PEL Pensamiento Educativo. Revista de Investigación Educacional Latinoamericana
2017, 54(2), 1-19

Pontificia Universidad Católica de Chile
Facultad de Educación

Calibración concurrente unidimensional y multidimensional dentro de la teoría de respuesta del Ítem

Concurrent Unidimensional and Multidimensional Calibration within Item Response Theory

Steffen Brandt
opencampus.sh, Kiel, Germany

Resumen

Actualmente, importantes estudios sobre logros educacionales, particularmente a gran escala, usan la Teoría de Respuesta al Ítem (TRI) como método para sus análisis. Uno de los supuestos más importantes y básicos de esta teoría es en la dimensionalidad de los test: Para ser interpretados unidimensionalmente los test deben ser unidimensionales y por lo mismo no pueden ser multidimensional. Aunque este supuesto pareciese muy básico, usualmente es ignorado. El *Programa para la Evaluación Internacional de Estudiantes* (PISA), por ejemplo, aplica un modelo TRI unidimensional para el análisis de logros matemáticos y a su vez aplica un modelo TRI multidimensional para analizar cuatro sub-escalas de matemáticas. Esta contradicción de uno de los supuestos más básicos del modelo TRI no es exclusivo de PISA. Este trabajo discute los acercamientos recientemente usados y presenta un nuevo planteamiento: el modelo de sub-dimensión generalizada. Este modelo permite el cálculo de calificaciones ponderadas dentro del marco TRI. Las características del modelo son comparadas a otros modelos, particularmente los de orden jerárquico. Más allá de las comparaciones sobre fiabilidad de resultados, la discusión se concentra particularmente en la diferencia de interpretaciones, quiere decir, en su validez.

Palabras clave: modelos jerárquicos, modelo bi-factorial, modelos de nivel superior,

Correspondencia a:
Steffen Brandt
opencampus.sh, Wissenschaftszentrum
Kiel, 24118 Kiel, Germany.
Email: steffen@opencampus.sh

© 2017 PEL, http://www.pensamientoeducativo.org - http://www.pel.cl

ISSN:0719-0409 DDI:203.262, Santiago, Chile
doi: 10.7764/PEL.54.2.2017.4

Abstract

Today, important educational achievement studies, particularly large-scale assessments, use item response theory (IRT) as the method for their analyses. An important and basic assumption of IRT is on the dimensionality of a test: In order to be interpreted unidimensionally a test has to be unidimensional and hence cannot be multidimensional. Though, this basic assumption is very often neglected. The *Programme for International Student Assessment* (PISA), for example, applies a unidimensional IRT-Model for the analysis of the mathematics achievement and at the same time applies a multidimensional IRT-model for the analysis of the four subscales of mathematics. This contradiction to one of the basic assumptions of IRT is not unique to PISA. This work, at first, discusses the currently used approaches, and presents a new approach: the generalized subdimension model (GSM). It allows the calculation of a weighted mean score within the IRT framework. The model's characteristics are compared to those of other models, particularly hierarchical models. Beyond the comparison of model fit, that is, the reliability of the results, the discussion particularly focuses on the difference in their interpretation, that is, on their validity.

Keywords: hierarchical models, bi-factor model, higher-order models, generalized sub-dimension model, local item dependence

Históricamente, las pruebas y cuestionarios de rendimiento[1] han sido analizados aplicando métodos basados en la teoría clásica de las pruebas (TCP) utilizando, por ejemplo, la suma o la media de los puntajes para interpretar los resultados, y gran parte de los análisis que se realizan en la actualidad todavía se basan en estos métodos. Sin embargo, la TCP presenta desventajas importantes: (a) la TCP no incluye una teoría sobre las dificultades del ítem y, por lo tanto, limita la investigación de las características de la prueba, así como la comparación o la relación de resultados de pruebas con diferentes conjuntos de ítems; y (b) la TCP incluye supuestos muy sólidos respecto de las características de la prueba que es analizada (ver, por ejemplo, Moosbrugger y Kelava, 2007; Rost, 1996).

