PROPUESTA TÉCNICA

Fecha: 28-03-2018

Aplicación de la prueba *PISA for Schools* en tableta: Estudio del impacto, validación y ajuste de las puntuaciones

Aplicación de la prueba PISA for Schools en tableta: Estudio del impacto, validación y ajuste de las puntuaciones

¿Están los alumnos de su centro educativo preparados para enfrentarse a los retos del futuro? ¿Pueden analizar, razonar y comunicar sus ideas de manera efectiva? ¿Han desarrollado el conocimiento y las destrezas esenciales para su participación con éxito en las sociedades del siglo XXI? La prueba *PISA for Schools* (PfS) busca dar respuesta a estas preguntas mediante una evaluación a los alumnos que está directamente fundamentada en el conocimiento acumulado por el internacionalmente reconocido Programa para la Evaluación Internacional de los Alumnos (PISA), administrada cada tres años a alumnos y centros de más de 80 países. Como en PISA, la prueba PfS mide el conocimiento aplicado y las competencias en lectura, matemáticas y ciencias de alumnos de quince años. La evaluación no solo pretende determinar si los alumnos saben reproducir el conocimiento adquirido, sino que también examina el grado en el que saben extrapolar lo que han aprendido y el grado en el que saben aplicarlo en situaciones no familiares, tanto dentro como fuera del centro educativo. Sin embargo, mientras que la evaluación internacional PISA pretende proporcionar resultados nacionales agregados para las comparaciones internacionales y proporcionar información para los debates políticos, la evaluación *PISA for Schools* se ha diseñado para proporcionar resultados a nivel de centro a efectos comparativos y de mejora de los centros educativos. PfS contiene 47 preguntas de lectura, 40 de matemáticas y 54 de ciencias. Cada alumno responde a uno de los siete cuadernillos, en un tiempo máximo de 120 minutos. Los alumnos responden además a un cuestionario que permite extraer información sobre las características socioculturales de los alumnos, sus intereses y actitudes hacia el aprendizaje.

La prueba, lanzada en 2012, fue proyectada para ser aplicada en formato papel. Sin embargo muchos centros y organismos demandan cada vez más una aplicación digital, ya sea mediante ordenadores o tabletas. De hecho, los alumnos que participaron en la última edición de PISA (OECD, 2015) realizaron la prueba íntegramente por ordenador.

Las ventajas que ofrece el uso de dispositivos tecnológicos frente a las pruebas impresas tradicionales son claras: reducción de costes asociados a la impresión, distribución, administración y corrección de las pruebas, mayores garantías de seguridad y confidencialidad de las respuestas, y mayor confortabilidad y preferencia del examinado (O´Malley et al., 2005). Aunque los beneficios son claros las pruebas digitales cuentan también con importantes inconvenientes. Los estudios de comparabilidad entre modos de aplicación (papel versus ordenador) han encontrado que las puntuaciones obtenidas bajo un modo de aplicación y otro no pueden ser equiparadas, es decir, no son comparables (Bennet, 2003; Pommerich, 2004; OECD, 2015). La forma de presentación de los ítems, los requerimientos de respuesta o las condicione generales de evaluación podrían explicar las discrepancias en las puntuaciones obtenidas (Kolen, 1999).

El interés generado en la comparabilidad de los modos de aplicación queda claramente recogido en las normas y directrices profesionales desde 1986 (American Psychological Association, APA,1986). Más recientemente, la última edición de los *Standards for Educational and Psychological Testing* (*American Educational Research Association*, [AERA]; APA; *National Council on Measurement in Education*, [NCME]) incluye una referencia explícita al problema de la equivalencia. Los estándares 9.7 y 9.9 reclaman evidencias empíricas de la fiabilidad y validez, a fin de que las puntuaciones de versiones distintas puedan ser utilizadas indistintamente.

Los responsables de Knotion® en México expresaron su interés de emplear la prueba PfS en formato tablet para evaluar el rendimiento y actitudes de los alumnos que están adscritos al programa. Knotion® es un ecosistema de contenidos transdisciplinarios estratégicamente creados en un universo digital, para apoyar y encauzar el aprendizaje significativo de los alumnos, mientras que fomenta el desarrollo de competencias que todo ciudadano global debe poseer. Este planteamiento, no basado en el curriculum educativo sino más bien en las competencias, encaja perfectamente con la visión que PISA tiene sobre los estudiantes del futuro.

El presente documento recoge procesos y procedimientos encaminados a estudiar la comparabilidad de las puntuaciones de la prueba PfS en función del modo de aplicación (papel o tableta), desde un nivel de análisis del ítem y un nivel de análisis de prueba total. Se utilizarán métodos de estimación clásicos y otros basados en la Teoría de Respuesta al Ítem (TRI). Asimismo, y solo en caso de encontrar no equivalencia de puntuaciones, se propondrá un método de ajuste que permita equiparar las puntuaciones de los alumnos de la muestra estudiada.
A continuación se presenta a modo de resumen un índice.

