)

Calidad de la traducción y dificultad del ítem

Para analizar el impacto que tiene la mala traducción de un reactivo en la ejecución de los estudiantes, correlacionamos el número de errores de traducción del ítem con su respectivo valor *p* (proporción de estudiantes que respondieron correctamente el reactivo). No utilizamos la *severidad* de los errores de traducción como medida de calidad del ítem por la multitud de características contextuales que intervienen en la comprensión de un ítem, lo que hace muy poco confiable este indicador

Por un lado, comparamos las frecuencias de reactivos marcados con *bandera roja* con los demás ítems. Como era de esperarse, los primeros tienen una media de errores más alta que los segundos (tabla V). Desafortunadamente, el número reducido de ítems con *bandera roja* impide evaluar la significancia estadística de dicha diferencia.

Tabla V. Número de errores de ítems con y sin *bandera roja*

Reactivos traducidos	K	Media	Desviación estándar	Error estándar	Rango
Con *bandera roja*	24	8.83	3.19	.64	3-20
Sin *bandera roja*	295	6.25	3.37	.19	0-17

Por otro lado, correlacionamos el número de errores con el valor *p* del reactivo, con el método de Pearson. El sentido común hace pensar que deben observarse correlaciones negativas cuando los errores de traducción tienen un efecto adverso sobre el desempeño de los estudiantes. Sin embargo, esto no es necesariamente cierto, pues hay errores de traducción que pueden atenuar la magnitud y el nivel de importancia de estas correlaciones. La traducción inadecuada no siempre está sesgada en contra de la población examinada en el idioma destinatario (Solano-Flores, Trumbull, y Nelson-Barber, 2002). 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 destinatorio. Además, mientras algunos errores de traducción importantes pueden tener un impacto sobre la validez de contenido de un reactivo, éstos no necesariamente tienen efectos notables en el desempeño del estudiante.

Estas razones y un poder estadístico limitado de las correlaciones que involucran a los ítems aplicados a la población dos (19 y 23, respectivamente, para matemáticas y ciencias naturales) nos obligan a concentrarnos en su magnitud y dirección y en el patrón de las correlaciones en las poblaciones y áreas de contenido, más que en su significancia estadística.

Observamos que los valores *p* de los ítems y el número de errores de traducción correlacionan de manera diferente dependiendo de la población y del área de contenido (tabla

VI). Para la población uno, se observan correlaciones negativas y positivas, respectivamente, para los ítems de Matemáticas y de Ciencias Naturales. Para la población dos, se observa el patrón opuesto: correlaciones negativas y positivas, respectivamente para los ítems de Ciencias Naturales y Matemáticas. Estos patrones 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 VI. Correlaciones entre el valor *p* del ítem y el número de errores de traducción del reactivo

Área de contenido	K	TIMSS-1995	SEP- 2000	TIMSS-1995	SEP- 2000
Población uno		**3o de primaria**		**4o de primaria**	
Matemáticas	42	-.233	-.262	-.245	-.333*
Ciencias	39	.026	.024	.114	.075
Población dos		**1º de secundaria**		**2º de secundaria**	
Matemáticas	19	.157	.130	.102	.132
Ciencias	23	-.245	-.267	-.187	-.252

De acuerdo con el patrón de correlaciones observado, la traducción defectuosa en el desempeño de los estudiantes tiene un mayor impacto en el desempeño de los estudiantes en reactivos de matemáticas que en los de ciencias naturales para población uno, y un mayor impacto para los reactivos de Ciencias Naturales que para los ítems de Matemáticas en la población dos. Este patrón puede interpretarse como una indicación de que existe una interacción entre los efectos de la traducción defectuosa y las demandas lingüísticas intrínsecas de cada área de contenido. Sin embargo, también puede interpretarse como un resultado de los distintos niveles de calidad en el trabajo de traducción de los ítems en las dos áreas de contenido y las dos poblaciones.