Como una forma de evitar estas desventajas se desarrolló la teoría de respuesta al ítem (TRI) (ver, p. ej., Moosbrugger y Kelava, 2007; Rost, 1996) y, en la actualidad, todos los estudios nacionales e internacionales importantes se basan en análisis TRI. Sin embargo, la TRI incluye numerosos supuestos, que a menudo son difíciles de cumplir. Existe un supuesto importante y muy básico respecto de la determinada dimensión de una prueba. En la TRI (así como en la teoría clásica de las pruebas [TCP]), es un hecho que para poder interpretar una prueba de manera unidimensional, esta tiene que ser unidimensional. Aunque esta pareciera ser una indicación redundante, en realidad, en muchos casos la misma prueba se interpreta tanto de manera unidimensional como multidimensional. En el Programa para la Evaluación Internacional de Alumnos (PISA), por ejemplo, se informa un puntaje unidimensional en matemáticas y al mismo tiempo asigna puntajes en las cuatro sub-dimensiones: *cambio y relaciones*, *cantidad*, *espacio y forma* e *incertidumbre y datos*, dando por hecho que las matemáticas son un constructo multidimensional. Lo mismo aplica para la lectura y para los constructos de ciencia que se investigan en PISA (OCDE, 2012b) y para constructos similares investigados por otros estudios internacionales, como el Estudio Internacional de Tendencias en Matemáticas y Ciencias (TIMSS) y el Estudio sobre el Progreso Internacional de la Competencia en Lectura (PIRLS) (Martin y Mullis, 2012). Al parecer, esto es

1 En adelante, los "cuestionarios" y las "pruebas de rendimiento académico" simplemente serán llamadas pruebas.

una contradicción para un supuesto básico de la TRI. Sin embargo, la necesidad práctica de generar resultados unidimensionales y multidimensionales se impone ante las necesidades de la teoría. Es notable, sin embargo, que ni en PISA, TIMSS, ni PIRLS se analice esta contradicción teórica y que se pasen por alto sus posibles efectos.

La Evaluación Nacional del Progreso Educativo de Estados Unidos (NAEP) adopta un enfoque distinto. En esta evaluación, la capacidad de lectura, por ejemplo, está compuesta de Lectura como experiencia literaria, Lectura para obtener información y Lectura para realizar una tarea (Donahue y Schoeps, 2001). Los puntajes de estas tres sub-dimensiones de lectura se calibran mediante un modelo de la TRI multidimensional (tres-). Sin embargo, el puntaje de comprensión de lectura se calcula como puntaje ponderado basado en los resultados estimados de la calibración de las sub-dimensiones (Allen, Carlson y Donoghue, 2001). Tanto el enfoque de usar un puntaje de escala de un modelo TRI unidimensional como el enfoque de usar un puntaje compuesto basado en una calibración multidimensional tienen sus ventajas y desventajas (cf. Tabla 1), las que se analizan en mayor detalle en la siguiente sección. A raíz de este análisis, se presentó un enfoque nuevo: el Modelo de sub-dimensión generalizada (GSM). El GSM es una combinación de los dos enfoques que se utilizan actualmente, una restricción de un modelo TRI multidimensional que genera un puntaje medio ponderado para la dimensión de comprensión como un parámetro adicional del modelo.

Tabla1

Ventajas y desventajas de obtener un puntaje unidimensional para datos multidimensionales a través de un puntaje compuesto versus una calibración unidimensional

Calibración unidimensional	
Ventajas	Desventajas
• Es posible calcular estimaciones de máxima probabilidad (WLE, MLE, etc.)	• Sobreestimación de confiabilidad debido a dependencia local del ítem errónea (DLI) • Dudosa validez del constructo multidimensional si la prueba fue elaborada para ser unidimensional • Ponderación implícita y confusa del puntaje unidimensional
Resultado compuesto basado en calibración multidimensional	
Ventajas	Desventajas
• Incluye a la DLI debido a la estructura multidimensional y por lo tanto, las estimaciones de confiabilidad son más apropiadas • Ponderación explícita y clara del puntaje unidimensional	• *No* es posible calcular estimaciones de máxima probabilidad. • No es apropiado para el modelo de Rasch (la normalización de los puntajes multidimensionales provoca un mayor error de medición)

2

2 Las estimaciones de probabilidad se basan en un conjunto de datos de respuesta determinado para un conjunto de ítems administrados y ofrecen una estimación de la probabilidad de los datos de respuesta bajo el supuesto de un modelo determinado y un conjunto de parámetros estimados del modelo. La aplicación de una calibración multidimensional reduce la información que proporciona el conjunto total de datos de respuesta a puntajes de las dimensiones consideradas. Debido a que un modelo para la construcción de la composición de estos puntajes solo se basa en estos resultados, pero no en el conjunto de datos de respuesta propiamente dicho, no permite realizar más estimaciones respecto de la probabilidad del conjunto de datos de respuesta.