Método

Participantes: sobre la muestra de análisis del estudio. Se prevén ligeros cambios en el número final de alumnos y centros participantes.

Instrumento: sobre la prueba PfS .

Procedimiento: sobre el procedimiento de recogida de datos en las muestras de análisis de modo de aplicación papel (MP) y modo de aplicación tableta (MT). Pendiente de concretar las especificaciones técnicas de las tabletas que se van a emplear en este estudio.

Análisis

Análisis descriptivos: análisis de equivalencia de los grupos, y descriptivos de ambas muestras. Medias, porcentajes y distribuciones. Tasas de respuesta, omisiones y participación.

Índices clásicos: puntuación directa, índice de dificultad, índice de discriminación.

Análisis de la fiabilidad: índices de consistencia interna (TCT) y función de información (TRI) para los grupos MP y MT.

Análisis de la validez: invarianza factorial (análisis factorial multigrupo), invarianza parámetros ítems, funcionamiento diferencial de los ítems

Ajuste del modo de aplicación: coeficientes beta para cada dominio de evaluación (lectura, matemáticas y ciencias) para cada modo de aplicación.

Ajuste de edad: coeficientes beta para cada dominio de evaluación (lectura, matemáticas y ciencias) en función de la edad (meses de desviación respecto a la edad de los alumnos en el estudio PISA).

Método

Participantes

La muestra de estudio está configurada por 1698 alumnos procedentes de 26 centros educativos adscritos al programa Knotion® en México, todos de titularidad privada. La fase de recogida de datos, es decir, la fase de aplicación de la prueba PfS se prevé durante la tercera semana de mayo y segunda semana de junio de 2018.
El estudio PISA y PfS basan el criterio de elegibilidad en la edad de los alumnos, de forma que éstos tengan aproximadamente 15 años (15 años y 3 meses a 16 años y 2 meses) y en el curso académico (grade 7th o superior). Sin embargo la muestra de alumnos elegibles que participan en este estudio está compuesta por alumnos que cursan Educación Secundaria, con el 61% cursando 3º Educación Secundaria (México), un 30% en 2º, y un 9% de los alumnos en 1º de Educación Secundaria. De acuerdo a los datos proporcionados, un 60% de los alumnos tienen 15 años, un 14% tienen 16 años y un 12% de los alumnos tienen 14 años en el momento de aplicación de la prueba (junio 2018). Un 14% no tiene aún edad identificada aunque está previsto completar esta información en las próximas semanas.

Centro Educativo	Edad de los alumnos (junio 2018)				
	14	15	16	NA	Total
Centro Escolar Balam	15,9%	56,3%	27,8%	0,0%	100,0%
Colegio Agnes Gonxha	9,1%	75,0%	11,4%	4,5%	100,0%
Colegio Bilingüe María Fernanda	15,6%	84,4%	0,0%	0,0%	100,0%
Colegio Franco Inglés	14,3%	60,2%	23,5%	2,0%	100,0%
Colegio Grecia	6,9%	82,8%	6,9%	3,4%	100,0%
Colegio Inglés (Saltillo)	2,5%	47,5%	37,5%	12,5%	100,0%
Colegio Ingles Americano	10,2%	62,5%	27,3%	0,0%	100,0%
Colegio México de Tehuacán	14,5%	73,7%	10,5%	1,3%	100,0%
Colegio México Roma	13,6%	52,5%	7,6%	26,3%	100,0%
Colegio Mirasierra	13,3%	66,7%	4,4%	15,6%	100,0%
Colegio Newland (Campus Juriquilla)	14,9%	70,1%	14,9%	0,0%	100,0%
Colegio Newland School (Campus Piamonte)	22,7%	68,2%	9,1%	0,0%	100,0%
Colegio Simón Bolívar Tepic	14,5%	76,8%	8,7%	0,0%	100,0%
Colegio Ypsilanti	11,5%	72,1%	16,4%	0,0%	100,0%
Cumbres Comunidad Educativa	17,9%	60,7%	17,9%	3,6%	100,0%
Educare	13,0%	66,7%	15,9%	4,3%	100,0%
Escuela Libertad	7,3%	43,9%	7,3%	41,5%	100,0%
Instituto Bilingüe Green Hills	7,4%	77,8%	11,1%	3,7%	100,0%
Instituto Bilingüe Ovidio Decroly	4,7%	15,9%	0,9%	78,5%	100,0%
Instituto J. Francisco Rodríguez	22,4%	62,7%	14,9%	0,0%	100,0%
Instituto México Inglés	14,0%	71,9%	14,0%	0,0%	100,0%
Instituto Rosner	7,8%	47,1%	9,8%	35,3%	100,0%
Instituto Santiago - Solecito	4,9%	61,7%	17,3%	16,0%	100,0%
Instituto Vélez	10,4%	64,6%	14,6%	10,4%	100,0%
Liceo los Cabos	10,8%	58,1%	8,1%	23,0%	100,0%
Varmond School Tres Marías	11,8%	48,4%	6,5%	33,3%	100,0%
Total general	**12,0%**	**59,8%**	**14,2%**	**14,1%**	**100,0%**

