INFORME: TIMSS 11 DATABASE – MATHEMATICS – EIGHT GRADE

Se va a analizar la base de datos correspondiente a la base de datos *Trends In Mathematics and Science Study* (TIMSS) de 2011, en estudiantes de octavo grado. El TIMSS es una prueba que evalúa cada 4 años a estudiantes de cuarto y octavo grado a nivel internacional, en competencias matemáticas y científicas. Durante este informe, se procederá a realizar una calibración de un modelo TRI a los datos del test, estudiando el ajuste del modelo a los datos, la posible existencia de DIF y la equiparación de dos cuestionarios formados por ítems teóricamente paralelos.

1. Descripción de los datos

Se cuenta con participantes estadounidenses y canadienses, y con dos cuadernillos distintos, compuestos por ítems teóricamente paralelos.

Tabla 1. Número de participantes por país y cuadernillo

Cuadernillo/ País	*Cuadernillo 1*	*Cuadernillo 2*	*Total Países*
Canadá	1132	1142	2274
Estados Unidos	731	743	1474
Total Cuadernillos	1863	1885	3748

Ambos cuadernillos evalúan el constructo de Competencia Matemática, dividida en las siguientes áreas: Números, Datos y Azar, Álgebra y Geometría. A su vez, cada una de estas áreas puede evaluar uno de los siguientes dominios cognitivos: Conocimiento, Razonamiento o Aplicación. Los cuadernillos se componen por ítems de rendimiento óptimo, tanto dicotómicos como politómicos, y de respuesta seleccionada y respuesta construida. El primer cuadernillo se compone de un total de 26 ítems, y el segundo, de 32 (ver Anexo A). Los ítems dicotómicos se han codificado como: 0 – Fallo, 1 – Acierto. Los politómicos se han codificado como 0 – Incorrecto, 1 – Parcialmente Correcto, 2 – Correcto.

A continuación, se realizará un análisis del Cuadernillo 1, estudiando sus propiedades psicométricas y ajustando un modelo de TRI.

2. Análisis psicométrico clásico

Se analiza la escala de acuerdo a la TCT. El coeficiente alfa de Cronbach de fiabilidad toma un valor de $\alpha = 0.84$, con un intervalo de confianza al 95% de $IC_{\alpha_{95}} = (0.83, 0.85)$. Los estadísticos obtenidos para cada ítem se incluyen en la siguiente tabla:

Tabla 2. Análisis psicométrico clásico de los ítems.

Ítem	Media	Desviación Típica	Correlación Ítem-Resto corregida	α si se elimina
m032166_r	.81	.39	.34	.84
m032721_r	.42	.31	.23	.84
m032757_r	1.41	.59	.48	.83
m032760a_r	.89	.7	.6	.83
m032760b_r	.3	.61	.56	.83
m032760c_r	.21	.58	.53	.83
m032761_r	.36	.63	.56	.83
m032692_r	.33	.5	.41	.84
m032626_r	.56	.39	.32	.84
m032595_r	.64	.5	.43	.84
m032673_r	.63	.48	.41	.84
m052216_r	.83	.4	.35	.84
m052231_r	.88	.23	.18	.84
m052061_r	.64	.52	.46	.84
m052228_r	.33	.45	.39	.83
m052214_r	.36	.26	.19	.84
m052173_r	.11	.29	.24	.84
m052302_r	.81	.37	.31	.84
m052002_r	.28	.53	.46	.84
m052362_r	.45	.52	.46	.83
m052408_r	.28	.41	.35	.83
m052084_r	.62	.45	.39	.84
m052206_r	.34	.49	.43	.84
m052429_r	.72	.43	.37	.84
m052503a_r	.22	.33	.27	.84
m052503b_r	.25	.26	.2	.84

3. Estudio de la unidimensionalidad.

Como se ha mencionado, el cuestionario empleado evalúa un constructo: competencia matemática. A pesar de estar dividido en las facetas que ya hemos mencionado (Números, Álgebra, Datos y Azar, y Geometría), este constructo es teóricamente unidimensional.

Para comprobar que el supuesto de unidimensionalidad puede mantenerse, en primer lugar se realiza un Análisis Paralelo. Los resultados pueden verse en la Figura 1 y en la Tabla 3. Como se ve, la recomendación es extraer 2 componentes:

Figura 1. Análisis Paralelo de las respuestas al cuadernillo 1.

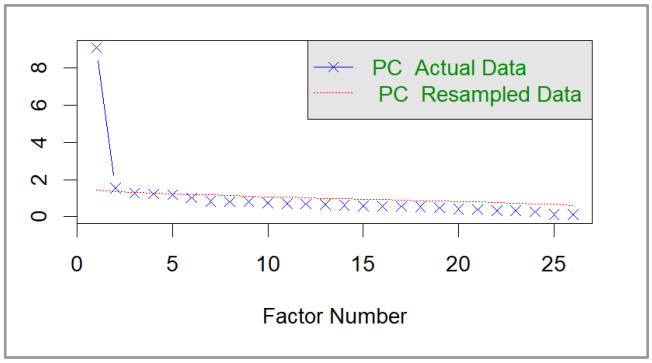


Tabla 3. Autovalores empíricos y simulados de las respuestas al cuadernillo 1.