Ventajas y desventajas de los enfoques utilizados actualmente en evaluaciones de gran escala

NAEP, TIMSS, PIRLS y PISA son evaluaciones a gran escala que se han aplicado de manera habitual durante las últimas dos décadas. Debido al fuerte interés político en los resultados y la considerable atención que reciben estos estudios, no solo de parte del público sino también de la comunidad científica, es una obligación natural realizar los estudios siempre en el estado del arte actual en medición. Debido a los recursos financieros recibidos, en ocasiones, los estudios incluso pueden aportar importantes avances a la investigación de medición. NAEP, por ejemplo, fue responsable de introducir el enfoque de valor plausible en la medición, y en la actualidad es una técnica estándar para calcular estimaciones objetivas de nivel de grupo (Beaton, 1987; Von Davier, Gonzalez y Mislevy, 2009).

Por lo tanto, el análisis de los enfoques que utilizan estas evaluaciones a gran escala es considerado una buena manera de ofrecer una descripción general del estado del arte actual del cálculo de puntajes unidimensionales para datos multidimensionales.

Dependencia local del ítem

Un supuesto subyacente básico de los modelos TRI es la interdependencia local del ítem (ILI). La ILI describe el hecho de que las respuestas observadas para una prueba se dan por sentado como condicionalmente independientes debido a los puntajes de los individuos en la variable latente que se está midiendo. La violación de este supuesto se denota como dependencia local del ítem (DLI). La DLI puede ocurrir por diversos motivos, el que se considera más común probablemente es la DLI debido a testlets, o paquetes de ítems. Los testlets se refieren a ítems que comparten un estímulo en común. Son populares debido a que permiten un uso más económico del tiempo de prueba; para responder varios ítems con un solo estímulo, las personas necesitan menos tiempo de respuesta por ítem si se compara con tener que leer un estímulo nuevo para cada ítem. Sin embargo, también existen desventajas. Si una persona contesta correctamente un ítem de un estímulo determinado, la probabilidad de que responda correctamente un ítem del mismo estímulo frecuentemente es levemente mayor que la probabilidad de que responda correctamente un ítem de un estímulo diferente, que ya ha sido contestado de manera incorrecta. Es decir, los ítems indican DLI. Lo mismo se aplica a las sub-dimensiones; si se asume que una dimensión determinada está compuesta por sub-dimensiones, se asume que los ítems pertenecientes a una sub-dimensión común se relacionan más fuertemente entre sí que con ítems de otras sub-dimensiones.

Los efectos de la DLI en análisis TRI han sido investigados por numerosos autores. De manera unánime, se establece que ignorar la DLI genera una estimación sesgada de los parámetros de dificultad, una sobreestimación del ítem de discriminación, un sesgo en la estimación de varianza y una sobreestimación de confiabilidad (ver, p. ej., Monseur, Baye, Lafontaine y Quittre, 2011; Tuerlinckx y De Boeck, 2001; Wainer, Bradlow y Wang, 2007; Wang y Wilson, 2005; Yen, 1984).

Las evaluaciones consideradas de gran escala toman en cuenta el supuesto de ILI de manera distinta. En NAEP, TIMSS y PIRLS las pruebas de matemáticas, por ejemplo, simplemente no usan paquetes de ítems, pero cada ítem tiene un estímulo individual. Solo en PISA se utilizan paquetes

de ítems para la prueba de matemáticas. Por otra parte, para la prueba de capacidad de lectura, todos los estudios utilizan paquetes de ítems. Por lo tanto, en NAEP la potencial DLI se investiga usando índices DLI disponibles y cuando varios ítems necesarios se agrupan en un solo ítem para evitar la dependencia local del ítem[3] (Allen y Carlson, 1987, pp. 236–237). En los informes técnicos de TIMSS, PIRLS y PISA no se mencionan ni analizan las posibles dependencias locales del ítem. Sin embargo, se sabe que los paquetes de ítems usados en PISA, por ejemplo, genera DLI en los respectivos ítems (Brandt, 2006; Monseur et al., 2011).

Cuando considera la DLI debido a sub-dimensiones, NAEP también adopta un enfoque diferente. En NAEP se usa modelo TRI multidimensional para calibrar los valores plausibles para cada persona y cada sub-dimensión, y se calculan los puntajes completos de las sub-dimensiones como medias ponderadas de los valores plausibles (Allen et al., 2001, p. 155). De esta manera, se evitan posibles efectos negativos en la calibración TRI debido a DLI causada por las sub-dimensiones. En TIMSS, PIRLS y PISA las sub-dimensiones se calibran conjuntamente mediante un modelo TRI unidimensional; no se consideran posibles efectos causados por DLI.

