Ajuste de modelos Rasch multidimensionales, de *testlets* y de diagnóstico cognitivo a las pruebas ECE 2015

Andrés Burga León
Ministerio de Educación del Perú

Recibido: 21 de noviembre del 2017 / Aceptado: 26 de enero del 2018
doi: https://doi.org/10.26439/persona2018.n021.1991

El estudio tuvo como objetivos aplicar y comparar los resultados de los modelos: a) multidimensional logístico multinomial de coeficientes aleatorios, b) Rasch de testlets *y c) diagnóstico cognitivo, los datos generados a partir de la aplicación de las pruebas de comprensión lectora y matemática en la Evaluación Censal de Estudiantes 2015. El estudio se realizó con 5 000 000 evaluaciones censales aplicadas a los estudiantes de segundo grado de primaria que asisten a instituciones educativas del programa de educación básica regular (Minedu, 2009). Los resultados indicaron que: 1. el modelo multidimensional, tanto en el caso de la prueba de lectura como la de matemática, las dimensiones modeladas están altamente correlacionadas, 2. es pertinente aplicar un modelo unidimensional a los datos derivados de aplicar las pruebas ECE, y 3. existe cierta consistencia entre los modelos Rasch y el Modelo de diagnóstico cognitivo.*

modelos Rasch multidimensionales / *testlets* / diagnóstico cognitivo / pruebas ECE

Adjustment of the Rasch Multidimensional Model, the Rasch Testlet Model and the Cognitive Diagnosis Model to the 2015 Student Census Evaluation Tests

This study aimed at applying the results of three models: a) the Multidimensional Random Coefficients Multinomial Logit Model, b) the Rasch Testlet Model, and c) the Cognitive Diagnosis Model, and comparing those results to the data generated from the math and reading comprehension tests conducted in the 2015 Student Census Evaluation (ECE). The sample consisted of 5,000,000 evaluations of students attending second grade of primary school in educational institutions belonging to the regular basic education program (Minedu, 2009). The results showed that: 1. there is a high correlation between modeled dimensions in the multidimensional model both in the math and reading comprehension tests, 2. it is convenient to apply a one-dimensional model to the data derived from the application of the ECE tests, and 3. there is some consistency between the Rasch Testlet Model and the Cognitive Diagnosis Model.

Rasch multidimensional models / *testlets* / cognitive diagnostic / ECE tests

Correo electrónico: aburgal@minedu.gob.pe

Desde el año 2007 se aplica en nuestro país la Evaluación Censal de Estudiantes (ECE), orientada a evaluar, en segundo grado de primaria, la lectoescritura básica y el dominio de algunos conceptos matemáticos fundamentales, como la estructura aditiva y la comprensión del sistema decimal de numeración (Minedu, 2009). Para ello, se aplican pruebas estandarizadas a casi medio millón de estudiantes (Minedu, 2014). El análisis psicométrico de esas pruebas se basa en el modelo Rasch unidimensional para ítems dicotómicos. Este modelo se centra en el análisis de la interacción entre una persona y un ítem. Asimismo, establece la probabilidad de respuesta de una persona ante un ítem en términos de la diferencia entre la medida de rasgo latente de la persona y la dificultad del ítem (Bond y Fox, 2007).

A pesar de las ventajas que suponen los modelos Rasch tradicionales, en los últimos veinte años han aparecido un gran número de modelos psicométricos que superan algunos de los problemas de los modelos unidimensionales, la presencia de dependencia local entre ítems que comparten un mismo estímulo o la multidimensionalidad (Reckase, 2009. Los modelos Rasch o los de teoría de respuesta al ítem (TRI) tampoco están libres de críticas. Por ejemplo, De la Torre (2009) señala que dichos modelos son útiles para establecer un orden relativo de estudiantes a lo largo de un continuo dado por un rasgo latente, pero la información que contienen estos puntajes no permite la evaluación de las fortalezas y debilidades específicas de los estudiantes y no pueden ser usados como un mecanismo de retroalimentación que permita a los profesores identificar métodos efectivos en clase y prácticas que pueden ayudar a los estudiantes a aprender mejor.

Considerando la gran diversidad de modelos psicométricos disponibles y la estructura sobre la que se elaboran las pruebas ECE (Minedu, 2009), en el presente estudio se plantea como objetivo general aplicar y comparar los resultados de distintos modelos psicométricos a los datos generados a partir de la aplicación de las pruebas de comprensión lectora y matemáticas en la Evaluación Censal de Estudiantes 2015. Dicho objetivo general puede ser desagregado en los siguientes objetivos específicos:

- Aplicar el modelo multidimensional logístico multinomial de coeficientes aleatorios a los datos generados a partir de las pruebas de comprensión lectora y matemáticas.
- Aplicar el modelo Rasch de testlets a los datos generados a partir de la prueba de comprensión lectora.
- Aplicar el modelo de diagnóstico cognitivo a los datos generados a partir de las pruebas de comprensión lectora y matemáticas.
- Comparar los resultados de los distintos modelos aplicados.

Aplicar modelos Rasch de análisis multidimensionales, de testlets y de diagnóstico cognitivo permitirán una mejor desagregación y comprensión de las distintas habilidades y capacidades

vinculadas al desempeño en las pruebas de comprensión lectora y matemáticas. Se pueden generar, a partir de los resultados de estos modelos psicométricos, nuevas propuestas para presentar los resultados de las ECE a los distintos actores vinculados al quehacer educativo, de tal manera que puedan canalizar mejor sus esfuerzos por mejorar los aprendizajes básicos de los estudiantes.

Modelo multidimensional Logit multinomial de coeficientes aleatorios

Un ítem puede requerir distintos tipos de conocimientos y un número distinto de habilidades para su correcta resolución. Es decir, en muchas situaciones puede ser más realista hipotetizar que las respuestas de una persona a un ítem se deben a su localización en diferentes variables latentes, con lo cual se tiene un espacio latente multidimensional para describir la interacción entre personas e ítems (De Ayala, 2009). De esta manera, surgen los modelos multidimensionales que postulan, tal y como lo señala Reckease (2009), que la relación entre la localización en el espacio multidimensional y la probabilidad de responder correctamente a un ítem puede ser representada mediante una función matemática continua, en la que la probabilidad de una respuesta correcta a un ítem aumenta conforme aumenta la localización de una persona en cualquiera de las coordenadas multidimensionales.