Autovalor	*Empírico*	*Simulados Media*	*Simulados* P_{95}
1	*9.1*	*1.4*	*1.46*
2	*1.55*	*1.36*	*1.4*
3	*1.28*	*1.31*	*1.35*

**Nota: Se incluyen tan solo los tres primeros autovalores, a efecto ilustrativo.*

Se puede observar que el ratio entre el primer y el segundo autovalor empíricos es de 5.8, superior a lo recomendado para poder mantener el supuesto de unidimensionalidad en TRI. Otra forma de comprobar que se cumple el supuesto de unidimensionalidad es mediante la aplicación de un Análisis Factorial Confirmatorio para un modelo unifactorial. Los índices de ajuste del modelo (CFI_R = .96; TLI_R = .96; $RMSEA_R$ = .037; SRMR = .065) muestran un ajuste adecuado. La mayor parte de los pesos factoriales (excepto 2) son mayores a .3, por lo que la mayor parte de los ítems tendría una correlación ítem-factor elevada. La proporción de varianza explicada por el factor es de .34 (mayor al .2 que se considera estándar para poder mantener el supuesto de unidimensionalidad).

Se efectúa también un análisis de los residuos del modelo factorial, que desvela que se encuentran residuos superiores a .1 en valor absoluto para 35 parejas (de las 351 posibles), vulnerando así uno de los criterios clásicos empleados para poder mantener el supuesto de unidimensionalidad estricta mediante CFA. Existen criterios que apoyan el supuesto de unidimensionalidad, y otros que tenderían a rechazarlo. Sin embargo, se decide mantenerlo dados los siguientes hechos: este criterio es sensible a pequeñas desviaciones del modelo (residuos correlacionados), los criterios recomendados de forma específica para TRI se han cumplido, y los modelos TRI son robustos al incumplimiento de la unidimensionalidad estricta si existe un factor dominante, como es nuestro caso.

4. Estimación de parámetros de un modelo TRI.

A continuación, se va a calibrar un modelo TRI para las respuestas a nuestro cuadernillo. Se plantea un modelo en el cuál se emplea el ML3P para los ítems dicotómicos de opción múltiple, el ML2P para los ítems dicotómicos de respuesta construida, y el MRG para los ítems politómicos. Este modelo ha demostrado mostrar un ajuste significativamente mejor a un modelo simplificado en el que se usa el ML2P para todos los ítems dicotómicos (ver Anexo C). Para la estimación del parámetro *c* de pseudoazar, se emplea

una distribución a priori normal para el logit de *c*, con media -1.099 y desviación típica de 1. La correlación entre los parámetros *c* estimados sin asumir una distribución previa y los estimados por el modelo descrito es de .86. A continuación se incluyen los parámetros de los ítems, empleando métrica logística ($D = 1$):

Tabla 4. Parámetros de los ítems.

	Modelo	a	b	c	b1	b2
m032166_r	*ML3P*	1.18	-1.55	.01	-	-
m032721_r	*ML3P*	.91	1.3	.21	-	-
m032757_r	*MRG*	1.71	-	-	-.89	-.65
m032760a_r	*MRG*	2.52	-	-	.1	.25
m032760b_r	*ML2P*	2.29	.66	-	-	-
m032760c_r	*ML2P*	2.41	.98	-	-	-
m032761_r	*MRG*	2.2	-	-	.83	1.64
m032692_r	*MRG*	1.3	-	-	1.45	1.8
m032626_r	*ML3P*	1.54	.64	.34	-	-
m032595_r	*ML3P*	1.54	-.31	.13	-	-
m032673_r	*ML3P*	1.17	-.56	0	-	-
m052216_r	*ML3P*	1.47	-1.13	.27	-	-
m052231_r	*ML2P*	.66	-3.31	-	-	-
m052061_r	*ML2P*	1.42	-.54	-	-	-
m052228_r	*ML3P*	1.76	1.04	.13	-	-
m052214_r	*ML3P*	.86	1.93	.21	-	-
m052173_r	*ML3P*	3.2	1.9	.07	-	-
m052302_r	*ML3P*	1.06	-1.62	.01	-	-
m052002_r	*MRG*	1.67	-	-	1.29	1.91
m052362_r	*ML2P*	1.27	.22	-	-	-
m052408_r	*ML2P*	.94	1.19	-	-	-
m052084_r	*ML3P*	1.02	-.57	0	-	-
m052206_r	*ML2P*	1.18	.73	-	-	-
m052429_r	*ML3P*	1.11	-1.05	0	-	-
m052503a_r	*ML2P*	.72	1.93	-	-	-
m052503b_r	ML2P	.5	2.4	-	-	-

Vamos a fijarnos en los parámetros del cuarto ítem: *m032760a_r*. Este es un ítem de codificación politómica en el que la mayor puntuación es un 2, y la menor un . Este ítem tiene un valor muy elevado del parámetro *a* de discriminación (de hecho, es el segundo mayor de todo el cuadernillo). Un valor alto de *a* nos indica que la discriminación será muy elevada, por lo que podemos esperar pendientes para las curvas logísticas muy inclinadas, y curvas características de respuesta estrechas. Por otro lado, los parámetros *b* del ítem están ordenados (lo que nos indica que, efectivamente, obtener una puntuación mayor en el ítem se relaciona con un mayor nivel de rasgo). Sin embargo, la distancia entre b_1 y b_2 es muy escasa, lo que nos indica que ambas curvas logísticas estarán muy próximas. Teniendo en cuenta que, en el MRG cada curva indica la probabilidad de obtener la puntuación *k* o superior (excepto la última), esto nos indica que la probabilidad de obtener la puntuación intermedia (1, en este caso) será muy baja, ya que ambas curvas

prácticamente se solapan. Esto se puede ver más claramente con las Curvas Características Operantes y las Curvas Características de Respuesta:

Figura 2. Curvas Características Operantes del ítem 4.