Ponderación de las sub-dimensiones

En NAEP, las sub-dimensiones se calibran como dimensiones independientes y expertos en la materia proporcionan una ponderación para cada sub-dimensión de la dimensión general. De esta manera, el puntaje de Lectura en el último año de secundaria, por ejemplo, está compuesto por *Lectura como experiencia literaria* con una ponderación de 35%, *Lectura para obtener información* con una ponderación de 45% y *Lectura para realizar una tarea* con una ponderación de 20% (Donahue y Schoeps, 2001).

Tabla 2
Distribuciones de puntajes en matemáticas por sub-dimensión para Estudio en papel PISA 2003 y Estudio digital y en papel PISA 2012

Sub-dimensión	PISA 2003 En papel	PISA 2012 En papel	PISA 2012 En computadora
Cambio y relaciones	30,4% (28 puntos)	26,1% (24 puntos)	29,2 (14 puntos)
Cantidad	23,9% (22 puntos)	23,9% (22 puntos)	20,8% (10 puntos)
Espacio y forma	22,8% (21 puntos)	25% (23 puntos)	31,3% (15 puntos)
Incertidumbre y datos	22,8% (21 puntos)	25% (23 puntos)	18,8% (9 puntos)
Total	100% (92 puntos)	100% (92 puntos)	100% (48 puntos)

Nota. Los puntajes se calcularon basándose en las clasificaciones de ítem que se indican en el Anexo 12 del Informe técnico de PISA 2003 y en el Anexo A del Informe técnico PISA 2012(OCDE, 2005, 2012b).

En TIMSS, PIRLS y PISA la ponderación de las sub-dimensiones son menos claras y de hecho, varían según la prueba. Para demostrar esto, se representan gráficamente las ponderaciones de la

3 Agrupar ítems en uno solo es una posible estrategia para evitar la DLI; sin embargo, el inconveniente es una pérdida de información, ya que en el modelo TRI ahora solo se considera el puntaje de la suma de los ítems y no los puntajes individuales de los ítems respectivos (Yen, 1993).

prueba de matemáticas de PISA en la Tabla 2. El modelo TRI usado en PISA para calibrar las escalas de rendimiento académico se basa en el modelo de Rasch (Rasch, 1980). En el modelo de Rasch, la ponderación de un ítem corresponde al puntaje máximo obtenible para dicho ítem. Por lo tanto, dividir el puntaje máximo para cada sub-dimensión por el puntaje total obtenible permite obtener las ponderaciones de cada sub-dimensión. En PISA 2003 y PISA 2012, el marco de evaluación para las pruebas de matemáticas (en estos dos ciclos de PISA, las matemáticas fueron el enfoque principal e incluyó la estimación de las sub-dimensiones) especificó una ponderación equivalente de 25% para cada una de las sub-dimensiones: *cambio y relaciones*, *cantidad*, *espacio y forma* e *incertidumbre y datos* (OCDE, 2012a, 2004). Las ponderaciones reales, sin embargo, variaron entre 22,8% y 30,4% en PISA 2003; entre 23,9% y 26,1% para la prueba en papel de PISA 2012 y entre 18,8% y 31,3% para la evaluación digital de PISA 2012. Una razón importante de la variación en las ponderaciones es el hecho que es muy difícil predecir el puntaje total final de una sub-dimensión en el momento en que se realiza la prueba. Los puntajes finales se fijan solo después de conocer los datos de respuesta y las características del ítem resultantes. En el caso de PISA 2012, por ejemplo, se eliminó un ítem de matemáticas incluido en el ensayo principal debido a problemas con respecto a la coherencia con la que se aplicó la regla de codificación planificada en los países (a pesar de que todos los ítems pasaron por un exhaustivo ensayo de campo); mientras que para seis países se eliminó un ítem a nivel nacional ya que el ítem (diferente en cada caso) mostró una dificultad que no era coherente con la dificultad observada en el resto de países (OCDE, 2012b, pp. 231–232). Por consiguiente, las sub-dimensiones de los ítems eliminados tendrán una menor ponderación en el puntaje total. De hecho, si se consideran las ponderaciones de las sub-dimensiones de los seis países en que se eliminó un ítem a nivel nacional, estas serán levemente diferentes a la ponderación de los otros países. Otra razón de que cambie el puntaje máximo final puede ser que un ítem que fue administrado para distinguir personas de tres categorías de puntaje (0 puntos, 1 punto y 2 puntos) no muestra suficiente variabilidad en las respuestas, por lo que las categorías de puntajes se reducen a dos (0 puntos y 1 punto).