Además de los 26 centros descritos, existen doce centros adicionales Knotion® que por ser centros con un número de alumnos reducido no se han considerado en la propuesta inicial. Sin embargo estos centros pueden contribuir con sus alumnos para el análisis de los distintos indicadores de validación de la prueba PfS en tabletas y para el cálculo del informe de grupo "Centros Knotion ®". No obstante estos centros no recibirán el informe PfS centro individualizado, ya que el tamaño de la muestra a nivel de centro no permite calcular indicadores de rendimiento ni de contexto fiables y válidos desde el punto de vista estadístico. Queda a decisión de Knotion® incluir estos centros en la muestra de estudio.

En cada centro se formarán dos grupos: modo de aplicación papel (MP) y modo de aplicación tableta (MT). Para asegurar la equivalencia de los grupos la asignación de alumnos a uno y otro grupo se llevarán a cabo a través de dos tipos de estrategias.
En primer lugar, y de forma previa a la recogida de datos, se realizará muestreo aleatorio estratificado para la asignación de los alumnos a los grupos, valorando los estratos de curso (1º, 2º y 3º de secundaria) y género (hombre y mujer). La proporción de alumnos en cada uno de los grupos se mantendrá idéntica, de forma que en cada centro el grupo MP estará formado por el 50% de los alumnos y el grupo MT estará formado por el 50% de alumnos restantes.
Así, tanto el grupo MP y MT contarán con 849 alumnos respectivamente (datos inciales).

En este estudio no se prevé el uso de pesos de estratificación, por lo que la configuración final de los grupos estará en función del tamaño de los centros. Por ejemplo, se prevé que los centros educativos de mayor tamaño contribuyan en mayor medida en la configuración de la muestra que los centros de menor tamaño. Por ejemplo, el centro con más alumnos de la muestra aportará de manera aproximada un 8% de alumnos, mientras que el centro más pequeño estará representado con un 2% de alumnos aproximadamente sobre el total de la muestra.

De manera posterior a la recogida de datos, y para valorar que los alumnos de los grupos MP y MT tengan características sociodemográficas equivalentes, se prevé realizar pruebas de homogeneidad (X^2) o pruebas de contraste de medias (t de student). Para cuantificar el impacto del modo de aplicación es fundamental asegurar y comprobar la equivalencia de los grupos, de manera que las diferencias observadas puedan ser atribuidas al efecto del modo de aplicación, y no a las diferencias en las características personales y sociales de los alumnos evaluados.

Las variables analizadas procederán del cuestionario del alumno, que se administra después de la prueba de rendimiento PfS. Entre otras variables se estudiará la equivalencia de los grupos MP y MT en las variables Edad (15-ST004), Curso (15-ST002), Estudios de la madre (15-ST005), Estudios del padre (15-ST007), Idioma hablado en casa (15-ST022), Recursos culturales en el hogar (15-ST011) o Recursos tecnológicos en casa (15-ST012). En función de los datos recogidos otras variables pueden ser añadidas o sustituidas.

Instrumento

La prueba PfS consta de siete cuadernillos con preguntas sobre lectura, matemáticas y ciencias, más un cuestionario del alumno al que todos los alumnos responden el día de la prueba. El director o un miembro del equipo directivo del centro también debe responder a un cuestionario del centro.
El número total de ítems al que responden los alumnos son 141: 47 de lectura, 40 de matemáticas y 54 de ciencias. Los ítems se presentan agrupados en siete clusters, con dos clusters de lectura (R), dos de matemáticas (M), dos de ciencias (S) y un cluster mixto (RMS), que agrupa ítems de lectura, matemáticas y ciencias. Sin embargo, y debido al diseño matricial de ítems (*balanced incomplete block*, BIB) que sigue PISA y PfS, los alumnos responden a tres clusters de ítems (media de 60 ítems) en un tiempo máximo de 120 minutos.