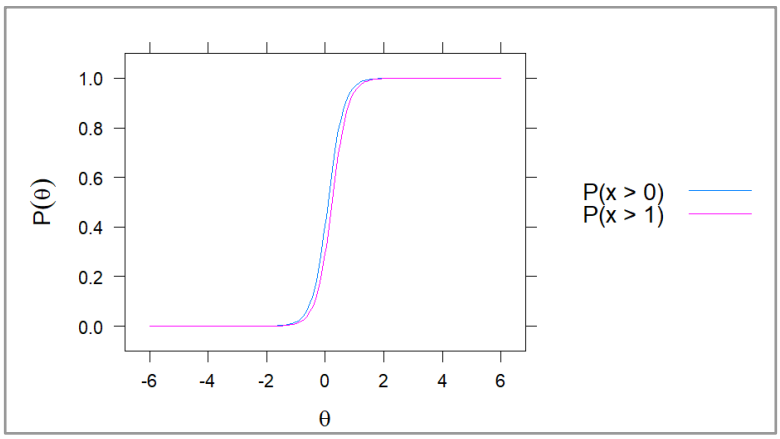
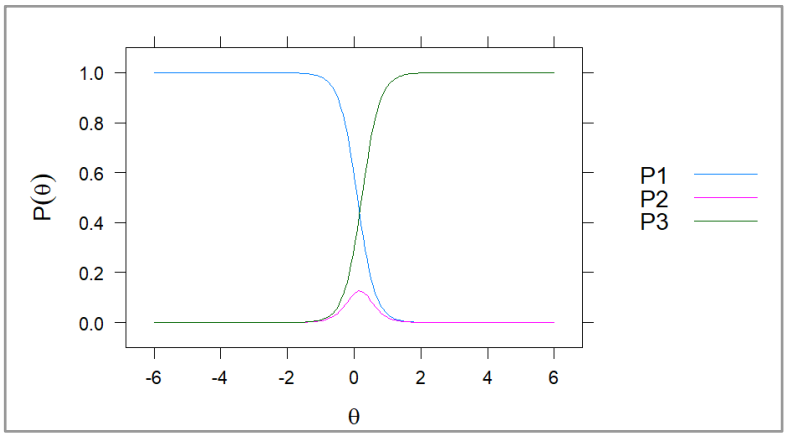


Figura 3. Curvas Características de Respuesta del ítem 4.



**Nota: En este caso, P1 hace referencia a obtener la menor puntuación (0), P2 a obtener 1, y P3 a obtener 2.*

5. Evaluación de la independencia local.

Para evaluar el ajuste del modelo, atendemos a tres tipos de índices de ajuste: aquellos que informan de la independencia local, aquellos que muestran el ajuste de las CCI, y aquellos que reportan el ajuste global.

En primer lugar, atendiendo al supuesto de independencia local, se emplean los estadísticos X^2 de Chen y Thissen (estandarizado), y V de Cramer, derivado del anterior. El primero, trabajando por pares de ítems, comprueba si la diferencia entre frecuencias observadas y esperadas para cada patrón de respuestas en la pareja es significativamente distinta de . El segundo es una transformación de X^2, y funciona como medida de tamaño del efecto. A priori, observaremos que se da dependencia local entre una pareja de ítems si el estadístico X^2 estandarizado toma un valor superior a 10, y esta será problemática si V de Cramer toma un valor superior a .2.

Tabla 5. Parejas de ítems con dependencia local.

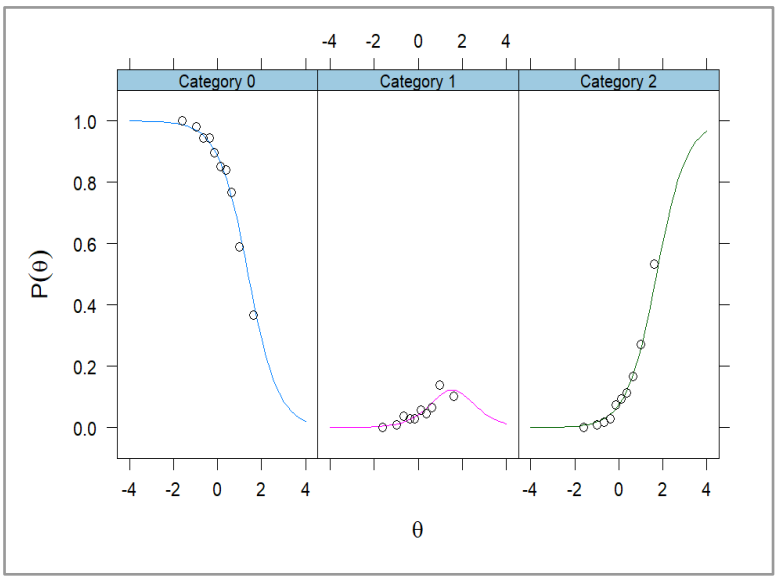
Ítem 1	Ítem 2	$X2_{estándar}$	V Cramer	P_{BH}
m052503b_r	m052503a_r	46.57	.2	0
m032760c_r	m032760b_r	38.56	.17	0
m032760b_r	m032760a_r	31.62	.19	0
m032760c_r	m032760a_r	16.73	.14	0
m032760a_r	m032757_r	15.51	.11	0
m052231_r	m052216_r	1.1	.09	.004

En total encontramos 6 parejas de ítems que mostrarían dependencia local según X^2, de las cuáles, atendiendo a V de Cramer, se podría considerar problemática 1, utilizando el punto de corte habitual de .2. De esta forma, la relación entre el acierto en el ítem *m052503b_r* y el acierto en el ítem *m052503a_r* será superior a lo esperado.