Si se consideran las ponderaciones finales en TIMSS y PIRLS, es aún más complicado definir las ponderaciones reales, ya que estos dos estudios utilizan un modelo TRI logístico de 2 parámetros (Birnbaum, 1968). Mientras que el modelo de Rasch solo estima un parámetro de dificultad para cada ítem, el modelo logístico de 2 parámetros estima de forma adicional un parámetro de discriminación para cada ítem, lo que permite modelar datos más precisos y aumentar la confiabilidad. El inconveniente, sin embargo, es que los ítems obtienen distintas ponderaciones para la calibración del puntaje final, es decir, los ítems con más discriminación obtienen una ponderación mayor y los ítems con menos discriminación obtienen una ponderación menor. En consecuencia, la ponderación de una sub-dimensión también cambia si la discriminación promedio de sus ítems está por encima o debajo del promedio de la prueba total.

Confiabilidad

En NAEP, el análisis TRI siempre considera las sub-dimensiones de manera independiente. Además, las estadísticas del ítem calculadas para el ensayo de campo, por ejemplo, se basan en los resultados de cada una de las respectivas sub-dimensiones. Por lo tanto, el proceso psicométrico de selección del ítem pretende maximizar la confiabilidad respecto de la medición de las sub-

dimensiones. Una posible desventaja de la construcción de esta prueba multidimensional puede ser que se reduce la confiabilidad del puntaje total general, puesto que los ítems no están diseñados para estar en una escala común. Esta reducción tiene dos razones: en primer lugar, es más fácil desarrollar una escala con alta confiabilidad si se pueden usar más ítems (si las capacidades existentes son muy diferentes, hasta cierto punto también es necesario abarcar todo el rango) y, segundo, mientras menos correlación exista entre las escalas independientes, menos podrán medir un constructo común y menor será la confiabilidad de la escala general que se está midiendo. Es decir, mientras más independientes sean las escalas que realmente pueden identificar y medir constructos diferentes, peor será para la escala general. Probablemente, estas son las razones por las que en PISA se prefiere el enfoque unidimensional para la construcción de pruebas. Sin embargo, probablemente la desventaja más seria es que el enfoque correspondiente no permite calcular de manera confiable los puntajes individuales, lo cual es necesario en cualquier examen de admisión o pruebas similares de gran importancia. Los valores plausibles permiten calcular los resultados confiables a nivel de grupo para el puntaje total, pero no es posible calcular estimaciones de puntos individuales, como WLE, MLE o EAP (para conocer más detalles sobre estas estimaciones, ver p. ej., Rost, 1996). Por lo tanto, el enfoque solo es apropiado si no es necesario realizar el cálculo de las estimaciones individuales. Además, el enfoque no es apropiado si la prueba se analizó mediante el modelo de Rasch. A diferencia del modelo logístico de 2 parámetros, la calibración del modelo de Rasch no permite limitar las sub-dimensiones a varianzas equivalentes sin imponer restricciones adicionales a los datos. Por tanto, en el caso del modelo de Rasch, se deben estandarizar los valores plausibles una vez realizada la calibración del modelo, utilizando las varianzas estimadas de las sub-dimensiones. Sin embargo, al hacerlo, el error estándar de la estimación de la varianza de una sub-dimensión se agregará a cada valor plausible de dicha sub-dimensión y los valores plausibles pierden su útil característica de ser objetivos debido a la estimación.

En PISA, TIMSS y PIRLS, el enfoque está claramente puesto en la construcción de una prueba unidimensional. El proceso de selección del ítem se basa en el modelo unidimensional de Rasch[4] y su objetivo es maximizar la confiabilidad de los puntajes totales generales. La ventaja de este enfoque es que ofrece la opción de calcular puntajes individuales confiables. Además, ajustar los ítems a la escala completa podría generar una mayor confiabilidad para esta escala. Sin embargo, si los ítems atribuidos a una sub-dimensión común comparten algo especial, y por lo tanto incluyen DLI debido a las sub-dimensiones, la confiabilidad estimada será parcial y mayor de lo que es en realidad.

