¿Modifica la traducción la exigencia en Ciencias? Efectos de las lenguas en las evaluaciones de Ciencias del Programa del Diploma del Bachillerato Internacional

Resumen ejecutivo

Joshua McGrane, Yasmine El Masri, Heather Kayton, Robert Woore y Kit Double

Centro de Evaluación Educativa de la Universidad de Oxford (OUCEA)

Departamento de Educación, Universidad de Oxford

Resumen ejecutivo

Contexto

Los programas del Bachillerato Internacional se imparten en más de 153 países de todo el mundo. Aunque el inglés es la lengua de instrucción en la mayor parte de los Colegios del Mundo del IB, muchos de ellos imparten las clases en español o francés. Por esta razón, las evaluaciones del Programa del Diploma (PD) del IB se realizan en español, francés e inglés, y algunas también en más de 75 idiomas adicionales. Las complejidades asociadas a la traducción de las evaluaciones y el compromiso del IB de ofrecer evaluaciones multilingües equivalentes plantean preguntas fundamentales. Se cuestiona si las versiones en las tres lenguas principales son equiparables en el plano lingüístico y, por extensión, en el nivel de exigencia cognitiva. Además, se plantean preguntas más generales en cuanto a la dificultad empírica de sus elementos. Estas preguntas son importantes, ya que tienen una relación directa con el grado de equiparabilidad de las puntuaciones del IB. La equivalencia de las puntuaciones obtenidas en las versiones en diferentes idiomas tiene repercusiones considerables en la equidad del acceso a la educación superior, especialmente en una titulación internacional como es el PD. Este estudio, que constituye una primera aproximación a la materia, analiza la equiparabilidad de las evaluaciones de Ciencias del PD de 2019 en español, francés e inglés.

Alcance y objetivos

El enfoque, los métodos y los hallazgos de este estudio de investigación se vieron motivados por los siguientes objetivos generales:

- Examinar las tendencias y los patrones de las diferencias observadas en el desempeño de los alumnos entre la lengua original (inglés) y las traducidas (español y francés) en cada una de las preguntas de las evaluaciones de Ciencias del PD en 2019, a fin de analizar si la exigencia de las preguntas era distinta en los distintos idiomas, y en qué grado.
- Investigar en qué medida estas diferencias observadas se debían a una modificación de las exigencias lingüísticas y cognitivas durante el proceso de traducción de las preguntas al español y al francés.
- Elaborar un modelo que permita explicar las diferencias observadas a partir de efectos lingüísticos, traductológicos y no traductológicos, utilizando métodos cualitativos y cuantitativos.
- Proponer mejoras a los procesos de traducción del IB basadas en los hallazgos de la investigación.

La investigación se llevó a cabo en tres fases:

- En la primera fase, se aplicaron técnicas cuantitativas a los datos de las evaluaciones de Ciencias del PD correspondientes a 2019 para determinar si existían diferencias sistemáticas en la exigencia

de las preguntas (a partir de su dificultad empírica en los diferentes idiomas), cuál era la magnitud de estas diferencias y si favorecían a la lengua origen o a las versiones traducidas.

- En la segunda fase, se profundizó en estos resultados mediante la evaluación por parte de revisores bilingües expertos de un subconjunto de las preguntas en las que se habían detectado diferencias sistemáticas de exigencia. El objetivo era analizar si las versiones en la lengua original y en las de destino discrepaban en determinados criterios lingüísticos y traductológicos clave.
- En la tercera fase, partiendo de los resultados de las dos primeras, se desarrolló un modelo explicativo que permitiera evaluar si las diferencias lingüísticas y traductológicas entre las diferentes versiones de las preguntas estaban relacionadas en un grado significativo con los diferentes niveles de exigencia de las versiones.

Enfoque metodológico

El estudio aplicó métodos de vanguardia en sus tres fases de investigación:

- En la primera fase se evaluaron las diferencias de exigencia entre las preguntas utilizando un campo de la modelización psicométrica conocido como teoría del rasgo latente, en concreto el modelo logit multinomial de coeficientes aleatorios. Los datos analizados en esta fase fueron las respuestas que dieron los alumnos en las evaluaciones de Ciencias del PD de 2019 en español, francés e inglés para las asignaturas de Física, Química y Biología del Nivel Medio (NM) y el Nivel Superior (NS). Los análisis se llevaron a cabo de forma independiente para cada combinación de asignatura y nivel, y para cada uno de los pares de idiomas (inglés-francés e inglés-español); por tanto, en total se realizaron 12 análisis. Las posibles diferencias de exigencia se analizaron utilizando una técnica llamada funcionamiento diferencial de los ítems (DIF, por su siglas en inglés). Esta técnica permitió comparar estadísticamente los resultados de los alumnos evaluados en los diferentes idiomas para cuantificar las posibles diferencias de exigencia entre las versiones en español, francé e inglés de las preguntas. Se adaptaron tres modelos concretos a los datos de las respuestas: uno que asumía que no había DIF entre los grupos de idiomas distintos, y dos que asumían que sí existía un DIF entre los grupos e incluían un parámetro específico para cada grupo de idiomas, así como un término de interacción entre el grupo y el parámetro del modelo asociado a la dificultad del ítem (es decir, la pregunta). En aquellos casos en los que el modelo con DIF se adaptaba mejor a los datos de las respuestas, este término de interacción del modelo estimaba la magnitud del DIF (ninguno, pequeño, moderado o grande) a nivel de pregunta, además de indicar si beneficiaba a los grupos que utilizaban el idioma original o a los de las traducciones. A continuación, estas estimaciones del DIF se compararon con otras propiedades psicométricas de las preguntas para cada asignatura del PD y se eligió un subconjunto de ellas para su análisis en las siguientes dos fases del estudio.

- La segunda fase consistió en una revisión cualitativa por expertos de aquellas preguntas para las que se había identificado un DIF pequeño, moderado o grande en la fase 1. Puesto que hubieran sido necesarios demasiados recursos para revisar todas las preguntas, se seleccionó un subconjunto de ellas pertenecientes a tres asignaturas (Física NM, Química NS y Biología NM) a partir de varios criterios, como que abarcaran los diferentes tipos de preguntas (de opción múltiple y de respuesta desarrollada) y que hubiera un equilibrio entre las preguntas que favorecían al grupo correspondiente al idioma original y las que lo perjudicaban. En colaboración con el IB, se seleccionaron diez revisores expertos bilingües y trilingües (dos para cada combinación de idioma y asignatura) para evaluar en qué grado las versiones traducidas de las preguntas seleccionadas eran equiparables a la versión original en inglés. Los expertos revisaron las preguntas empleando una encuesta de 14 a 15 puntos especialmente elaborada por los investigadores a partir de un prestigioso marco de traducción y verificación. Los puntos de la encuesta abordaban ocho criterios clave de este marco y de los procesos internos del IB en materia de la veracidad de las traducciones respecto a la versión original. Se calculó la fiabilidad entre calificadores de las valoraciones realizadas por los expertos en las encuestas y se recopilaron las respuestas con el objetivo de evaluar si las preguntas de Ciencias del PD con un DIF entre idiomas detectado en la fase 1 mostraban diferencias lingüísticas y de traducción congruentes con dicho DIF. Por otra parte, estas variables de la revisión por expertos se utilizaron en la tercera fase del estudio en la elaboración de un modelo explicativo del DIF.
- La tercera fase de la investigación consistió en crear un modelo explicativo del DIF entre idiomas para las mismas tres asignaturas analizadas en la fase 2. Esto se llevó a cabo en dos etapas. En la primera etapa, únicamente se modelizó el subconjunto de preguntas seleccionadas para la fase 2 a fin de incluir en el modelo las variables de la revisión por expertos. Además de estas variables, en la fase 3 se incluyeron también índices para las preguntas provenientes de un campo de la lingüística computacional conocido como procesamiento del lenguaje natural (PLN), así como características no lingüísticas, como la asignatura, la prueba (como equivalente del tipo de ítem) y el idioma al que se había traducido la pregunta. Se calcularon los índices de PLN para cada combinación de asignatura, idioma y pregunta utilizando un marco de procesamiento de textos multilingüe y de código abierto llamado ReaderBench. En estudios anteriores realizados con estos índices se había observado que están asociados con la complejidad textual, por lo que se esperaba que las posibles diferencias entre estos índices en las versiones en la lengua original y traducidas contribuyeran a explicar el DIF entre idiomas. En la segunda etapa, se incluyeron todas las preguntas con estimaciones de DIF pertenecientes a las tres asignaturas; por tanto, se eliminaron del modelo las variables de la revisión por expertos y se centró la atención en la capacidad de los