Regresando a los modelos unidimensionales, desde que George Rasch introduce su modelo psicométrico en 1960 ha habido una gran proliferación de trabajos que extienden esa propuesta original (Adams, Wilson y Wang, 1997; Adams y Wu, 2007). En ese contexto, marcado por la aparición de distintos modelos Rasch, Adams y Wilson (1996) presentan el modelo Logit multinomial de coeficientes aleatorios (RCML, por sus siglas en inglés) que pretende integrar, bajo un único marco de referencia, varios modelos Rasch preexistentes y proveer un mecanismo flexible para general y ajustar nuevos modelos. El modelo RCML fue extendido luego por Wang, Wilson y Adams (1997) al contexto multidimensional, dando origen al modelo multidimensional Logit multinomial de coeficientes aleatorios (MRCML, por sus siglas en inglés). Este modelo generalizado permite, a través de las matrices de diseño y de puntuación, representar la parametrización de los distintos modelos Rasch. Para ello, los patrones de respuesta a un conjunto de ítems son considerados una variable dependiente. Las variables independientes son la dificultad de los ítems y la habilidad de las personas, que se relacionan con la variable dependiente a través de la regresión logística. Además, se asume que los ítems son descritos a partir de un conjunto fijo y desconocido de parámetros ξ, mientras que la habilidad de los estudiantes (la variable latente θ) es un efecto aleatorio (Adams y Wu, 2007).

Wang, Wilson y Adams (1997) señalan que para formular matemáticamente el MRCLM se asume que hay dimensiones subyacentes al conjunto de respuestas que un grupo de personas da a un conjunto de ítems. La posición de esas personas en

el espacio latente es recogida mediante $\boldsymbol{\theta} = (. \theta_1,..., \theta_D)$. En la mayoría de casos se asume una distribución normal multivariada como modelo poblacional para θ; sin embargo, estos autores señalan que es posible utilizar otras formas de distribución poblacional.

Además se suponen *I* ítems, indexados como $i = 1, ..., I$, cada uno con K_i+1 alternativas de respuesta, indexadas como $K = 0,1,...,Ki$. Una respuesta en la categoría *K*, de la dimensión *d*, en un ítem i, es calificada como b_{ikd}. Las puntuaciones en las *D*dimensiones se recogen en un vector columna, $\boldsymbol{b}_{ik} = (b_{ik1},..., b_{ikD})'$, y luego pueden conformar la submatriz del ítem *i*, $\boldsymbol{B}_i = (\boldsymbol{b}_{i1},...,\boldsymbol{b}_{iD})'$, las cuales finalmente se integran en la matriz de puntuaciones $B = (\boldsymbol{B}'_1,...,\boldsymbol{B}'_I)'$.

Siguiendo con la propuesta de estos autores, para describir las propiedades de los ítems se utiliza un vector de *p* parámetros $\xi = (\xi_1,...,\xi_p)'$. Las combinaciones lineales de esos parámetros son utilizadas en los modelos de respuesta probabilística para describir las características empíricas de las categorías de respuesta de cada ítem. Esas combinaciones lineales se definen por un conjunto de vectores de diseño $\mathbf{a}_{ik}$, $(i = 1,...,I; K=1,K_i)$ cada uno de largo *p*. Esos vectores, también, pueden ser recogidos en la matriz de diseño $\boldsymbol{A} = (\mathbf{a}_{11},...,\mathbf{a}_{1k1}, \mathbf{a}_{21},..., \mathbf{a}_{2k1},..., \mathbf{a}_{Iki})'$. Además, se debe definir una variable indicador X_{ik}, de tal manera que:

$$X_{ik} = \begin{cases} 1, \text{ si la respuesta del ítem } i \text{ está en la categoría } K \\ 0, \text{ en cualquier otro caso} \end{cases}$$

A partir de todo lo anterior, se define el modelo (Adams, Wilson y Wang, 1997; Wang, Wilson y Adams, 1997):

$$f(X_{ik} = 1; \mathbf{A}; \mathbf{B}; \xi| \boldsymbol{\theta}) = \frac{\exp(\mathbf{b}' K\boldsymbol{\theta} + \mathbf{a}'_{ik} \xi)}{\sum_{u=1}^{K_i} \exp(\mathbf{b}'_{iu} \boldsymbol{\theta} + \mathbf{a}'_{iu} \xi)}$$

Modelo Rasch de testlets

Algunos de los test que se aplican en el contexto de la evaluación psicológica o educativa pueden estar estructurados a modo de testlets. Tanto Wainer y Lewis (1990) como Lee, Brennan y Frisbie (2000) señalan que el concepto de testlet fue presentado por primera vez en el contexto de los test adaptativos informatizados (TAI) por Wainer y Kiely en 1987. Sin embargo, Wilson y Adams (1995) identifican un antecedente más antiguo, al señalar que Cureton, en 1965, introdujo el término superítem —que parece no haber ganado tanta difusión— para referirse a la manera de estructurar los ítems dentro de un test, que actualmente suelen denominarse testlets.

Más allá del origen del término, es importante considerar su significado. Específicamente, se usa para referirse a aquellas situaciones en las cuales algunos ítems no son independientes entre sí, pues comparten algo en común, más allá del constructo o rasgo latente que se pretende medir con ellos. Jiao, Wang y He (2013) señalan que la respuesta de una persona a un ítem en un testlet puede afectar la probabilidad de respuesta de ese mismo evaluado a otro ítem en el testlet, violando de esta manera el supuesto de independencia local. Por ejemplo,

algunos tests consisten en pasajes, párrafos, mapas u otros materiales de estímulo que tienen un conjunto de ítems asociados, de tal manera que dicho conjunto es dependiente de ese estímulo común (Lee, Brennan y Frisbie, 2000; Wilson y Adams, 1995).

Los modelos básicos de la TRI asumen la independencia local y la estimación de la confiabilidad de las medidas es realizada bajo dicho supuesto (Sireci, Thissen y Wainer, 1991). Sin embargo, en el caso de los testlets hay fuentes obvias de dependencia local. Por ejemplo, en un test de comprensión lectora el conocimiento previo del tópico tratado en el texto, las habilidades específicas requeridas para comprender ese pasaje de lectura, el nivel de interés u otros factores motivacionales específicos generan dependencia entre los ítems derivados del mismo texto (De Mars, 2006).