Tabla 1. Distribución de los clusters en la prueba PfS

Cuadernillo	Cluster 1	Cluster 1	Cluster 3
1	R1	RMS	M1
2	RMS	M2	S2
3	M2	M1	R2
4	M1	S2	S1
5	S2	R2	R1
6	R2	S1	RMS
7	S1	R1	M2

Las preguntas de la prueba varían en cuanto a su formato. Alrededor de la mitad requieren que los alumnos construyan sus propias respuestas. Algunas requieren una respuesta breve, mientras que otras permiten respuestas individuales diferentes y, a veces, una evaluación de la justificación de los puntos de vista de los alumnos. La otra mitad son preguntas de respuesta múltiple en las que los alumnos hacen una elección entre cuatro o cinco alternativas, o bien eligen una de dos posibles respuestas ("sí" o "no") a una serie de proposiciones o afirmaciones. La tabla 2 presenta el porcentaje de ítems en función del tipo de respuesta que exigen.

Tabla 2. Ítems de la prueba PfS según tipo de respuesta y domino

Tipo de respuesta	Lectura	Matemáticas	Ciencias	Total
Elección múltiple simple	38 %	28 %	33 %	34 %
Elección múltiple compleja	15 %	7 %	30 %	18 %
Respuesta construida cerrada	9 %	47 %	0 %	22 %
Respuesta construida abierta	36 %	18 %	37 %	26 %
Total	100%	100%	100%	100%

Tanto para las respuestas de los alumnos del grupo MP y del grupo MT la corrección de los ítems de respuesta construida se llevará a cabo por codificadores previamente formados y que cuenten con amplia experiencia en la corrección de ítems y preguntas de rendimiento. Los correctores deben ser docentes y expertos en el dominio que corrijan. La codificación de las respuestas de elección manual se realizará de forma manual en el grupo MP y de forma automatizada en el grupo MT.

Se contempla la posibilidad de excluir del estudio algún ítem que no permita una adaptación sencilla y directa al formato de aplicación en tableta. Es el caso de la unidad PR6026- Bus Timetable, que podría plantear ciertas dificultades técnicas para ser insertado en la plataforma de tableta. Los ítems que sean finalmente excluidos han de estar previamente identificados y justificados el motivo de exclusión. Los ítems serán excluidos del análisis, pero no necesariamente de la aplicación.

Procedimiento

Se prevé aplicar la prueba en una ventana temporal de dos semanas como máximo, en concreto desde el día 14 de mayo y hasta el día 14 de junio de 2018.
Los análisis estadísticos y los resultados de este estudio dependen directamente de la calidad de los datos recogidos, por lo que se recomienda una planificación en detalle de la fase de recogida de datos. Para facilitar la aplicación de la prueba cada centro Knotion® debe designar un coordinador de centro que facilite las tareas de seguimiento y gestión de la prueba.

La aplicación de la prueba PfS en los grupos MP y MT debe realizarse estrictamente bajo las mismas condiciones. Esto significa que ambos grupos deben contestar la prueba en aulas o espacios similares, en el mismo horario (primera o segunda hora lectiva) y con similares condiciones de confortabilidad. En concreto debe prestarse especial atención a la práctica y la familiaridad que tengan los alumnos en el uso manejo de las tabletas. Todos ellos deben tener suficiente experiencia en el uso de este dispositivo. Las condiciones ergonómicas del grupo MT deben considerarse especialmente, teniendo en cuenta que dado el tiempo de aplicación de la prueba pueden aparecer factores de fatiga en los alumnos. Con anterioridad a la aplicación de la prueba se recomienda proporcionar a los alumnos algunas sugerencias para mejorar la confortabilidad.

Cada grupo, tanto MP y MT, debe contar con un aplicador previamente formado por Knotion®. Los aplicadores serán preferentemente profesores, pero en ningún caso podrán ser los profesores de los alumnos. Se recomienda que los aplicadores sea personal totalmente externo al centro educativo. Los aplicadores proporcionarán a los alumnos, según el grupo que corresponda, las instrucciones para realizar la prueba, que serán levemente distintas en el grupo MT con el fin de adaptarlas al modo de uso e interacción con la tableta.

Modo papel: La prueba PfS será aplicada en el grupo MP en un cuadernillo (de los siete existentes) por alumno. Los cuadernillos tendrán un formato de tamaño A4 y serán impresos en blanco y negro. De media, cada cuadernillo tiene 60 páginas. Las adaptaciones idiomáticas para la muestra de alumnos mexicanos han sido supervisadas y están pendiente de aprobación por parte de Capstan y la OECD [aproximadamente 15 abril de 2018].