índices del PLN para explicar los resultados. Esta segunda etapa se llevó a cabo porque el enfoque de modelización cuantitativa de esta fase requería utilizar grandes cantidades de datos para que las estimaciones fueran sólidas; también por este motivo, en el análisis se incluyeron además preguntas de 2018 pertenecientes a las mismas tres asignaturas, junto con sus estimaciones de DIF entre idiomas. Los modelos explicativos empleados en esta fase proceden del aprendizaje automático. En concreto, se aplicaron tres modelos (regresión paso a paso, regresión Elastic Net y regresión de bosques aleatorios), puesto que cada uno de ellos tiene ventajas e inconvenientes, como poderse interpretar de forma más transparente (regresión paso a paso y Elastic Net) o ser más opacos, pero más flexibles en cuanto a las relaciones complejas y no lineales establecidas entre las variables de los modelos. En ambas etapas, los modelos se evaluaron en función del error de sus predicciones y su capacidad explicatoria. El modelo que proporcionó un mejor rendimiento se evaluó en función de las variables concretas más importantes del modelo para explicar el DIF entre idiomas. En todos los casos, los modelos se aplicaron utilizando un enfoque de validación cruzada para mejorar la capacidad de generalizar los resultados.

Hallazgos principales

A continuación, se indican los principales hallazgos de la primera fase de la investigación:

- Los análisis mostraron que uno de los modelos con DIF era el que mejor se adaptaba a todas las combinaciones de asignatura, nivel e idioma, lo que puso de manifiesto que existía un DIF entre idiomas en todos los exámenes de Ciencias del PD y en las versiones de las tres lenguas.
- Más positivo para el proceso actual de traducción del IB es que tan solo una pequeña proporción, aunque sustancial, de los elementos de todas las asignaturas presentaron un DIF moderado o grande, y los mayores DIF tendían a ser más prevalentes en las respuestas de desarrollo de las pruebas 2 y 3. En general, las asignaturas de Química eran las que presentaban una mayor proporción de DIF moderados y grandes a nivel de elemento, seguidas por las de Física y, finalmente, por las de Biología.
- Como tendencia general, las preguntas que presentaban un DIF significativo que favorecía a los idiomas traducidos solían ser las más difíciles y menos diferenciadoras; esto ocurría especialmente con las preguntas de opción múltiple. La relación entre las estimaciones de DIF y estas otras propiedades psicométricas de los elementos ofreció indicios de que algunos de los DIF entre idiomas podrían deberse a problemas generales propios de los elementos, más que al propio idioma. En concreto, parte de este DIF podría atribuirse a la selección de respuestas al azar, sobre todo teniendo en cuenta que los alumnos que respondieron a las versiones traducidas tendían, en general, a tener un peor desempeño en las asignaturas.

- Se seleccionó la asignatura de Física NM para su inclusión en las siguientes fases de la investigación, ya que la muestra de Física NS de francés era muy pequeña y, por tanto, las estimaciones del DIF de Física NM eran más sólidas (aunque de menor magnitud en general). Las asignaturas de Química NS y Biología NM se seleccionaron para su inclusión en las siguientes fases porque las magnitudes de DIF observado en ellas eran, en general, mayores que las correspondientes al otro nivel de las mismas asignaturas.

Los siguientes fueron los principales hallazgos de la segunda fase de la investigación:

- Los resultados obtenidos en la revisión por expertos cualitativa de las preguntas fueron muy positivos para el modelo de traducción adoptado actualmente por el IB, ya que se consideró que las versiones traducidas en francés y español y la versión original inglesa eran muy equiparables en la mayor parte de las preguntas. Aparecieron algunas inconsistencias en preguntas concretas, pero en general eran menores y no sistemáticas en cuanto a su magnitud o al grupo favorecido por el DIF. Por ejemplo, en las pruebas de Química NS hubo muchas más preguntas que se clasificaron con un DIF entre idiomas de entre medio a grande, pero los revisores expertos consideraron que las traducciones de dichas preguntas eran más comparables a la versión inglesa.
- De entre los criterios utilizados en la revisión por expertos que sí mostraron una cierta desviación entre la versión original y las traducidas, los que lo hacían en mayor grado y de manera más uniforme eran los criterios de **búsqueda de patrones** y **precisión de la formulación,** aunque en términos absolutos estas desviaciones eran pequeñas.
- En general, los revisores expertos pudieron utilizar con fiabilidad la encuesta desarrollada recientemente para evaluar las posibles diferencias entre la versión original y las traducidas de los elementos. Esta buena fiabilidad proporcionó confianza en el uso de estas variables en la modelización de la fase 3. Sin embargo, algunos criterios mostraron una fiabilidad menor que el umbral del 70 % de forma uniforme en todas las asignaturas e idiomas. Entre estos se encontraban los criterios de formulación y longitud de los sintagmas, por lo que en futuras aplicaciones de esta encuesta debería intentarse mejorar la comprensión estandarizada de su significado.