Dadas estas consideraciones, el análisis psicométrico es más adecuado si se modela explícitamente dicha dependencia local. A pesar de la diversidad de modelos disponibles para ello, autores como Paek, Yon, Wilson y Kang (2009) consideran que, comparado con otros modelos de testlets como el bayesiano, que necesita supuestos de distribución para todos los parámetros del modelo, el modelo testlet de Rasch requiere solo supuestos de distribución para la habilidad de las personas y el efecto de los testlets (γ). Dicho modelo puede ser considerado un submodelo del Modelo Multidimensional Logístico Multinomial de Coeficientes Aleatorios (MRCMLM; Adams, Wilson y Wang, 1997) y, por lo tanto, ser calibrado y estimado dentro de dicho marco de referencia, aplicando el método de máxima verosimilitud marginal para estimar sus parámetros (Wilson y Adams, 1995). Por ello, ese será el modelo que se tomará para analizar los datos provenientes de la prueba de comprensión lectora aplicada a los estudiantes de segundo grado de primaria en la Evaluación Censal de Estudiantes del año 2015 (ECE 2015). Wang y Wilson (2005) formulan el modelo Rasch de testlets para ítems dicotómicos de la siguiente manera:

$$P_{ni1} = \frac{\exp(\theta_n - b_i + Y_{nd(i)})}{1+\exp(\theta n - b_i + Y_{nd(i)})}$$

Estos mismos autores indican que la ecuación anterior también puede ser expresada así:

$$\log(P_{ni} / P_{ni0}) = \theta n - b_i + Y_{nd(i)}$$

donde P_{ni1} y P_{ni0} son las probabilidades de puntuar 1 y 0 en el ítem *i* para la persona *n*, respectivamente. Si *Ynd(I)*=0 (sin efectos testlet), la ecuación se reduce al modelo Rasch para ítems dicotómicos. Además, dicha ecuación puede ser extendida a ítems politómicos y expresada como:

$$\log(p_{nij} / (p_{ni(j-1)}) = \theta n - b_i + \gamma_{nd(i)})$$

donde p_{nij} y $P_{ni(j-1)}$ son las probabilidades de puntuar *j* y *j-1* para el ítem *i* en la persona *n*, respectivamente, y es el paso *j-ésimo* de dificultad en el ítem *i*.

Wang y Wilson (2005) señalan que en este modelo los ítems son tratados como efectos fijos; el rasgo latente θ y el efecto del testlet γs, asumidos como variables aleatorias distribuidas normal e independientemente, y no se realizan suposiciones distribucionales sobre $\sigma^2_{\gamma d(i)}$. Dado que θ y γ son asumidos como independientes y distribuidos normalmente bajo el modelo testlet, $\theta' = [\theta,\gamma 1,\ldots,\gamma 2,\ldots,\gamma_D]$, tiene una distribución normal multivariada $N(\mu,\Sigma)$ y, por necesidad de identificación del modelo, μ es fijado en cero y Σ restringido a ser una matriz diagonal. Además, Jiao, Wang y He (2013) señalan que la magnitud del efecto de un testlet está representada por $\sigma^2_{\gamma jd(i)}$, que es la varianza del parámetro $\gamma^2_{jd(i)}$. Si esta varianza es cero, los ítems en el testlet son condicionalmente independientes; mientras mayor es la varianza, también lo es la dependencia local (Wainer y Wang, 2000).

Modelos de diagnóstico cognitivo

Según De la Torre (2009), los modelos de diagnóstico cognitivo (en adelante, MDC) tienen como objetivo diagnosticar la competencia de los evaluados a lo largo de un conjunto de múltiples habilidades discretas/dicotómicas; específicamente predicen la probabilidad de una respuesta categórica observable (acertar o fallar un ítem) a partir de variables categóricas inobservables (es decir, latentes), que han sido denominadas de diferentes maneras: habilidad, subhabilidad, atributo, conocimiento, capacidad, procesos y estrategias. A partir de ello, el perfil individual de dominio de atributos de cada estudiante puede ser producido y comunicado usando métodos escritos y gráficos; con ello se busca incluir información de retroalimentación diagnóstica sobre la fortalezas y debilidades del estudiante; además, pueden recomendarse estrategias remediales que buscan superar dichas debilidades (Lee y Sawaki, 2009). Otra característica importante es que a diferencia de los modelos Rasch o TRI, los cuales se basan en las expectativas que tienen los investigadores sobre qué procesos cognitivos siguen los evaluados para resolver los problemas en los test, los MDC buscan recolectar evidencia empírica de los procesos y estrategias realmente seguidos en estas situaciones (Junker y Sijtsma, 2001). En esta línea, Lee y Sawaki (2009) señalan que, de forma general, los MDC suelen tener dos componentes principales: a) el análisis de contenido de ítems del test para identificar sus relaciones con los atributos cognitivos de interés, y b) el modelado psicométrico de estos atributos e ítems.

En cuanto a la identificación de los atributos cognitivos, es importante considerar que estos son definidos *a priori* por una matriz *Q* que representa la estructura de cargas de las habilidades requeridas para resolver los ítems (Tatsuoka, 1985; Chiu, 2013). La especificación de una matriz *Q* se hace usando un esquema binario 0 / 1 para indicar si una habilidad está o no presente en la solución de un ítem. Integrando las propuestas de diversos autores (Chiu, 2013; Lee y Sawaki, 2009; Liu, Xu y Ying, 2012; Ravand y Robitzsch, 2015), sus principales características son:

- Tiene tantas filas como la cantidad de ítems (J) del test y tantas columnas como la cantidad de atributos (K) subyacentes al desempeño en el test.
- La matriz binaria de *J x K* dimensiones indica si el ítem *j (j=1,...,J)* requiere que un examinado posea un atributo *k* (k=1,…K) para responder correctamente a ese ítem.
- Cada celda q_{jk} de la matriz *Q* es codificada como *0* (habilidad no requerida) o *1* (habilidad requerida).
- Una fila o vector q_j especifica todos los atributos que un examinado debe poseer para responder correctamente al ítem *j*.
- Para un test que mide *K* atributos, hay 2^K clases de competencia (perfil de atributos) posibles.