Modo tableta: La prueba PfS será aplicada en el grupo MT mediante tableta. Cada alumno recibirá una clave personal que le permitirá acceder al cuadernillo asignado. Las tabletas serán proporcionadas por Knotion® con las siguientes especificaciones técnicas: [añadir especificaciones técnicas tabletas]. Todos los alumnos usarán el teclado virtual interno de las tabletas. Como en el caso de la prueba PfS en formato papel, los alumnos del grupo MT tendrán disponibles durante todo el tiempo las instrucciones de la prueba, así como un formulario que podrán consultar en cualquier momento durante el transcurso de la prueba.

Con anterioridad al día de aplicación de la prueba (idealmente 7 días antes de la prueba) cada centro participante recibirá un Formulario de Seguimiento del Alumnado (FSA), donde aparecerán los alumnos asignados a los grupos MP y MT, así como sus datos de nacimiento, curso donde están matriculados, programas específicos o si son alumnos con necesidades educativas especiales. Cada sesión (cada grupo MP y PT) contará con un FSA que permitirá a los aplicadores identificar a los alumnos y un control estricto de la asistencia.

La administración de la prueba PfS se estructurará en dos sesiones, la sesión de la prueba de rendimiento y la sesión del cuestionario de contexto. La primera sesión estará dedicada a la primera parte de la prueba, es decir, los primeros 60 minutos del tiempo de aplicación. Agotados los primeros 60 minutos se concederá a los alumnos un tiempo de descanso no superior a 5 minutos, y tras éstos se reanudará la prueba durante otros 60 minutos más. Cumplidos los 120 minutos de aplicación los alumnos podrán disponer de un descanso más extenso, entre 15 y 30 minutos. La segunda sesión de la prueba PfS se dedicará a la recogida de información contextual de los alumnos. Los aplicadores deben prestar especial atención en esta fase para garantizar la máxima información posible.

Tabla 3. Estructura de la prueba PfS

Estructura de la aplicación PfS	Tiempo
Primera sesión	
Primera parte prueba	60 minutos
Descanso	5 minutos
Segunda parte prueba	60 minutos
Segunda sesión	
Cuestionario de contexto	35 minutos

Análisis

Análisis descriptivo

El estudio del impacto del modo de aplicación se iniciará con un estudio psicométrico clásico. La Teoría Clásica de los Test (TCT) ofrece un marco familiar para el análisis preliminar de la calidad de los ítems. Aunque los ítems de PfS ya han sido estudiados y validados en su versión internacional, es imprescindible valorar el funcionamiento de los ítems en la muestra de estudio, especialmente en la muestra de alumnos que responden a la prueba en formato tableta. Los análisis que se planean y describen a continuación son dependientes de la calidad de los datos recogidos, por lo que pueden ser modificados o sustituidos por otros si se considera conveniente. También están planteados en función de la muestra entregada. Una reducción significativa en el número final de participantes en el estudio puede imposibilitar total o parcialmente la realización de los análisis descritos. El análisis bajo la perspectiva clásica incluirá los siguientes cálculos, separados para cada grupo, MP y MP:

a) Frecuencias de respuesta: para estudiar los patrones de respuesta de los alumnos se presentarán los porcentajes de respuesta para cada ítem,

b) Puntuación directa: la puntuación directa obtenida en cada dominio, y sus diferencias entre los grupos;

c) Índice de dificultad como el cociente entre el número de alumnos de cada grupo que aciertan el ítem en cuestión y el número total de alumnos que lo han respondido en cada grupo. Se presentaran los índices de dificultad de cada ítem y para cada dominio de evaluación (Lectura, Matemáticas y Ciencias). Los ítems con índices de dificultad extremos serán propuestos para revisión;

d) Índice de discriminación como la correlación entre la puntuación del ítem y la puntuación total en la prueba, para cada ítem y para cada dominio. Cuando sea pertinente se analizara además el efecto del género para valorar posibles interacciones entre género y modo de aplicación;

e) Tasas de respuesta. Las diferencias se estudiarán mediante pruebas de significación estadística y de tamaño del efecto.

Análisis de la fiabilidad

Siguiendo la recomendación de los *Standars* y para estudiar la equivalencia en la precisión (fiabilidad) de la prueba PfS en formato papel y de la prueba PfS en formato tableta, se calcularán los índices clásicos de consistencia interna para cada versión, y la diferencia entre los índices obtenidos se valorará con el estadístico *W* de Feldt (Feldt, 1969). El índice de consistencia interna por el que se evaluará la fiabilidad será el coeficiente α (Cronbach, 1951), según

$$\alpha = \frac{J}{J-1}\left(1 - \frac{\sum S_{X_j}^2}{S_X^2}\right)$$

donde J es el número de ítems y $\sum S_{X_j}^2$ es la suma de las varianzas de los ítems.