6. Ajuste de las Curvas Características del Ítem empíricas a las teóricas.

A continuación, empleamos el índice chi-cuadrado de Orlando y Thissen para evaluar el ajuste a las CCI observadas. Empleando el método de corrección por comparaciones múltiples de Benjamini-Hochberg, obtenemos que ningún ítem muestra un desajuste significativo. Se incluye la Figura 4, que compara las CCR esperadas y observadas para el ítem *m032692_r*, que es aquel que muestra un desajuste mayor:

Figura 4. CCR teóricas para el ítem *m032692_r*, superponiendo los valores empíricos.



7. Ajuste global

Respecto al ajuste global, se calcula el estadístico M2, que contrasta la hipótesis nula de que las proporciones esperadas son iguales a las observadas, así como los índices RMSEA, SRMSR, TLI y CFI.

Tabla 6. Estadísticos de ajuste global

M2	*gl*	*p*	RMSEA	SRMSR	TLI	CFI
1114.038	282	< .001	.043	.038	.948	.954

El índice M2 indicaría que el modelo desajusta significativamente, lo cual no es raro, teniendo en cuenta el elevado tamaño muestral. Sin embargo, el índice RMSEA nos mostraría que el ajuste es adecuado, al ser inferior al punto de corte habitual de .05.

8. Distribución del nivel de rasgo y relación con la puntuación directa

La distribución del nivel de rasgo estimado (mediante estimador MAP) en nuestra muestra es la siguiente:

Figura 5. Distribución del nivel de rasgo en la muestra.

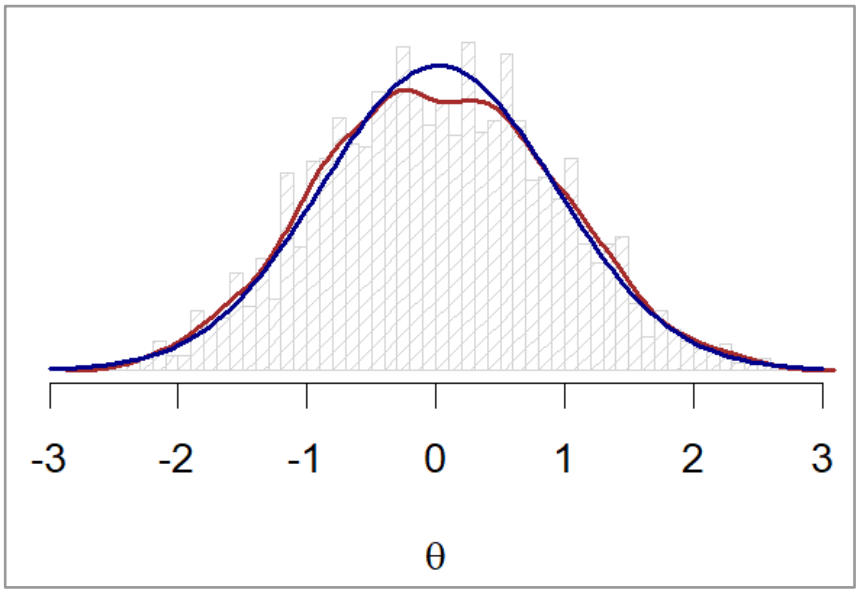
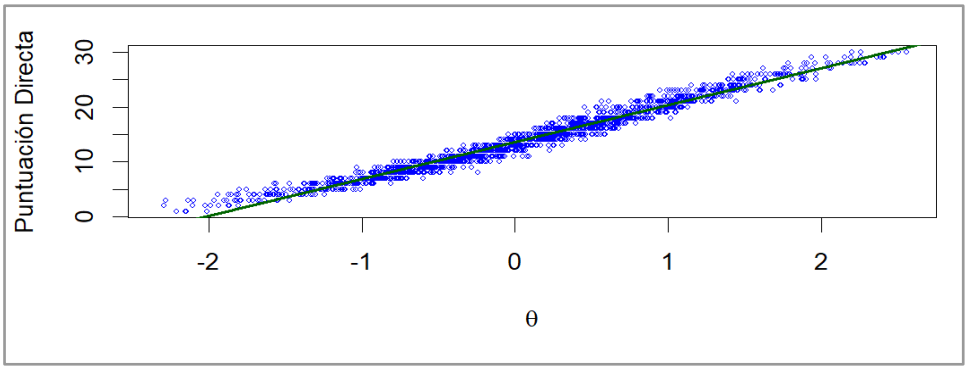


Tabla 7. Estadísticos descriptivos del nivel de rasgo en la muestra.

Media	*Desviación Típica*	*Mínimo*	*Máximo*	*Asimetría*	*Curtosis*
.02	.9	-2,33	2.56	.06	-.36

A pesar de que la distribución no es estrictamente normal (p = .033 para el test Kolmogorvo-Smirnov), se aproxima a ella, con estadísticos de asimetría y curtosis que toman un valor absoluto muy inferior a 2, siendo este el criterio habitual para evaluar si la violación de la normalidad es problemática. La correlación entre el nivel de rasgo estimado y la puntuación directa obtenida en el test es de .98. La Figura 7 muestra un gráfico de dispersión para ambas variables, junto con la línea de tendencia ajustada a una regresión lineal. Como se observa, la predicción es muy buena, si bien nos puede llamar la atención que empeora ligeramente en los valores más bajos de theta, al predecir puntuaciones de . Es importante recordar en este punto que se han eliminado los casos en los que un sujeto puntuaba 0 en todos los ítems, por no considerarlos informativos.