Los siguientes fueron los principales hallazgos de la tercera fase de la investigación:

- Se obtuvieron resultados variables sobre cómo las diferencias lingüísticas y traductológicas entre el idioma original y las traducciones explicaban las diferencias en dificultad de las versiones. En primer lugar, ninguna de las variables de idioma utilizadas en la revisión por expertos de la fase 2 pudo identificarse como un predictor significativo del DIF entre idiomas, aunque esto era congruente con los resultados descriptivos para estas variables en la fase 2. Es probable que esto se debiera, al menos en parte, a la falta de variación de estas variables en la revisión por expertos.

- Se observó que las diferencias en los índices de complejidad del PLN para las versiones original y traducidas de los elementos explicaban hasta un cierto punto los diferentes grados de DIF observados. El comportamiento del modelo de bosques aleatorios, el que mejor se adaptaba a los datos en ambas etapas, era mejor para el subconjunto menor de elementos de la fase 2, siendo responsable del 11 % de la varianza de la variable de salida del DIF entre idiomas, mientras que era del 4 % para el conjunto mayor de datos (que incluía elementos de 2018 y 2019 correspondientes a tres asignaturas, Física NM, Química NS y Biología NM).
- Las características más importantes del PLN para predecir la variable de salida del DIF entre idiomas en el modelo de regresión de bosques aleatorios se pudieron clasificar en los siguientes tres temas, dados en el orden de importancia general dentro del modelo: elección de las palabras, longitud de las oraciones y complejidad estructural.
    - Los índices de elección de las palabras representan diferentes aspectos de cómo la información nueva o poco conocida presente en el texto podría crear dificultades para los lectores en cualquiera de los idiomas. Cuanto más esperable o predecible es una frase para el lector, más fácil es de entender. Esta información podría darse en forma de palabras, letras, oraciones o incluso puntuación.
    - Los índices asociados a la longitud de las oraciones reflejan diferentes aspectos de cómo, a medida que aumenta la longitud de una frase, mayor es la carga cognitiva asociada con su procesamiento y cómo esto podría afectar al grado en que los lectores son capaces de comprender dicha frase.
    - Los índices de complejidad estructural reflejan cómo las diferentes características de un texto, en cuanto a su gramática y sintaxis, pueden crear diferentes grados de complejidad para el lector.

Recomendaciones

En cada fase de la investigación se obtuvieron recomendaciones específicas. Entre las recomendaciones generales de la fase 1 se incluyen:

- Revisar las preguntas de opción múltiple que muestran un grado diferente de respuestas marcadas al azar para las versiones en los diferentes idiomas a fin de entender qué características de estos elementos podrían provocar un aumento de las respuestas al azar, tanto en general como para idiomas concretos.
- Revisar los elementos que presentaban un DIF entre idiomas medio o grande pertenecientes a las tres asignaturas no analizadas en las fases posteriores del estudio y a exámenes de otros años.

Entre las recomendaciones extraídas de la segunda fase están:

- Asegurarse de que los procesos de traducción y garantía de calidad están estandarizados para cada asignatura y entre diferentes asignaturas.
- Reducir la subjetividad de las evaluaciones introduciendo menos elementos culturales y menos dialectos. Para hacer esto, se crean dos versiones de la evaluación en dos idiomas de partida (por ejemplo, inglés y español) y ambas versiones de partida se utilizan para crear una versión en el idioma de destino (por ejemplo, francés).
- Evaluar el uso de procesos de revisión o de garantía de calidad de las traducciones que permitan que, a partir de los problemas identificados en la versión traducida, se pueda reescribir o comprobar la de origen. Por ejemplo, podría ocurrir que un problema identificado en la versión traducida sea también pertinente para la versión en el idioma de origen y que sea necesario adaptar ambas.
- Revisar los términos de instrucción para asegurarse de que las listas de términos se traduzcan a los idiomas de destino de forma que no suenen extraños ni haya matices de significado entre idiomas.
- Traducir los esquemas de calificación para ampliar la investigación a un rango mayor de asignaturas y evaluar si el no contar con una traducción de los esquemas de calificación está afectando a la validez de los exámenes multilingües.