Se espera obtener valores adecuados y semejantes del coeficiente de alfa de Cronbach para las muestras los grupos MP y MT. Diferencias significativas entre los índices α del MP y MT podrían revelar problemas de la fiabilidad y validez. Para explorar con más detalle el aspecto de la fiabilidad de las versiones se estimará bajo la perspectiva de la Teoría de Respuesta al Ítem (TRI) la Función de Información (FI), que permitirá evaluar la precisión de los modos MP y MT a lo largo del continuo de competencia lectora, matemática y científica.

Análisis de la validez

Invarianza factorial. Dimensionalidad. La equivalencia en validez de las puntuaciones será explorada en detalle mediante diversos procedimientos y desde una perspectiva global (prueba completa) y otra particular (ítem).

Sobre cada modo de aplicación se prevé realizar un análisis factorial para valorar el supuesto de unidimensionalidad de las pruebas. Comprobar la unidimensionalidad es fundamental para el análisis posterior y la calibración de las respuestas puesto que los modelos de Rasch (Rasch, 1960) y Crédito Parcial (Masters, 1982) se asumen unidimensionales. Se prestará especial atención a la estructura dimensional de la prueba en formato tableta ya que estudios previos han encontrado factores adicionales al rendimiento. El ajuste se valorará mediante índices de ajuste absoluto (SRMSR, Hu y Bentler, 1998; RMSEA, Steiger y Lind, 1980; GFI, Joreskog y Soborm, 1989).

Invarianza Parámetros b_i. La validez a nivel de ítem de cada uno de los modos de aplicación se estudiará mediante el estudio de los parámetros de los ítems, en concreto el parámetro b_i que indica la dificultad del ítem *i* en el contexto de la TRI y que se expresa en la misma escala que la competencia o aptitud del alumno (θ). En primer lugar se estudiará la invarianza de los parámetros b_i en los grupos MP y MT, y para cada dominio. Para obtener las estimaciones de los parámetros en cada grupo las respuestas de los alumnos serán calibradas con el modelo *logit* multinomial de coeficientes mixtos (Adams, Wu y Wilson, 1997), que combina un modelo de respuesta al ítem con el modelo poblacional. Los modelos TRI empleados será el de Rasch para los ítems dicotómicos y el modelo de Masters para los politómicos. Bajo este modelo la probabilidad de que un estudiante con un nivel de aptitud θ obtenga una puntuación de *j* en el ítem *i* se expresa de la siguiente manera:

$$P_{ij}(\theta) = P\left(X_{ij} = 1 \middle| \theta, \Delta\right) = \frac{\exp(\theta - \delta_i + \tau_{ij})}{\sum_{k=0}^{M_i} \exp(\theta - \delta_i + \tau_{ij})}, j = 0,1, \dots, M_i$$

Donde el ítem *i* tiene M_i steps, y *j* es el número de pasos (*steps*) que los alumnos completan correctamente o el número de puntos obtenidos. El parámetro $\delta_i$proporciona la localización del ítem en el espacio latente θ y τ_{ij} es la dificultad del paso *j* en el ítem *i*.
Este modelo es un modelo condicional en cuanto que describe el proceso de respuesta condicionado un nivel de aptitud θ. Por tanto, la definición completa del modelo a estimar en este estudio requiere la especificación de la función de densidad $f_\theta(\theta)$, para la variable latente, θ. Se asume que la aptitud latente de los alumnos se muestrea de una población normal con media μ y varianza σ^2,

$$f_\theta(\theta; \mu; \sigma^2) = \frac{1}{\sqrt{2\pi\sigma^2}} exp\left(-\frac{(\theta - \mu)^2}{2\sigma^2}\right)$$

Para lograr la especificación final del modelo (modelo *logit* multinomial de coeficientes mixtos), el modelo de respuesta y el modelo poblacional han de combinarse para obtener el modelo mixto:

$$f_x(X;\Delta) = \int_\theta f_x(X;\Delta|\theta) f_\theta(\theta;\beta,\sigma^2)\, d\theta$$

Tras la calibración de las respuestas de los alumnos bajo el modelo descrito, se estudiarán los parámetros obtenidos para los modos de aplicación papel y tableta y se estudiarán la relación entre los parámetros mediante pruebas de correlación y mediante inspección gráfica (gráficos de puntos o dispersión). Se prevé obtener un grado de correlación entre los parámetros altos, y que las diferencias entre éstos sean de pequeñas a moderadas. Diferencias de aproximadamente 0.3 o más logits (criterio propuesto por Hyunh y Rawls, 2009) serán consideradas evidencias del impacto del modo de aplicación. El ajuste de los parámetros estimados libremente para los modos de aplicación papel y tableta también será analizado mediante el índice MNSQ. Los análisis serán realizados para cada modo de aplicación y para cada dominio de evaluación, esto es, lectura, matemáticas y ciencias.