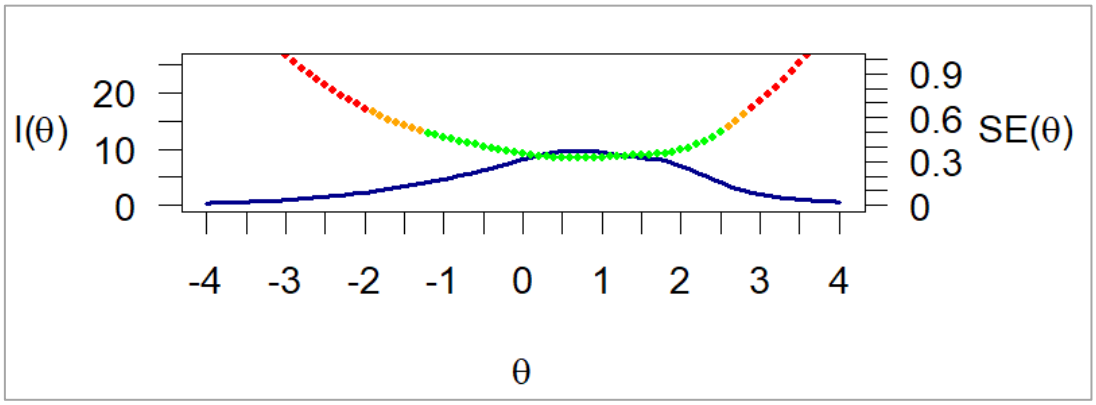
Figura 6. Relación entre el nivel de theta estimado y la puntuación directa.



9. Información del test

La función de información para los distintos niveles de rasgo es la siguiente:

Figura 7. Función de información del test.



Esta función alcanza su máximo en $\theta = 0.7$, por lo que el test será máximamente informativo para personas con un nivel de rasgo alrededor de este valor. La fiabilidad empírica obtenida es de .87, lo que nos indica que un 87% de la varianza total representaría varianza "verdadera", en promedio. Este es un porcentaje elevado, que nos asegura ciertas garantías de fiabilidad para la aplicación del test. La fiabilidad será mayor para las personas cuyo nivel de rasgo se encuentre entre (-1.2, 2.5), superando un valor de .8.

10. Análisis del Funcionamiento Diferencial de los Ítems (DIF) y del Test (DTF)

En este apartado se va a estudiar si existe una diferencia significativa entre las puntuaciones obtenidas por las muestras canadiense y estadounidense, y si existe DIF/DTF. Con DIF nos referimos a una diferencia en la probabilidad de acertar cada ítem para los miembros de ambos grupos, condicionando al nivel de rasgo estimado. DTF hace referencia al mismo fenómeno, pero para la probabilidad de obtener una determinada puntuación en el test. El DIF/DTF viene causado por la presencia de sesgos que afectan negativamente a alguno de los grupos, por ejemplo, un enunciado puede contener una palabra de uso menos común en uno de los dos países, y por tanto el ítem será más complicado para un grupo. Por tanto, al estudiar el DIF/DTF estamos comprobando que

no existan sesgos, y que la interpretación de las puntuaciones en términos de diferencias entre países sea válida.

Para detectar el DIF se emplea la técnica de Razón de Verosimilitudes, por ser la que mejores resultados ofrece en estudios de simulación. Para tener en cuenta el tamaño del efecto, se han incluido además las medidas SIDS, UIDS y ESSD. El grupo estadounidense se ha tomado como grupo de referencia (R), y el canadiense como focal (F). En el código se ha incluido también la prueba SIBTEST para detectar DIF, que no se incluye aquí por ofrecer unos resultados similares.

Tanto para este apartado como para el siguiente, se han dicotomizado las respuestas (los valores de 2 en los ítems politómicos han pasado a valer 1). Aunque esto elimina variabilidad, y por tanto nos hace perder información, facilita mucho operar con los datos, y el número de ítems politómicos es tan pequeño que no debería afectar a los resultados finales.

Nivel descriptivo

Existe una diferencia significativa (p = .0015) en la puntuación directa del test entre ambos grupos, siendo la media de Canadá 12.445, y la de Estados Unidos 11.665. El tamaño del efecto se encuentra entre .058 y .24, con una confianza del 95%. Parece, por tanto, que hay una pequeña tendencia a obtener mayor número de aciertos en Canadá.

Razón de Verosimilitudes

Se llevó a cabo en dos pasos. En el primer paso, se realizó por cada ítem un contraste por razón de verosimilitudes para evaluar el ajuste de dos modelos, uno en el que se fija que los parámetros del ítem tomen el mismo valor en ambos grupos, y otro en el que se estiman para cada grupo por separado. Esto se hace asumiendo que en el resto de los ítems no hay DIF. Como este supuesto puede no ser verdadero, se realiza un segundo paso.

En el segundo paso, se eligieron 5 ítems de anclaje que se fijaron como invariantes, en concreto: *m032760a_r, m032760c_r, m032760b_r, m032761_r* y *m032757_r*. Estos 5 ítems se eligieron a partir de los resultados del paso 1, entre aquellos más discriminativos y con menos probabilidades de tener DIF. A partir de este test de anclaje, se evalúa de nuevo por razón de verosimilitudes como se incrementa el desajuste al fijar los parámetros de un ítem como invariantes en ambos grupos, frente al dejarlos libres.

Los resultados se muestran en la Tabla 7. Se incluyen también las medidas del tamaño del efecto SIDS, UIDS y ESSD. Vemos que existen 8 ítems con DIF, y atendiendo al ESSD, en la mayoría de ellos el tamaño del efecto sería al menos medio. También vemos, sin embargo, que la mitad de los ítems favorecen a un grupo, y la otra mitad al otro, por lo que es posible que el efecto a nivel de test no sea relevante.