Si bien existen diversos MDC, las pruebas ECE 2015 serán trabajadas a partir del modelo generalizado G-DINA, cuya formulación matemática es la siguiente:

$$P(X_{ij}|\ \alpha_1, \ldots, \alpha_k) = \delta_{j0} + \sum_{k=1}^{k} \delta_{jk}\alpha_{lk} + \sum_{k'=k+1}^{k} \sum_{k=1}^{k-1} \delta_{jkk'}\alpha_{lk}\alpha_{lk'} \ldots + \delta_{j12\ldots k} \prod_{k=1}^{k} \alpha_{lk}$$

Con $P(X_{ij}|\alpha_1,\ldots,\alpha_k)$ se presenta la probabilidad de responder correctamente un ítem, dadas *k* habilidades cognitivas (α). El parámetro δ_{j0} representa la dificultad general de un ítem; $\sum^k = \delta_{jk}\ \alpha_{lk}$, los efectos principales de las habilidades cognitivas consideradas en el modelo, y $\sum_{k'=k+1}\sum_{k=1}\delta_{jkk'}\alpha_{lk}\alpha_{lk'}\ldots+\delta_{j12}\ldots k\ \prod_{k=1}\alpha_{lk}$ resume los efectos de las interacciones entre dichas habilidades cognitivas.

Método

Tipo de investigación

La presente investigación sigue un diseño no experimental, de tipo psicométrico. Dicho tipo de investigación es definido, por Alarcón (2008), como el que se focaliza en el análisis y estandarización de test psicológicos o educativos. En este caso, se trata de probar y comparar el ajuste de distintos modelos psicométricos al mismo conjunto de datos.

Participantes

Las evaluaciones censales están dirigidas a todos los estudiantes de segundo grado de primaria que asisten a instituciones educativas del programa de educación básica regular; las cuales, a su vez, atienden a cinco o más estudiantes (Minedu, 2009). En el caso de la ECE 2015, se planificó evaluar a un poco más de 500 000 estudiantes a nivel nacional.

Del conjunto total de estudiantes evaluados, se tomó una muestra aleatoria de 5000 estudiantes, con la finalidad

de que las calibraciones de los ítems con los diferentes modelos psicométricos no demanden demasiados recursos computacionales.

Instrumentos

Las pruebas aplicadas en una evaluación censal corresponden a las áreas de comprensión lectora y matemática (Minedu, 2009). Esos instrumentos tienen 46 y 42 ítems respectivamente y poseen evidencias de validez vinculadas al contenido dadas por el juicio de expertos de la Dirección de Educación Primaria y del Instituto Peruano de Evaluación, Acreditación y Certificación de la Calidad de la Educación Básica:

> Ellos evalúan aspectos como la calidad, actualidad y veracidad de la información según cada disciplina científica, la correspondencia con la tabla de especificaciones, la adecuación de la complejidad del ítem a la población evaluada, y la construcción del enunciado y las alternativas, tanto en sus aspectos formales como en su eficacia para la medición del constructo. También se toman en cuenta posibles sesgos socioeconómicos, culturales y de género en la construcción". (Minedu, 2014, p. 48)

Asimismo, cuentan con evidencias de validez vinculadas a la estructura interna, dadas por el ajuste de los datos a un modelo Rasch unidimensional, tal y como lo muestra el resultado del análisis de componentes principales de los residuos del modelo, cuyo primer autovalor es de 2.3 o menor, y representa menos del 3.5 % de la varianza (Minedu, 2014). Finalmente, la confiabilidad de las medidas es estimada mediante el índice de separación de personas, que tiene un valor de .85 para comprensión lectora y .88 en matemáticas.

Resultados

Todos los análisis psicométricos propuestos se realizaron mediante *software* libre, ya que se encuentran disponibles diferentes paquetes de *R* (https://cran.r-project.org/) para aplicarlos, como el caso de TAM v.1.15 (https://cran.r-project.org/web/packages/TAM/index.html), SIRT v.1.9 (https://cran.r-project.org/web/packages/sirt/index.html) y CDM v.4.6 (https://cran.r-project.org/web/packages/CDM/index.html). Se mostrarán los resultados de cada una de las pruebas analizadas. En primer lugar, se presentarán los resultados de lectura y, luego, los de matemáticas.

Prueba de lectura

Es posible modelar de distintas maneras la multidimensionalidad que podría estar contenida en un instrumento de medición. En la tabla 1 se muestra la confiabilidad estimada para las distintas dimensiones consideradas en la prueba de lectura, al aplicar la estrategia secuencial, que implica modelar cada dimensión de forma independiente, y la del modelo multidimensional. En el caso del modelo secuencial, los coeficientes fluctúan entre .27 y .84; mientras que en el modelo multidimensional, entre .67 y .87.

Tabla 1

Confiabilidad de las dimensiones de la prueba de lectura

Dimensión	Secuencial	Multidimensional
1. Obtiene información	.63	.87
2. Hace inferencias	.84	.84
3. Reflexión	.27	.67

Nota: La estimación de la confiabilidad para el modelo unidimensional fue .88.

De igual modo, en la tabla 2 se pueden apreciar los coeficientes de correlación de Pearson entre las medidas derivadas de aplicar los ítems que conforman cada una de las dimensiones de la prueba de lectura. Esas correlaciones varían entre .80 y .95.

Tabla 2

Correlación entre las dimensiones de la prueba de lectura

Dimensión	1	2	3
1. Obtiene información	1.205		
2. Hace inferencias	.95	1.589	
3. Reflexión	.83	.80	1.007

Nota: la diagonal muestra la desviación estándar de las medidas de cada dimensión, cuya media aritmética es 0 en todos los casos.

Otro de los objetivos de esta investigación fue aplicar un modelo Rasch de testlets a la prueba de lectura. Ello se debe a que está organizada de tal manera que de un texto se desprenden varias preguntas, lo cual genera dependencia local. Concretamente, existen doce textos y cuarentaiséis preguntas. La cantidad de dependencia local fue analizada mediante el estadístico *Q3*, cuyos resultados aparecen en las tablas 3 y 4:

Tabla 3

Análisis de dependencia local con el método Q3 de la prueba de lectura

Estadístico	Valor
M	- 0.019
DE	0.050
mínimo	-0.110
máximo	0.401

Tabla 4

Análisis de dependencia local con el método Q3 de la prueba de lectura

Testlet	N.ítems	Q3 promedio
A	3	.28
B	4	< .01
C	7	.02
D	2	.18
E	5	.01
F	2	.17
G	3	.33
H	3	.05
I	2	.12
J	4	.04
K	6	.04
L	5	.04

Por último, la tabla 5 presenta el coeficiente de confiabilidad estimado mediante el modelo unidimensional y el estimado con el modelo de testlets.