El análisis descrito será repetido para las dos muestras, pero anclando los parámetros de los ítems a sus valores internacionales. De esta manera se podrá evaluar el ajuste de los datos de las muestras papel y tableta frente a la predicción del modelo. De este análisis se espera mayor desajuste, puesto que los parámetros internacionales están calibrado con una muestra de alumnos de edad superior a la muestra de alumnos de este estudio.

También se evaluará la dificultad global por cada dominio en los grupos MP y MT. Se calcularan los parámetros de dificultad promedios para lectura, matemáticas y ciencias, y mediante pruebas de significación estadística y de tamaño del efecto se estudiará la presencia del efecto de aplicación y si lo hubiera, se cuantificará su impacto real (Cohen, 1988).

Funcionamiento diferencial del ítem. Un ítem presenta funcionamiento diferencial (FDI) cuando la probabilidad de acertar el ítem es diferente para alumnos que pertenecen a distintos grupos aún teniendo los alumnos el mismo nivel de competencia (Fidalgo, 1996). Es decir, un ítem en particular de la prueba PfS mostrará FDI si los alumnos del grupo MP tienen más probabilidad de acertarlo que los alumnos que contestan al mismo ítem del grupo MT, teniendo el mismo nivel de aptitud. El estudio de los ítems con posible FDI se llevará a cabo sobre los ítems que los análisis previos señalaron como ítems afectados por el modo de aplicación. Sobre los grupos se estudiará el efecto del modo de aplicación, así como su interacción con las variables de género y curso. Estudios previos han mostrado no solo FDI en los ítems de aplicación informatizada, sino efectos de interacción con el curso académico en el que se encontraban los alumnos (Choi y Tinkler, 2002) y efectos diferenciales del modo de aplicación según el género de los alumnos (OECD, 2015). Los resultados se presentarán a nivel de ítem: para cada uno de los ítems analizados, junto con sus correspondientes curvas características del ítem (CCI) para inspección gráfica, y a nivel de dominio: se examinará la puntuación promedio de cada uno de los grupos en Lectura, Matemáticas y Ciencias.

Ajuste de las puntuaciones modo de aplicación. En caso de que los resultados de los análisis descritos apunten hacia la incomparabilidad de las puntuaciones de los modos papel y tableta, será necesario realizar algunos ajustes sobre las puntuaciones para compensar las diferencias.

Para este ajuste el modo de aplicación será introducido en el modelo como regresor directo, y codificado como variable dummy previamente. Así los valores de 0 identificarán a los alumnos que contestaron a la prueba en papel, y el valor 1 identificara a los alumnos que lo hicieron en formato tableta. El modelo será estimado fijando los parámetros de los ítems internacionales. Para cada dominio se estimarán los coeficientes de regresión β y sus correspondientes errores estándar. Los coeficientes β determinarán el valor del ajuste.

Ajuste de las puntuaciones edad de los alumnos. Puesto que alrededor del 20% de los alumnos de esta muestra tienen una edad inferior a 15 años y 3 meses es esperable que puntúen por debajo de los alumnos de mayor edad, debido al efecto madurativo y al efecto del tiempo de instrucción. Los promedios de rendimiento para los centros estarían infraestimados y no serían directamente comparables a los rendimientos de los países participantes en el estudio PISA.
Para el ajuste de edad el rendimiento de los alumnos en la prueba será regresado a la edad en el momento de aplicación de la prueba. Para cada dominio se estimarán los coeficientes de regresión β y sus correspondientes errores estándar. Los coeficientes β determinarán el valor del ajuste para cada mes de desviación.

Para el ajuste de las puntuaciones es fundamental considerar al mismo tiempo la representatividad de la muestra. La muestra de estudio Knotion® está compuesta en su mayoría por centros educativos privados, y supone una muestra claramente no representativa de la muestra de alumnos de México en PISA 2015, especialmente en el índice socioeconómico y cultural (ESCS), índice clave que mejor predice el rendimiento de los alumnos.

En PISA 2015 la muestra nacional mexicana de centros públicos tuvieron un ESCS de -1,37 puntos, es decir, más de una desviación típica por debajo del promedio de la OCDE. Los centros públicos de la misma muestra contaban con un ESCS superior (-0,19) y más cercano al promedio del conjunto de todos los países de la OCDE. La importante relación entre este índice y el rendimiento se entiende mejor si se considera que en promedio un aumento de 1 punto en el índice ESCS, se corresponde con 39 puntos en la escala de ciencias.

Para corregir este efecto se calcularán pesos post-estratificación asignando a cada alumno un peso específico en función de la categoría o rango percentil del ESCS en la que se encuentre.
Tras estos análisis, y realizados los ajustes descritos, se transformará la puntuación de los alumnos de la escala *logit* a la escala PISA, de acuerdo a los coeficientes proporcionados por la OECD (2016). Por último, se calcularan los promedios para cada centro y se evaluarán las diferencias.