Tabla 8. DIF por ítem por razón de verosimilitudes

Ítem	*X2*	*p*	Tipo	Grupo favorecido	SIDS	UIDS	ESSD
m032166_r	*24.147*	.000	Unidireccional	R	-.084	.084	-.61
m032721_r	*11.837*	.012	Unidireccional	F	.081	.081	.77
m032757_r	*4.056*	.246	-	-	-	-	-

m032760a_r	*.570*	.815	-	-	-	-	-
m032760b_r	*.743*	.795	-	-	-	-	-
m032760c_r	*3.387*	.281	-	-	-	-	-
m032761_r	*.705*	.795	-	-	-	-	-
m032692_r	*.352*	.838	-	-	-	-	-
m032626_r	*51.092*	.000	Unidireccional	R	-.164	.164	-1.02
m032595_r	*8.681*	.042	No Unidireccional	R	-.047	.054	-.2
m032673_r	*4.743*	.202	-	-	-	-	-
m052216_r	*6.562*	.109	-	-	-	-	-
m052231_r	*3.749*	.249	-	-	-	-	-
m052061_r	*1.183*	.685	-	-	-	-	-
m052228_r	*4.041*	.246	-	-	-	-	-
m052214_r	*6.156*	.120	-	-	-	-	-
m052173_r	*1.501*	.646	-	-	-	-	-
m052302_r	*76.620*	.000	Unidireccional	R	-.152	.152	-1.14
m052002_r	*1.507*	,656	-	-	-	-	-
m052362_r	*1.328*	.669	-	-	-	-	-
m052408_r	*5.960*	.123	-	-	-	-	-
m052084_r	*2.829*	.000	No Unidireccional	F	.093	.094	.48
m052206_r	*11.854*	.012	Unidireccional	F	.076	.076	.37
m052429_r	*3.847*	.249	-	-	-	-	-
m052503a_r	*.422*	.828	-	-	-	-	-
m052503b_r	9.844	.027	No Unidireccional	F	.055	.056	.66

Repercusión del DIF a nivel de test

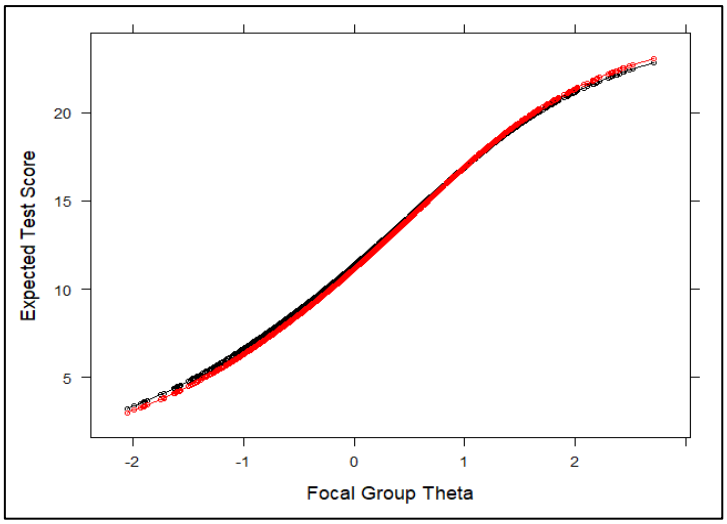
Se calcularon las medidas de tamaño del efecto STDS, UETDS y ETSSD. Los resultados pueden verse en la Tabla 8.

Tabla 9. Tamaño del efecto DTT

STDS	UETSDS	ETSSD
-.1413	.1772	-.032

El STDS nos dice la diferencia que podemos esperar para personas con el mismo nivel de habilidad en el test. Como vemos, es una puntuación muy pequeña, teniendo en cuenta que la puntuación máxima es 26. El ETSSD, que funciona como medida de tamaño del efecto, nos indica que la diferencia es despreciable. Si bien el valor esperado es menor en el grupo de referencia (Estados Unidos), fijándonos en el UETSDS vemos que hay una pequeña cancelación a través de los sujetos. Esto se aprecia bien en la siguiente Figura, donde la línea roja representa al grupo canadiense:

Figura 8. Puntuación esperada en el test por nivel de rasgo, dividido por grupos.



11. Detección de patrones aberrantes

Un patrón aberrante es aquel en el que las respuestas de un sujeto no se ajustan al modelo teórico. Esto es, la persona se comporta de forma incongruente con su nivel de rasgo estimado. Es, como un ejemplo extremo, el caso de la persona que acierta todos los ítems menos el más fácil. Un alto porcentaje de patrones aberrantes en la muestra será problemático, en tanto que las estimaciones del nivel de rasgo estarán sesgadas.

En primer lugar, se dicotomizan los datos, ya que las funciones implementadas en la librería *PerFit* exigen que los ítems tengan el mismo número de categorías. Se calculan los estadísticos Lzstar, Ht y U3, para adoptar distintos enfoques y poder comprobar si convergen. De cara a la selección de puntos de corte para cada estadístico, se decidió seguir el punto de corte teórico de -1.64 para Lzstar, y utilizar simulación por *bootstrap* para los otros dos (aunque se emplea la simulación para Lzstar también, obteniendo resultados muy parecidos a los de la simulación). En la Tabla 9 puede verse cada estadístico con su punto de corte y la tasa de patrones aberrantes obtenida con éste.

Tabla 9. Puntos de corte y porcentaje de patrones aberrantes

	Lzstar	**Ht**	**U3**
Punto de corte	< -1.64	< .215	> .343
% de patrones aberrantes	5.23%	6.36%	5.98%

Aunque los resultados que ofrece cada método son ligeramente distintos, vemos que tienden a converger en torno a un 6% de patrones aberrantes, valor que no resulta preocupante. Para ver si los casos detectados como aberrantes por cada estadístico se solapan, se calcula la correlación entre los tres estadísticos. Se pueden ver los resultados en la Tabla 10:

Tabla 10. Correlaciones entre estadísticos de detección de patrones aberrantes.