Tabla 5

Confiabilidad de los modelos aplicados a la prueba de lectura

Modelo	Confiabilidad
Unidimensional con independencia local	.88
Rasch de testlets	.85

En cuanto a los análisis efectuados con el modelo de diagnóstico cognitivo, uno de los requisitos previos es poseer una matriz *Q*. Para ello, se tuvieron entrevistas con expertos de la Unidad de Medición de la Calidad de los Aprendizajes (UMC), del Ministerio de Educación del Perú. Considerando que tanto las pruebas de lectura como las de matemáticas se agrupan en capacidades, y tomando en cuenta parte de las demandas que han podido recoger de los docentes que asisten a los talleres de devolución de resultados a nivel de dichas capacidades; se decidió estructurar la matriz *Q* estimando las capacidades evaluadas mediante los ítems que conforman las pruebas aplicadas durante la ECE 2015. Es decir, se siguió la misma estructura considerada al momento de ajustar el modelo Rasch multidimensional.

En general, se puede observar, en la tabla 6, que todas las capacidades evaluadas en esta muestra tienen una probabilidad

mayor a .50 de ser dominadas. Además, los perfiles de capacidades más usuales son aquellos en el que se dominan todas las competencias o aquel en el cual no se domina ninguna de ellas, tal y como aparece en la tabla 7.

Tabla 6

Probabilidad de dominio de las capacidades consideradas en la prueba de lectura

Capacidades	P(0)	P(1)
1. Obtiene información	.360	.640
2. Hace inferencias	.286	.714
3. Reflexión	.411	.589

Tabla 7

Probabilidad de ocurrencia de los perfiles de capacidades en la prueba de lectura

Perfil	P
000	.280
100	.003
010	.073
110	.054
001	.003
101	.000
011	.004
111	.583

A continuación, la tabla 8 presenta las correlaciones entre las capacidades estimadas con el modelo G-DINA, las cuales son bastante altas y similares entre sí.

La Tabla 9 contiene los indicadores de ajuste de los ítems a los diferentes modelos considerados como parte de esta investigación.

Tabla 8

Correlación entre las dimensiones de la prueba de lectura

Capacidades	1	2	3
1. Obtiene información	-		
2. Hace inferencias	.98	-	
3. Reflexión	.98	.97	-

Tabla 9

Indicadores de ajuste de los ítems, considerando los distintos modelos psicométricos aplicados a la prueba de lectura

	UNIDIM.		SECUENCIAL		MULTIDIM.		TESTLET		MDC
item	infit	outfit	infit	outfit	infit	outfit	infit	outfit	rmsea
i01	0.93	0.57	0.97	0.76	0.95	0.68	0.47	0.94	0.004
i02	0.89	0.61	0.90	0.63	0.89	0.55	0.56	0.86	0.038
i03	0.87	0.65	0.91	0.80	0.90	0.71	0.66	0.88	0.016
i04	0.90	0.65	0.91	0.66	0.91	0.69	0.63	0.88	0.039
i05	0.97	0.91	0.96	0.91	0.96	0.89	0.93	0.98	0.050
i06	0.94	0.90	0.98	0.92	0.99	1.08	1.01	0.94	0.009
i07	1.16	1.22	1.14	1.20	1.15	1.21	1.24	1.16	0.022
i08	1.11	1.16	1.10	1.14	1.12	1.16	1.17	1.11	0.025
i09	0.93	0.86	0.95	0.90	0.92	0.83	0.86	0.94	0.058
i10	1.00	0.98	1.09	1.16	1.10	1.15	1.00	1.01	0.082
i11	0.97	0.89	0.96	0.88	0.96	0.89	0.91	0.97	0.034
i12	1.08	1.12	1.06	1.09	1.07	1.11	1.11	1.09	0.018
i13	1.01	1.06	1.01	1.04	1.01	1.03	1.04	1.01	0.025
i14	1.00	1.00	1.00	1.00	0.99	0.99	1.03	1.02	0.028
i15	0.97	0.94	0.96	0.92	0.97	0.93	0.95	0.98	0.025
i16	0.94	0.81	0.98	0.97	0.99	1.04	0.81	0.94	0.004
i17	0.97	0.90	1.03	0.98	1.04	1.03	0.89	0.97	0.017
i18	0.96	0.95	0.97	0.97	0.95	0.96	0.99	0.96	0.072
i19	0.87	0.59	0.89	0.61	0.87	0.61	0.43	0.82	0.056
i20	1.05	1.05	1.04	1.04	1.04	1.03	1.07	1.06	0.040
i21	1.17	1.25	1.15	1.23	1.16	1.25	1.30	1.17	0.016
i22	0.87	0.70	0.87	0.71	0.87	0.71	0.73	0.90	0.057
i23	1.06	1.08	1.05	1.06	1.06	1.09	1.10	1.05	0.025
i24	1.00	0.99	0.98	0.98	0.99	0.99	1.01	0.99	0.016
i25	1.00	0.95	0.99	0.94	1.00	0.94	0.92	1.00	0.018
i26	0.93	0.85	0.98	0.95	0.97	1.02	0.82	0.93	0.007
i27	1.21	1.36	1.19	1.32	1.19	1.33	1.42	1.17	0.028
i28	0.97	0.92	0.97	0.92	0.97	0.93	0.96	0.97	0.050
i29	0.86	0.68	0.91	0.80	0.90	0.80	0.70	0.87	0.014
i30	1.00	0.97	0.99	0.96	0.99	0.96	1.01	1.01	0.034
i31	0.89	0.64	0.90	0.65	0.89	0.65	0.52	0.87	0.048
i32	0.94	0.79	1.01	0.95	1.01	0.88	0.79	0.93	0.017

(continúa)

(continuación)

	UNIDIM.		SECUENCIAL		MULTIDIM.		TESTLET		MDC
item	infit	outfit	infit	outfit	infit	outfit	infit	outfit	rmsea
i33	0.88	0.79	0.90	0.81	0.88	0.77	0.78	0.89	0.083
i34	0.99	1.00	1.04	1.05	1.06	1.35	0.97	0.99	0.005
i35	1.03	1.01	1.03	1.00	1.02	1.00	0.97	1.01	0.049
i36	0.99	0.98	1.06	1.14	1.08	1.15	0.99	1.00	0.023
i37	0.87	0.69	0.87	0.68	0.86	0.67	0.63	0.86	0.063
i38	0.92	0.76	0.91	0.76	0.91	0.75	0.78	0.92	0.036
i39	1.04	1.06	1.02	1.04	1.03	1.06	1.09	1.06	0.020
i40	0.88	0.55	0.91	0.55	0.89	0.72	0.40	0.88	0.009
i41	1.15	1.25	1.00	1.00	1.03	1.04	1.31	1.16	0.015
i42	1.01	1.03	1.00	1.02	1.00	1.01	1.02	1.00	0.039
i43	0.98	0.94	0.98	0.93	0.97	0.92	0.94	0.97	0.053
i44	1.06	1.06	0.99	0.99	0.97	0.94	1.08	1.07	0.021
i45	1.15	1.26	1.00	1.00	1.03	1.04	1.27	1.15	0.015
i46	1.03	1.08	1.03	1.08	1.03	1.06	1.10	1.04	0.036

Prueba de matemáticas

Igual que en el caso de la prueba de lectura, los datos fueron modelados aplicando la estrategia secuencial y el modelo multidimensional. En el caso del modelo secuencial, los coeficientes fluctúan entre .34 y .89, mientras que en el multidimensional varían entre .76 y .90.