Todos los cálculos se realizarán con el paquete *Test Analysis Module* (Robitzsch, 2017) del software R (R Core team, 2017), SPSS v22.0 (IBM, 2013) y Mplus (Muthén, 2007).

Bibliografía

Adams, R. J., Wilson, M., y Wang, W. C. (1997). The Multidimensional Random Coefficients Multinomial Logit Mode. *Applied Psychological Measurement, 21(1)*, 1-23.

American Educational Research Association, American Psychological Association, y National Council on Measurement in Education. (2014). *Standards for educational and psychological testing*. Washington, DC: American Educational Research Association.

American Psychological Association. (1986). Guidelines for Computer-Based Test and Interpretations. Washington, DC: American Psychological Association.

Bennett, R. E. (2003). Online assessment and the comparability of score meaning. Princeton, NJ:

Educational Testing Service, ETS. Informe de investigación. Descargado desde http://www.ets.org/media/Research/pdf/RM-03-05-Bennett.pdf

Choi S.W. y Tinkler, T. (2002). *Evaluating comparability of paper and computer-based assessment in a K-12 setting*. Articulo presentado en el congreso annual de la NAcinoal Council on Measurement in Education, New Orleans, LA.

Cohen, J. (1988). *Statistical power analysis for the behavioral sciences* (2º ed.). New Jersey: Lawrence Erlbaum.

Cronbach, L.J. (1951). Coefficient alpha and the internal structure of test. *Psychometrika*, *16*, 297-334.

Feldt ,L.S. (1969). A test of the hypothesis that Cronbach´s alpha or Kurder-Richardson coefficient twenty is the same for two test. *Psychometrika*, *34*, 363-373.

Fidalgo, A.M. (1996). Funcionamiento diferencial del ítem. En J. Muñiz (Ed.), *Psicometría*. Madrid: Universitas.

Hu, L., y Bentler, P.M. (1999). Cutoff criteria for fit indexes in covariance structure analysis: Conventional criteria versus new alternatives. S*tructural Equation Modeling: A multidisciplinary Journal, 6,* 1-55.

Huynh, H, y Rawls, A. (2009). A comparison between robust z and 0.3-logit difference procedures in assessing stability of linking items for the Rasch model. En Everett V. Smith Jr. y Greg E. Stone (Eds.) *Applications of Rasch Measurement in Criterion-Referenced Testing: Practice Analysis to Score Reporting.* Maple Grove, MN: JAM Press.

IBM Corp. Released 2013. IBM SPSS Statistics for Windows, Version 22.0. Armonk, NY: IBM Corp.

Joreskog, K. y Soborm, D. (1989). *LISREL 7, a guide to the program and applications*. Chicago: SPSS Publications.

Kolen, M.J. (1999). Threats to score comparability with applications to performance assessment and computerized adaptative test. *Educational Assessment*, *6(2)*, 73-96.

Masters, G. N. (1982). A Rasch model for partial credit scoring. *Psychometrika*, Vol. 47, pp. 149-174.

Muthén, L. K.., y Muthén, B. O. (2007). Mplus User's Guide (Sixth Edition). Los Angeles, CA: Muthén & Muthén.

O´Malley, K. J., Kirkpatrick , R., Sherwood, W., Burdick, H.J., Hsieh, M.C., y Sanford, E. (2005). *Comparability of a paper based and computer based reading test in early elementary grades*. Articulo presentado en el congreso anual de la American Educational Research Association, Montreal, Canadá.

Organización para la Cooperación y el Desarrollo Económicos, OECD. (2015). *PISA 2015 Volume I*. OECD Publishing, Paris.

Organización para la Cooperación y el Desarrollo Económicos, OECD. (2017). *Pisa-based Test for Schools Technical Report.* OECD Publishing, Paris.

Pommerich, M. (2004). Developing computerized versions of paper-and-pencil tests: Mode effects for passage-based tests. *The Journal of Technology, Learning, and Assessment, 2(6)*.

R Core Team (2017). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. URL https://www.R-project.org/.

Rasch, G. (1960), Probabilistic Models for Some Intelligence and Attainment Tests, Copenhagen, Denmark: Nielsen and Lydiche (Expanded edition, Chicago, University of Chicago Press, 1980).

Robitzsch, A., Kiefer, T., & Wu, M. (2017). TAM: Test analysis modules. R package version 2.8-21.

Steiger y Lind, (1980). *Statistically-based tests for the number of common factors.* Trabajo presentado en la Conferencia anual de la Sociedad de psicometría.