	Lzstar	***Ht***	***U3***
Lzstar	1	-	-
Ht	.945	1	-
U3	-.954	-.966	1

12. Equiparación

La equiparación nos permite transformar las puntuaciones de un test (habitualmente llamado X) a otro (habitualmente, Y). Para ello, hay dos métodos básicos. El primero es obtener unas constantes A y B, tales que el nivel de rasgo estimado por un test X puede expresarse en la métrica del test Y mediante la transformación $\theta_X^{(Y)} = \mathrm{A} \cdot \theta_X + \mathrm{B}$. La segunda forma, es obteniendo la puntuación que habría obtenido una persona en el test Y, a partir de su puntuación en X. En este apartado vamos a equiparar las puntuaciones del cuadernillo 2 (formado por los bloques M02 y M03) a la métrica del cuadernillo 1. Es importante observar que ambos cuadernillos comparten el bloque M02, lo que nos permite usar estos ítems como un test de anclaje proceder con la equiparación.

Empezando por el primero de los métodos, se calibra el segundo cuadernillo utilizando un modelo análogo al empleado para el primer cuadernillo (apartado 4), si bien los ítems politómicos se han dicotomizado. Se emplea el método de Haebara para realizar el proceso de equiparación, que busca calcular unas constantes A y B que minimicen las diferencias en las Curvas Características de los Ítems del test de anclaje en el primer y segundo cuadernillo.

Se obtienen unas constantes A = .999 y B = -.12. Además de permitir la equiparación, estas constantes resumen las diferencias en variabilidad y nivel de habilidad entre ambos grupos. Como vemos, estas diferencias son escasas. La Tabla 11 refleja los parámetros de los 26 primeros ítems del cuadernillo 2 tras el proceso de equiparación métrica. A partir de estas constantes, podemos trasladar tanto las estimaciones del nivel de asgo como los parámetros de los ítems de la métrica del grupo X a la métrica del grupo Y, permitiendo así realizar comparaciones entre personas de ambos grupos.

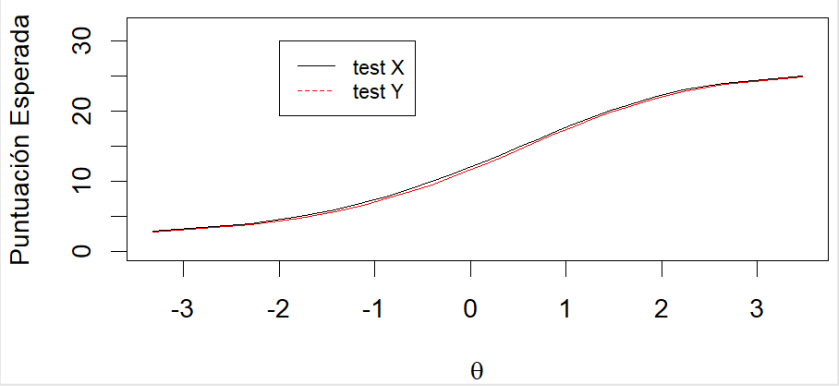
El segundo método se lleva a cabo mediante la equiparación de puntuaciones verdaderas. La puntuación verdadera, en este contexto, es la puntuación directa esperada en cada test según el nivel de rasgo. Aprovechando las constantes A y B que hemos calculado previamente, se puede modelar la puntuación esperada en el test como una función del nivel de rasgo. Esto se observa en las CCT de la Figura 9.

Tabla 12. Equivalencia en puntuaciones entre los cuadernillos y el nivel de rasgo estimado.

X	$X^{(Y)}$
0	0
1	1.01
2	1.98
3	2.82
4	3.81
5	4.75
6	5.7
7	6.67
8	7.62
9	8.6
10	9.58
11	10.57
12	11.58
13	12.59
14	13.61
15	14.63
16	15.66
17	16.69
18	17.71
19	18.73
20	19.73
21	20.74
22	21.74
23	22.75
24	23.79
25	24.84
26	26

**Nota: X denota la puntuación en el primer cuadernillo y $X^{(Y)}$ la puntuación en el segundo cuadernillo tras la equiparación.*

Figura 9. Curvas Características del Test para ambos cuadernillos.



13. Rendimiento de un TAI simulado.

Un TAI, o Test Adaptativo Informatizado, es una herramienta que permite administrar a cada sujeto el ítem que mejor le corresponda, dado su nivel de rasgo estimado. En esencia, los pasos que sigue el algoritmo de un TAI son los siguientes:

- Asignar un valor inicial al nivel de habilidad del sujeto, generalmente un número aleatorio cercano a 0.
- Aplicar el ítem más adecuado para ese nivel de rasgo, que puede seleccionarse en función de distintos criterios.
- En función de la respuesta, estimar nuevamente el nivel de habilidad.
- Continuar hasta alcanzar un criterio de parada. Puede ser alcanzar un cierto número de ítems, un determinado error típico, etc.

Los TAIs tienen varias ventajas frente a los test tradicionales, siendo la principal ofrecer una estimación del nivel de rasgo igualmente o más fiable, con menor número de ítems.

Al tener ya las respuestas, no podemos aplicar un TAI a los sujetos. Sin embargo, sí que podemos simular como funcionaría un TAI presentado a sujetos con un nivel de rasgo similar al de la muestra. De esta forma, en base a los niveles de rasgo estimados a partir de las respuestas reales, se simulan respuestas a los ítems que aplica el TAI.