Tabla 10

Confiabilidad de las dimensiones de la prueba de matemáticas

Dimensión	Secuencial	Multidimensional
1. Comunicación matemática	.54	.81
2. Razonamiento y demostración	.34	.76
3. Resolución de problemas	.89	.90

Nota: la estimación de la confiabilidad para el modelo unidimensional fue .90.

La tabla 11 presenta los coeficientes de correlación de Pearson entre las medidas que operacionalizan las tres dimensiones evaluadas mediante la prueba de matemáticas. Esas correlaciones fluctúan entre .86 y .93.

Tabla 11

Correlación entre las dimensiones de la prueba de matemáticas

Dimensión	8. 1	9. 2	10.3
11.1.Comunicación matemática	1.079		
2. Razonamiento y demostración	.86	1.706	
3. Resolución de problemas	.93	.90	1.324

Nota: la diagonal muestra la desviación estándar de las medidas de cada dimensión, cuya media aritmética es 0 en todos los casos.

En el caso de la prueba de matemática, al momento de ajustar el modelo de diagnóstico cognitivo, también se decidió estructurar la matriz *Q* usando la misma estructura considerada al momento de aplicar el modelo Rasch multidimensional. En la tabla 12 se observa que todas las capacidades evaluadas en esta muestra tienen una probabilidad mayor a .50 de ser dominadas. Además, los perfiles de capacidades más usuales son aquellos en los que se dominan todas las competencias o aquel en el cual no se domina ninguna de ellas, tal y como aparece en la tabla 13.

Tabla 12

Probabilidad de dominio de las capacidades consideradas en la prueba de matemáticas

Capacidades	12. P(0)	13. P(1)
Comunicación matemática	.399	.601
2. Razonamiento y demostración	.310	.690
3. Resolución de problemas	.402	.598

Tabla 13

Probabilidad de ocurrencia de los perfiles de capacidades en la prueba de matemáticas

Peril	P		
000	.297	001	.008
100	.002	101	.003
010	.078	011	.016
110	.025	111	.571

La tabla 14 presenta las correlaciones entre las capacidades estimadas con el modelo G-DINA, las cuales son bastante altas y similares entre sí.

Tabla 14

Correlación entre las dimensiones de la prueba de matemáticas

Capacidades	1	2	3
1. Comunicación matemática	-		
2. Razonamiento y demostración	.98	-	
3. Resolución de problemas	.99	.96	-

La tabla 15 contiene los indicadores de ajuste de los ítems a los diferentes modelos considerados como parte de esta investigación.

Tabla 15

Indicadores de ajuste de los ítems, considerando los distintos modelos psicométricos aplicados a la prueba de matemáticas

	UNIDIM.		SECUENCIAL		MULTIDIM.		MDC
item	infit	outfit	infit	outfit	infit	outfit	rmsea
i01	1.06	1.28	1.09	1.32	1.08	1.35	0.044
i02	0.95	0.93	0.96	0.96	0.96	0.95	0.048
i03	0.86	0.76	0.87	0.76	0.86	0.77	0.062
i04	0.90	0.80	0.90	0.80	0.90	0.80	0.024
i05	0.81	0.71	0.82	0.70	0.82	0.71	0.028
i06	0.94	0.88	0.95	0.90	0.95	0.90	0.055
i07	0.85	0.80	0.93	0.91	0.83	0.77	0.023
i08	0.84	0.73	0.86	0.76	0.85	0.74	0.049
i09	0.95	0.97	0.96	1.00	0.96	0.97	0.036
i10	0.94	0.90	0.96	0.92	0.96	0.92	0.032
i11	1.02	1.03	1.03	1.05	1.04	1.08	0.045
i12	0.98	0.93	1.00	0.98	1.00	0.96	0.057
i13	0.94	0.71	0.96	0.73	0.95	0.72	0.031
i14	1.02	1.00	1.04	1.04	1.04	1.05	0.035
i15	1.03	1.17	1.05	1.21	1.06	1.23	0.032
i16	0.95	0.85	0.94	0.87	0.91	0.77	0.030
i17	0.98	0.96	1.00	0.98	1.00	0.98	0.035
i18	1.14	1.23	1.01	1.02	1.05	1.08	0.013
i19	0.82	0.71	0.83	0.73	0.82	0.72	0.054

(continúa)

(continuación)

i20	0.85	0.77	0.86	0.78	0.86	0.77	0.030
i21	1.03	1.06	1.05	1.08	1.05	1.09	0.030
i22	0.85	0.65	0.85	0.67	0.85	0.67	0.063
i23	0.91	0.89	0.93	0.91	0.92	0.92	0.071
i24	1.00	0.98	1.00	0.98	1.02	0.99	0.025
i25	0.90	0.79	0.99	0.97	0.94	0.86	0.013
i26	0.96	0.89	0.97	0.86	0.94	0.77	0.030
i27	0.93	0.89	0.94	0.90	0.95	0.91	0.024
i28	1.04	0.98	0.97	0.90	0.99	0.88	0.011
i29	0.90	0.82	0.91	0.84	0.91	0.82	0.024
i30	0.96	0.94	1.00	0.98	1.00	1.05	0.006
i31	0.85	0.78	0.86	0.78	0.86	0.79	0.067
i32	1.30	2.45	1.34	2.59	1.34	2.67	0.030
i33	0.93	0.84	0.94	0.85	0.94	0.88	0.022
i34	0.97	0.92	0.99	0.95	0.99	0.93	0.037
i35	1.43	1.75	1.46	1.85	1.48	1.88	0.040
i36	1.06	1.05	0.99	0.98	0.98	0.96	0.024
i37	1.06	1.07	1.07	1.09	1.08	1.10	0.030
i38	1.20	1.36	1.04	1.05	1.09	1.17	0.017
i39	1.06	1.09	1.09	1.14	1.08	1.13	0.028
i40	1.31	2.08	1.10	1.28	1.20	1.57	0.021
i41	1.10	1.18	1.13	1.23	1.13	1.23	0.034
i42	1.03	1.03	1.04	1.04	1.04	1.05	0.025