Entre las recomendaciones generales de la fase 3 están:

- A la hora de elegir las palabras empleadas en las traducciones, debería prestarse especial atención a la frecuencia relativa de las palabras con contenido (es decir, sustantivos, verbos, adjetivos y adverbios).
- En cuanto a la longitud de las oraciones, en el diseño de los elementos y la traducción, analizar siempre si la adición de palabras y sintagmas mejorará la claridad o aumentará la complejidad. Cuando se decida emplear oraciones más largas para mejorar la claridad, esto debería hacerse de forma uniforme en todos los idiomas.
- En el diseño de los elementos, evitar siempre que sea posible las oraciones largas y complejas que contengan varios signos de puntuación. Siempre que sea posible, deberían utilizarse oraciones cortas para mejorar la claridad y reducir la carga cognitiva asociada con el procesamiento de frases largas.
- En el desarrollo de los elementos, tener cuidado con aquellos componentes del discurso que puedan añadir complejidad, como adverbios o adjetivos. En los casos en que estos componentes del discurso se utilicen para aumentar la claridad, debería prestarse una especial atención a la frecuencia relativa con que se utilizan en los diferentes idiomas.

- En cuanto a la estructura de las oraciones, en el diseño de los elementos y la traducción debería tenerse siempre en cuenta si la adición de palabras y sintagmas mejorará la claridad o aumentará la complejidad. Si se decide emplear oraciones largas para mejorar la claridad, debería hacerse de manera uniforme en todos los idiomas.
- El software de análisis de textos puede ayudar a analizar sintácticamente los componentes de las oraciones. Esta información puede permitir comparar la complejidad de la estructura de los textos. En la medida de lo posible, la complejidad relativa de los elementos debería ser equivalente en todos los idiomas.
- Para tener en cuenta estas características de complejidad de los textos en los diferentes idiomas puede ser útil utilizar software de PLN como ReaderBench. El análisis previo de los elementos con software de análisis de textos puede ayudar a identificar si algunos de ellos podrían presentar más dificultades de compresión de lectura en un idioma concreto.
- Combinando las recomendaciones de las fases 2 y 3, una recomendación final general sería elaborar de forma simultánea las versiones en los idiomas de partida y de destino de los exámenes. Esto permitiría resolver las posibles discrepancias entre las versiones identificadas por expertos o durante la revisión con PLN modificando la versión original en inglés y trasladando estos cambios a las traducciones, con lo que las versiones en los diferentes idiomas pasarían a ser más uniformes en términos lingüísticos.

Conclusión

La conclusión general de este estudio de investigación es que las evaluaciones de Ciencias no se vieron afectadas por la traducción. En el estudio de las seis evaluaciones de asignaturas de Ciencias del PD de 2019 se observó un alto grado de equiparabilidad entre las versiones en español, francés e inglés. Parece que los procesos actuales de traducción del IB, que incluyen la traducción directa, una comprobación de la traducción y una revisión final respecto del original, y en los que se utilizan tanto conocimientos sobre traducción como de las evaluaciones del IB, resultan eficaces para elaborar evaluaciones en estos tres idiomas con una dificultad equiparable. Sin embargo, había una cantidad sustancial de elementos de las seis asignaturas de Ciencias del PD que mostraban un DIF entre idiomas moderado o grande, así que está claro que aún podría mejorarse la traducción de los elementos.

La relación sistemática entre las diferencias de dificultad que presentaban los elementos en los diferentes idiomas y otras de sus características psicométricas pusieron de relevancia la conexión existente entre el diseño/comportamiento general de los elementos y los problemas de traducción y, especialmente, la necesidad de seguir analizando por qué motivo en algunos idiomas era más frecuente que los alumnos dieran respuestas al azar. Por otra parte, la revisión por expertos sugirió que la traducción de los elementos podría ser más precisa en cuanto a seguir los patrones de los

elementos y utilizar formulaciones equiparables al transmitir la información en las traducciones. Por último, el análisis mediante PLN de los elementos en los diferentes idiomas mostró innumerables, aunque sutiles, diferencias lingüísticas, relacionadas en cierto grado con el DIF entre idiomas.

El uso de análisis de la complejidad de los textos de los elementos en diferentes idiomas utilizando PLN en combinación con técnicas de modelización de aprendizaje automático a fin de explicar el DIF entre idiomas observado (o la no existencia de él) ha sido una contribución muy innovadora del presente estudio de investigación. Esta metodología ha permitido identificar diferencias lingüísticas en los elementos con DIF que no hubieran podido detectarse mediante métodos convencionales. Este enfoque podría ser más eficaz si se aplicara a asignaturas del PD cuyos exámenes y elementos contengan más texto, de manera que tenga sentido el uso de los índices del PLN relacionados con la cohesión textual y el discurso. A partir de los hallazgos de este estudio, creemos que el uso de estas tecnologías de inteligencia artificial para predecir y explicar DIF entre idiomas seguirá siendo un campo de investigación fructífero que proporcionará información valiosa para diferentes evaluaciones multilingües internacionales.