De qué manera el área de contenido y el grado escolar moldean el efecto de la traducción inadecuada sobre la ejecución del estudiante, permanece como una pregunta de investigación para el futuro. Sin embargo, los análisis efectuados muestran que la metodología descrita hace posible evaluar la calidad de la traducción de pruebas.

Discusión

Los procedimientos utilizados en la traducción de pruebas internacionales han evolucionado en los últimos años. 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, ya no se acepta como prueba irrefutable de equivalencia entre idiomas el uso del procedimiento de *traducción hacia atrás*; en, el que se traduce la traducción de regreso al idioma original y se compara con las dos versiones del idioma fuente para identificar y corregir cualquier diferencia sustancial de contenido entre la versión en el idioma fuente y la versión en el idioma destinatario.

A pesar de tales avances en el procedimiento de traducción de pruebas, existen serias diferencias en el rigor metodológico con que se desarrolla una prueba y el rigor metodológico empleado para su traducción. Por lo general, el desarrollo de una prueba llega a tomar una considerable cantidad de tiempo (meses e incluso años). Sin embargo, es común que al trabajo de traducción de la misma prueba se le asignen poco tiempo y pocos recursos. En el mejor de los casos, esta traducción se realiza con una sola iteración de revisiones (Valdés y Figueroa, 1994; Solano-Flores, Trumbull y Nelson-Barber, 2002).

La necesidad de contar con un adecuado procedimiento de revisión de traducción de pruebas queda de manifiesto en los resultados de investigación en el campo de los estudios internacionales de logro académico. Existe evidencia de que una simple variación en la traducción de un reactivo puede ser suficiente para afectar su funcionamiento diferencial (Ercikan, 1998).

En este trabajo, hemos reportado un estudio sobre la calidad de la traducción de la versión mexicana del TIMSS-1995. Aunque esta prueba tiene diez años de que se tradujo y se aplicó en nuestro país, nos sirvió de plataforma para poner a prueba un marco conceptual de revisión de traducciones y también para valorar la calidad con que se realizó dicha traducción. Una parte fundamental de este marco conceptual es el uso de un panel interdisciplinario de revisores que identifiquen y codifiquen los errores de traducción de los reactivos de una prueba. Igualmente importante es su sensibilidad, ya que permite la clasificación de muchos tipos de errores de traducción, desde errores menores de producción e impresión, hasta errores de semántica y de representación curricular.

Los resultados de este estudio señalan que en la inmensa mayoría de los ítems de TIMSS-1995 hay errores de traducción. Los errores más frecuentemente observados pertenecieron a las dimensiones *Formato* y *Semántica*. Un análisis de correlaciones entre el número de errores de traducción de los ítems y sus valores *p* sugiere que estos errores afectan la dificultad del reactivo diferencialmente. Para los estudiantes de primaria, las correlaciones fueron negativas para los reactivos de Matemáticas y positivas para los ítems de Ciencias Naturales; el efecto opuesto ocurrió para los estudiantes de secundaria.

Es importante señalar que las correlaciones negativas fueron consistentemente más altas que las correlaciones positivas, lo cual apoya la noción de que los errores de traducción tienden a sesgar a los reactivos contra los estudiantes examinados en la lengua destinataria. Es posible que este efecto diferencial de la calidad de la traducción por área de contenido y grado educativo se deba al hecho de que diferentes áreas de contenido (y los ítems que pretenden evaluar conocimientos en dichas áreas) imponen distintas demandas lingüísticas diferentes en distintos grados educativos. Sin embargo, la falta de datos adicionales nos impide efectuar los análisis apropiados para poner a prueba tal hipótesis. Una interpretación alternativa de los resultados es que dichas diferencias son el resultado de distintos niveles de calidad de la traducción realizada por diversos traductores.

El hecho de que la mayoría de los ítems examinados tuvieron errores de traducción puede parecer sorprendente en una primera impresión. Sin embargo, estas frecuencias altas pueden ser engañosas si no se les interpreta apropiadamente. En primer lugar, nuestro marco conceptual parece ser más sensible a ciertos tipos de error de traducción que otros procedimientos empleados en pruebas internacionales. Por ejemplo, nuestro marco conceptual considera aspectos como la semántica, la representación curricular y el uso del idioma en diferentes contextos sociales y educativos.