Discusión

Al analizar los resultados del modelo multidimenional, se puede apreciar que tanto en el caso de la prueba de lectura como en la de matemáticas, las tres dimensiones modeladas están altamente correlacionadas, pues todos los coeficientes superan el valor de .80. Es decir, no son dimensiones radicalmente diferenciadas entre sí. Hay una alta consistencia entre los diferentes aspectos que se combinan para dar lugar a las capacidades lectoras y matemáticas. Además, tal y como lo señalan Adams, Wilson y Wang (1997), la confiabilidad de las medidas estimadas mediante el método de EAP de cada dimensión son mayores si se aplica un modelo multidimensional, que si se considerara un modelo secuencial. Si bien hay notables mejoras en la confiabilidad de las medidas, en especial de la dimensión *reflexión* de la prueba de lectura, esta aún sigue siendo algo baja.

Esto podría explicarse por el reducido número de ítems que conforman esta subdimensión.

Al aplicar el modelo de testlets, tal y como se puede apreciar en los trabajos de diversos autores, como Sireci, Thissen y Wainer (1991) o Wang y Wilson (2005), la confiabilidad estimada para las medidas es mayor cuando no se modela explícitamente la dependencia de los ítems dentro de un testlet. Sin embargo, la diferencia entre ambas estimaciones no es muy grande. Al analizar los resultados del estadístico *Q3*, se puede apreciar que en promedio no hay mucha dependencia local entre los ítems de la prueba de lectura.

El testlet con mayor dependencia local fue el *G* (.33), seguido por el *A* (.28). El resto de testlets muestra un *Q3* promedio inferior a .20. Esto podría explicar la poca diferencia entre la estimación de la confiabilidad entre ambos métodos, pues el modelo de testlets en general trata de captar la dependencia entre los ítems y, al ser poca, no habría mucha diferencia entre ambas estimaciones, como ocurre en el presente caso. En ese sentido, si la confiabilidad basada en la consistencia interna es calculada a nivel de ítem, los dos niveles de correlación (dentro del testlet y entre testlets) son promediados, y el resultado generalmente es un valor mayor que la correlación promedio, que sería obtenida entre ítems que no muestran esos dos niveles de correlación (Sireci, Thissen y Wainer, 1991). Por lo tanto, es importante cuantificar la cantidad de dependencia local que se encuentra entre un conjunto de ítems.

Al analizar los resultados del modelo de diagnóstico cognitivo, se aprecia que, tanto en el caso de la prueba de lectura como en la de matemáticas, las tres capacidades modeladas están altamente correlacionadas. Además, en cuanto a las probabilidades de cada uno de los perfiles posibles de dominio de las competencias consideradas, tanto en lectura como matemáticas, podemos apreciar que los perfiles más usuales son los de dominio de todas las capacidades o de ninguna de ellas. Este resultado puede explicarse por la alta correlación que existe entre las habilidades consideradas en el modelo aplicado. Este resultado es muy similar al encontrado en el caso del modelo multidimensional y aporta evidencias adicionales a favor respecto a que resulta totalmente pertinente aplicar un modelo unidimensional a los datos derivados de aplicar las pruebas ECE, pues las distintas capacidades funcionan de manera homogénea.

Con relación al ajuste de los ítems, en el caso de los modelos Rasch, este se evaluó mediante los indicadores *infit* y *outfit*, y se prefirió los valores que están entre 0.70 y 1.30 (Bond y Fox, 2007). En el caso del modelo de diagnóstico cognitivo, se utilizó como indicador de ajuste el RMSEA. Según este indicador, un valor menor indica una mejor adecuación del ítem al modelo. En términos generales, se prefieren valores inferiores a .05.

En todos los modelos Rasch, se observa congruencia en cuanto a la detección de los ítems con desajuste. En el caso de lectura, todos los modelos Rasch

detectan el desajuste del ítem 27, mientras que el ítem 34 solo desajusta al considerar el modelo multidimensional. Destaca el hecho de que el modelo de diagnóstico cognitivo detecta el desajuste en muchos más ítems, concretamente en nueve de ellos. Cabe destacar que ningunos de estos ítems con desajuste coincide con alguno de los dos detectados mediante los modelos Rasch.

Una situación similar se aprecia con la prueba de matemática. Los ítems 32 y 35 tienen desajuste en todos los modelos Rasch considerados, el ítem 38 solo desajusta en el modelo unidimensional, y el ítem 40, en el caso del modelo unidimensional y el multidimensional. Ya que los modelos tienen supuestos distintos y están formulados desde perspectivas diferentes (en especial, el modelo de diagnóstico cognitivo), no es de extrañar este resultado.

Para finalizar, es importante señalar que un modelo, en general, no es una representación verdadera de la realidad. Tal como señala Reckease (2009), lo importante es mostrar cómo esos modelos son idealizaciones útiles de las relaciones hipotetizadas entre los rasgos latentes y las repuestas a los ítems de un test.

Considerando que las pruebas ECE se construyen todos los años siguiendo las mismas tablas de especificaciones, es posible que los resultados observados en las pruebas del año 2015 puedan ser replicados, si se aplican a las pruebas de los años anteriores o a la del 2016. Sin embargo, no es posible generalizar estos resultados a otras pruebas de lectura o matemáticas.

Conclusiones

- Dadas las altas correlaciones entre las dimensiones estimadas con el modelo multidimensional y las altas correlaciones entre las capacidades estimadas con el modelo de diagnóstico cognitivo, resulta totalmente pertinente aplicar un modelo unidimensional a los datos derivados de aplicar las pruebas ECE, estrategia que desde hace varios años viene aplicando la Unidad de Medición de la Calidad de los Aprendizajes del Ministerio de Educación.
- Tal y como se ha demostrado en otras investigaciones precedentes, el dejar de lado la estructura de testlets que presentan los ítems de una prueba lleva a la subestimación de las medidas derivadas de aplicar dicha prueba.
- Si existiera la necesidad política o pedagógica de dar resultados más desagregados (por dimensiones), se debería aumentar el número de ítems de la dimensión de reflexión de la prueba de lectura y de la de *razonamiento* y demostración de la prueba de matemáticas, con la finalidad de tener mediciones más confiables para referidas dimensiones.
- Si bien existe cierta consistencia entre los modelos Rasch, al señalar qué ítems no se ajustan bien al modelo psicométrico, el modelo de diagnóstico cognitivo señala el desajuste en conjuntos distintos de ítems, pues tiene supuestos muy diferentes a los de los modelos Rasch.