Nuestro TAI comienza asignando un nivel de rasgo aleatorio entre (-0.5, 0.5), y utiliza el criterio *b-óptima* para elegir el ítem a aplicar. Este criterio busca elegir el ítem que tenga una *b* más cercana al nivel de rasgo del sujeto (o un valor un poco más elevado, en el ML3P), ya que estos ítems serán los más informativos para este sujeto. El método de estimación seleccionado es Máxima Verosimilitud, tanto durante el test como al final del mismo, y se impone un criterio de parada de haber respondido a 10 ítems. Para formar el banco de ítems, se toman los de ambos cuadernillos empleados hasta ahora.

Los índices de rendimiento del TAI pueden observarse en el archivo catR.main.txt. Nos vamos a fijar en cuatro aspectos principales: correlación entre theta estimada a partir de los datos completos, y la theta estimada a partir del TAI; *item overlap rate*, que nos indica qué proporción esperada de ítems del TAI de dos sujetos aleatorios sean similares; el RMSE (*Root Mean Square Error*), que nos informa de la precisión; y el sesgo, que nos informa de si el nivel de rasgo tiende a sobreestimarse o subestimarse.

En el caso del primer indicador, encontramos una correlación de 0.85 para el nivel de rasgo estimado con todos los ítems, y para el estimado con 10 ítems mediante TAI. Hay que recordar que originalmente se disponía de 26 ítems para algunos sujetos, y de 32 para otros, por lo que la reducción en el tamaño del test ha sido importante. Aún así, la correlación obtenida parece suficientemente alta.

El *ítem overlap rate* obtenido es de 0.28, un valor reducido para lo habitual en estos casos. Hemos buscado reducir este solapamiento aplicando el método Randomesque, como puede verse en el código. Este método selecciona, en vez del mejor ítem para cada punto de aplicación, los 5 mejores, y aplica uno al azar. Reducir el solapamiento es adecuado para evitar posibles problemas como la copia, la filtración de ítems, etc. Sin embargo, puede reducir ligeramente la precisión de las estimaciones.

El RMSE es elevado, tomando un valor de 0.64, siendo lo habitual recomendarse un RMSE inferior a 0.5. Esto implica que el TAI será poco fiable, por lo que puede ser recomendable ampliarlo, y quizás permitir un solapamiento mayor.

Por último, el sesgo medio es -0.014, por lo que no parece que, en promedio, el nivel de rasgo se esté ni sobreestimando ni subestimando de forma relevante.

Anexo A.

Estructura de los ítems de los cuadernillos

Tabla A1. Ítems del primer cuadernillo.

Ítem	Formato de respuesta	Número de opciones	Codificación	Contenido	Dominio
m032166	Seleccionada	4	Dicotómica	Números	Conocimiento
m032721	Seleccionada	4	Dicotómica	Datos y Azar	Razonamiento
m032757	Construida	-	Politómica	Álgebra	Razonamiento
m032760a	Construida	-	Politómica	Álgebra	Razonamiento
m032760b	Construida	-	Dicotómica	Álgebra	Razonamiento
m032760c	Construida	-	Dicotómica	Álgebra	Razonamiento
m032761	Construida	-	Politómica	Álgebra	Razonamiento
m032692	Construida	-	Politómica	Geometría	Razonamiento
m032626	Seleccionada	4	Dicotómica	Números	Conocimiento
m032595	Seleccionada	4	Dicotómica	Números	Aplicación
m032673	Seleccionada	4	Dicotómica	Álgebra	Conocimiento
m052216	Seleccionada	4	Dicotómica	Números	Conocimiento
m052231	Construida	-	Politómica	Números	Conocimiento
m052061	Construida	-	Politómica	Números	Aplicación
m052228	Seleccionada	4	Dicotómica	Números	Aplicación
m052214	Seleccionada	4	Dicotómica	Números	Conocimiento
m052173	Seleccionada	4	Dicotómica	Álgebra	Aplicación
m052302	Seleccionada	4	Dicotómica	Álgebra	Conocimiento
m052002	Construida	-	Politómica	Álgebra	Aplicación
m052362	Construida	-	Politómica	Geometría	Razonamiento
m052408	Construida	-	Politómica	Geometría	Razonamiento
m052084	Seleccionada	4	Dicotómica	Geometría	Aplicación
m052206	Construida	-	Politómica	Geometría	Razonamiento
m052429	Seleccionada	4	Dicotómica	Datos y Azar	Razonamiento
m052503a	Construida	-	Politómica	Datos y Azar	Razonamiento
m052503b	Construida	-	Politómica	Datos y Azar	Razonamiento

Anexo B

Tabla B. Pesos del modelo unifactorial para las respuestas al primer cuadernillo.

Ítem	Peso
m032166_r	0,54
m032721_r	0,32
m032757_r	0,74
m032760a_r	0,87
m032760b_r	0,85
m032760c_r	0,84
m032761_r	0,76
m032692_r	0,61
m032626_r	0,43
m032595_r	0,62
m032673_r	0,58
m052216_r	0,57
m052231_r	0,31
m052061_r	0,64
m052228_r	0,52
m052214_r	0,26
m052173_r	0,43
m052302_r	0,50
m052002_r	0,68
m052362_r	0,60
m052408_r	0,49
m052084_r	0,56
m052206_r	0,58
m052429_r	0,54
m052503a_r	0,38
m052503b_r	0,29

Anexo C

Comparación de modelos TRI por razón de verosimilitudes

Tabla 1

Razón de verosimilitudes entre el ML3P y el ML2P para ítems de opción múltiple

	AIC	SABIC	HQ	BIC	logLik	X2	df	p
ML3P	51226.96	5136.83	51343.04	51541.92	-25556.48	-	-	-
ML2P	51166.53	51328.58	51307.05	51547.8	-25514.26	84.43	12	0

Como el desajuste es significativo al pasar del modelo general al simplificado, mantendremos el M3PL en ítems de opción múltiple.