En segundo lugar, hay que recordar que nuestro procedimiento de calificación está basado en un modelo deficitario; es decir, está diseñado para detectar errores, 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 prácticamente imposible. De acuerdo con tales razonamientos, es el efecto aditivo de muchos errores o la presencia de errores críticos, lo que hace que un ítem sea cuestionable en el idioma destinatario.

En tercer lugar, los resultados no solamente hablan de la baja calidad de la traducción mexicana de la prueba, sino también de la pobreza de los mecanismos de revisión de traducción. Veinticuatro ítems identificados por los revisores como ítems de *bandera roja* (más del siete por ciento del total de ítems revisados) son demasiados tratándose de la versión final de una prueba. Sorprendente e inexplicablemente, esos ítems de *bandera roja* pasaron por todos los filtros evaluativos usados en la comparación internacional TIMSS-1995. Uno de los informes de TIMSS-1995 incluye un capítulo sobre los procedimientos empleados para revisar la calidad de las traducciones de la prueba (Mullis, Kelly y Haley, 1996). Según dicho informe, ninguno de los ítems utilizados por México fue identificado como *problemático* (¡!).

La gran discrepancia entre ese informe y nuestros resultados pone de manifiesto la necesidad de desarrollar metodologías más robustas para la traducción de pruebas y para la revisión de las traducciones. En tanto las organizaciones internacionales no generen lineamientos de traducción y de revisión más rigurosos, los países deben emplear sus propios procedimientos de revisión y de traducción, además de cumplir con los lineamientos de

traducción acordados por los países participantes. Nuestro marco conceptual puede ser base para garantizar la validez de la traducción de las pruebas.

Deseamos concluir con un comentario final: el costo no debería ser un obstáculo para garantizar el uso de procedimientos robustos para la traducción de pruebas. El uso de paneles de expertos, como el descrito en este trabajo, representa una opción de bajo costo y alta efectividad para avaluar las traducciones de las pruebas internacionales antes de su aplicación.

REFERENCIAS

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

Backhoff, E., y Solano-Flores, G. (2003). *Tercer Estudio Internacional de Matemáticas y Ciencias (TIMSS): Resultados de México en 1995 y 2000*. México, D.F.: INEE.

Baxter, G.P., Shavelson, R.J., Goldman, S.R. y Pine, J. (1992). *Evaluation of procedure-based scoring for hand-on science assessment*. Journal of Educational Measurement, 29(1), 1-17.

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*. PISA Technical Report, 42-54.

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

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

Messick, S. (1995). *Standards of validity and the validity of standards in performance assessment*. Eduactional Measurement: Issues and Practice, 14(4), 5-8.

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

Solano-Flores, G., 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.: INEE.

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, abril). *The metrics of linguistic complexity and the metrics of student performance in the testing of english language learners*. Symposium paper presented at the 2003 Annual Meeting of the American Evaluation Research Association. Chicago, IL.

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

Van de Vijver, F. y Poortinga, Y.H. (1997). *Toward an integrated análisis of bias in cross-cultural assessment*. European Journal of Psychological Assessment, 13(1), 29-37.

Wise, L.K., Hauser, R.M., Mitchell, K.J. y Feuer, M.J. (1999). *Evaluation of the voluntary national tests*: Phase I. Washington, D.C.: National Academy Press.

Nota de los autores

Una versión preliminar de este artículo se presentó originalmente en el congreso anual de la *National Council on Measurement in Education*. Montreal, Quebec, Canada, April 12-14, 2005. 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.

Correos electrónicos de los autores:

Guillermo Solano-Flores: Guillermo.Solano@colorado.edu
Eduardo Backhoff: backhoff@inee.edu.mx
Luis Ángel Contreras-Niño: angel@uabc.mx

Títulos de esta Colección:

1.- *Los resultados de las pruebas de PISA*
Martínez Rizo, Felipe.