Referencias

Adams, R. J., y Wilson, M. (1996). Formulating the Rasch model as a mixed coefficients multinomial logit. En G. Engelhard y M. Wilson (Eds.), *Objective measurement: theory into practice*, (vol. 3, pp. 143-166). Norwood, NJ: Ablex.

Adams, R. J., y Wu, M. L. (2007). The mixed-coefficients multinomial logit model: a generalized form of the Rasch model. En M. von Davier y C. H. Carstensen (Eds.), *Multivariate and mixture distribution Rasch models. Extension and applications*, (pp. 57-76). New York: Springer.

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

Alarcón, R. (2008). Métodos y diseños de investigación del comportamiento. Lima: Universidad Ricardo Palma.

Bond, T. G., y Fox, C. M. (2007). Applying the Rasch Model: Fundamental measurement in the Human Sciences (2.ª ed.). Nueva Jersey: Lawrence Earlbaum Associates.

Chiu, C. (2013). Statistical refinement of the Q-matrix in cognitive diagnosis. *Applied Psychological Measurement*, *37*(8), 598-618. http://doi.org/10.1177/ 0146621613488436.

De Ayala, R. J. (2009). *The theory and practice of item response theory*. New York, NY: The Guilford Press.

De la Torre, J. (2009). A cognitive diagnosis model for cognitively based multiple-choice options. *Applied Psychological Measurement*, *33*(3), 163-183. http://doi.org/10.1177/0146621608320523.

De Mars, C. E. (2006). Application of the bi-factor multidimensional Item Response Theory model to testlet-based tests. *Journal of Educational Measurement, 43*(2), 145-68. http://dx.doi.org/10.1111/j.1745-3984.2006.00010.x.

Jiao, H., Wang, S., y He, W. (2013). Estimation methods for one-parameter testlet models. *Journal of Educational Measurement, 50*(2), 186-203. http://dx.doi.org/ 10.1111/jedm.12010.

Junker, B. W., y Sijtsma, K. (2001). Cognitive assessment models with few assumptions, and connections with nonparametric item response theory. *Applied Psychological Measurement, 25(*3), 258-272. http://doi.org/10.1177/01466210122032064.

Lee, G., Brennan, R. L, y Frisbie, D. A. (2000). Incorporating the testlet concept in test score analyses. *Educational Measurement: Issues and Practice, 19*(4), 9-15. http://dx.doi.org/10.1111/j.1745-3992.2000.tb00041.x.

Lee, Y.-W., y Sawaki, Y. (2009). Cognitive diagnosis approaches to language assessment: an overview. *Language Assessment Quarterly, 6*(3), 172-189. http://doi.org/10.1080/15434300902985108.

Liu, J., Xu, G. y Ying, Z. (2012). Data-driven learning of Q-matrix. *Applied Psychological Measurement, 36(*7), 548-564. http://doi.org/10.1177/ 0146621612456591.

Ministerio de Educación del Perú (2009). Evaluación Censal de Estudiantes (ECE) Segundo grado de primaria y cuarto grado de primaria de IE EIB. Marco de trabajo. Recuperado de http://www2.minedu.gob.pe/umc/ece/Marco_de_Trabajo_ECE.pdf.

Ministerio de Educación del Perú (2014). Reporte técnico de la Evaluación Censal de Estudiantes (ECE 2014). Segundo y Cuarto (EIB) de primaria. Recuperado de http://umc.minedu.gob.pe/wp-content/uploads/2015/05/Reporte-T%C3%A9cnico-ECE-20141.pdf.

Paek, I., Yon, H., Wilson, M. y Kang, T. (2009). Random parameter structure and the testlet model: extension of the Rasch Testlet Model. *Journal of Applied Measurement, 10*(4), 394-407.

Ravand, H. y Robitzsch, A. (2015). Cognitive Diagnostic Modeling Using R. *Practical Assessment, Research & Evaluation, 20*(11), 1-12.

Ravand, H. (2015). Application of a cognitive diagnostic model to a high-Ravand, H. (2015). Application of a cognitive diagnostic model to a high-stakes reading comprehension test. *Journal of Psychoeducational Assessment*, 1-18. http://doi.org/10.1177/0734282915623053

Reckease, M. D. (2009). *Multidimensional Item Response Theory.* Nueva York: Springer.

Sireci, S. G., Thissen, D., y Wainer, H. (1991). On the reliability of testlet-based tests. *Journal of Educational Measurement, 28*(3), 237-247. http://dx.doi.org/10.1111/j.1745-3984.1991.tb00356.x.

Tatsuoka, K. K. (1985). A probabilistic model for diagnosing misconceptions by the pattern classification approach. *Journal of Educational and Behavioral Statistics, 10*, 55-73.

Wainer, H., y Lewis, C. (1990). Toward a psychometrics for testlets. *Journal of Educational Measurement, 27*(1), 1–14. http://dx.doi.org/10.1111/j.1745-3984.1990.tb00730.x.

Wainer, H., y Wang, X. (2000). Using a new statistical model for testlets to score TOEFL. *Journal of Educational Measurement, 37*(3), 203-220. http://dx.doi.org/10.1111/j.1745-3984.2000.tb01083.x.

Wang, W.-C., Wilson, M., y Adams, R. J. (1997). Rasch models for multidimensionality between and within items. En G. Engelhard, M. Wilson y K. Draney (Eds.), *Objective measurement: theory into practice* (vol. 4, pp. 139-156). Norwood, NJ: Ablex.

Wang, W.-C., y Wilson, M. (2005). The Rasch testlet model. *Applied Psychological Measurement, 29*(2), 126–149. http://dx.doi.org/10.1177/0146621604271053.

Wilson, M., y Adams, R. J. (1995). Rasch models for item bundles. *Psychometrika, 60*(2), 181–198. http://dx.doi.org/10.1007/BF02301412.