2.- *Factores externos e internos a las escuelas que influyen en el logro académico de los estudiantes de nivel primaria en México, 1998-2002*
Muñoz I, Carlos *et al*.

3.- *Contextualización sociocultural de las escuelas de la muestra de estándares nacionales (1998-2002).*
_____ *Determinantes sociales y organizacionales del aprendizaje en la Educación Primaria de México. Un análisis de tres niveles*.
_____ *Perfil de las escuelas primarias eficaces de México (2001)*
Fernández, Tabaré.

4.- *Tercer Estudio Internacional de Matemáticas y Ciencias Naturales (TIMSS): resultados de México en 1995 y 2000*
Backhoff, Eduardo y G. Solano.

5.- *Estudio Sobre las Desigualdades Educativas en México: la Incidencia de la Escuela en el Desempeño Académico de los Alumnos y el rol de los Docentes*
Treviño, Ernesto y G. Treviño.

6.- *Evaluación inicial de los procesos de calibración y equiparación de las pruebas del proyecto de estándares nacionales.*
Magriñá, Antonio.

7.- *Factores asociados al aprendizaje del lenguaje y las matemáticas en 13 estados de México*.
Cervini, Rubén.

8.- *Acciones de Evaluación en las Instituciones Públicas de Educación Media Superior*
Antonio, Rocío.

9.- *El proyecto PISA: su aplicación en México*
Vidal, Rafael, *et. al.*

10.- *La Comparabilidad de los Resultados de las Evaluaciones. Importancia y Dificultad de la Equiparación*
Martínez Rizo, Felipe.

11.- *Marginación y rezago educativo en México.*
Ávila, José Luis.

12.- *Pruebas y rendición de cuentas*
Martínez Rizo, Felipe.

13.- *Panorama Educativo 2004. La edición 2004 de* Education at a Glance de la OCDE
Martínez Rizo, Felipe.

14.- *El Diseño de Sistemas de Indicadores Educativos: Consideraciones Teórico-Metodológicas*
Martínez Rizo, Felipe.

15.- *La telesecundaria mexicana: desarrollo y problemática actual*
Martínez Rizo, Felipe.

16.- *Sobre la difusión de resultados por escuela*
Martínez Rizo, Felipe.

17.- *Exámenes de la Calidad y el Logro Educativos (Excale):nueva generación de pruebas nacionales*
Backhoff Escudero, Eduardo.

18.- La Educación Mexicana en *Education at a Glance 2005*
Martínez Rizo, Felipe.

Los Cuadernos de Investigación

Durante la década pasada nacieron y se fortalecieron en Latinoamérica los sistemas nacionales de evaluación educativa ante el desafío de contar con información apropiada sobre los conocimientos y competencias que los estudiantes adquieren en sus escuelas.

La mayoría de estos sistemas de evaluación han venido justificando su creación bajo la premisa de contribuir a la mejora de la calidad y equidad del sistema educativo. Así, el propósito fundamental es utilizar la información que arroja la evaluación, para rediseñar o ajustar políticas, planes, programas y prácticas pedagógicas y de gestión escolar.

Hacer bien la evaluación y difundirla suficientemente para que sus resultados sean utilizados en la toma de decisiones apropiadas, es de gran valor para el mejoramiento de las escuelas. Esto es lo que el Instituto Nacional para la Evaluación de la Educación (INEE) aporta a la educación mexicana.

Es necesario el diálogo entre personas e instituciones de diferentes sectores y de distintos países, capaces de desarrollar pensamiento crítico, promover debates, crear y fortalecer propuestas innovadoras, y unir esfuerzos encaminados a la búsqueda de una educación de calidad. A tal empeño contribuye también el INEE con la publicación de esta *Colección de Cuadernos de Investigación*, integrada por estudios técnicos, en los cuales convergen sustantivas aportaciones de especialistas en evaluación educativa de México y otros países.

Evaluar para transformar

inee@inee.edu.mx **www.inee.edu.mx**

José María Velasco 101, Col. San José Insurgentes, CP 03900, México, D.F. Tels. 5482 0900, Exts. 1034 y 1050