Validez

Además de los requisitos técnicos de medición indicados en las secciones anteriores, también es importante considerar hasta qué punto las medidas de construcción son válidas, es decir, que permiten el uso previsto (American Educational Research Association, American Psychological Association y National Council on Measurement in Education, 2014). En matemáticas, por ejemplo, el uso previsto de las sub-dimensiones es diferenciar el rendimiento de las personas en áreas específicas de las matemáticas con el fin de identificar debilidades y fortalezas en el área matemática. Sin embargo, esa distinción solo es útil si las sub-dimensiones son realmente diferentes. En NAEP se

4 TIMSS y PIRLS usan el modelo logístico de 2 parámetros para calibrar los resultados finales, sin embargo, el análisis de las características del ítem y su proceso de selección se basa en el modelo de Rasch (Martin y Mullis, 2012).

realizaron diversos análisis de dimensión y se concluyó que es razonable interpretar las matemáticas como un constructo multidimensional (Allen et al., 2001, pp. 155–156). Los informes técnicos de PISA, TIMSS y PIRLS no incluyen resultados de los análisis que investigan las estructuras de dimensión supuestas de las escalas de matemáticas, lectura o ciencias. Aquí en cambio, los resultados informados de las sub-dimensiones se pueden tomar como una indicación de las diferencias existentes entre las sub-dimensiones. En la prueba de rendimiento académico de matemáticas PISA 2012, por ejemplo, los Países Bajos alcanzaron un puntaje promedio de 532 puntos para la sub-dimensión*incertidumbre y datos* y un puntaje de 507 puntos para *espacio y forma* (OCDE, 2014). Además de su importancia estadística, que se puede dar por sentado si se consideran el tamaño de la muestra en PISA, la diferencia corresponde a un tamaño de efecto de 0.25 usando la d[5] de Cohen y por lo tanto, también se puede considerar como significativa (Cohen, 1988). Es decir, en este caso, para la prueba de matemáticas PISA, los resultados indican que las sub-dimensiones en realidad miden aptitudes diferentes y, por lo tanto, son multidimensionales. Sin embargo, debemos tener en cuenta que la prueba fue construida para ser unidimensional; de hecho, en todos los estudios piloto, ensayos de campo y también en el estudio principal, los ítems fueron elaborados y seleccionados para ajustarse a una escala unidimensional común. Por lo tanto, la pregunta importante para la validez de los resultados multidimensionales es: ¿se seleccionan los ítems para medir las sub-dimensiones representativas de las aptitudes definidas realmente? Bien podría ser que los resultados de las sub-dimensiones sean sesgados debido a la construcción de la prueba y que, por lo tanto, se reduzca la validez de los resultados de la sub-dimensión.

También podría verse afectada la validez de los resultados unidimensionales por otra razón si las sub-dimensiones realmente son distintas (y solo entonces tiene sentido informarlas por separado). Todas las evaluaciones a gran escala mencionadas utilizan un diseño de cuadernillo rotado, en la que no todos los ítems se administran a todas las personas, pero cada persona solo responde una muestra de ítems. En PISA 2012, por ejemplo, el estudio principal incluye 13 cuadernillos diferentes, cada uno compuesto de cuatro, llamados, grupos de ítems, que incluyen ítems para 30 minutos del tiempo de la prueba (es decir, cada cuadernillo incluye ítems para 2 horas del tiempo de la prueba). Para matemáticas, la prueba incluye siete grupos diferentes distribuidos en los 13 cuadernillos, algunos incluyen solo uno de estos grupos y otros incluyen hasta tres (OCDE, 2012b, p. 31). Los cuadernillos 2, 8, 12 y 13, incluyen solo un grupo. La Tabla 3 muestra la ponderación de las sub-dimensiones dentro de estos grupos. Las ponderaciones varían desde 7,7% hasta 40%, lo que deja claro que una persona, por ejemplo, con una fortaleza relativa en *cambio y relaciones* y una debilidad relativa en *incertidumbre y datos* obtendrá un puntaje diferente dependiendo de si completa el cuadernillo 8 o el 13. Por lo tanto, si la prueba de matemáticas incluye multidimensionalidad debido a las sub-dimensiones, no producirá resultados válidos para el rendimiento general en matemáticas en el nivel individual. Si consideramos los resultados del nivel de grupo, aun así podrían considerarse válidos, ya que se equipararán las diferencias del cuadernillo si los grupos son suficientemente grandes. Sin embargo, debido a que la calibración unidimensional supone que los puntajes individuales en matemáticas son independientes de las ponderaciones de los cuadernillos de acuerdo con las sub-dimensiones, es plausible suponer que el diseño de cuadernillo rotado causa una sobreestimación de la estimación de la confiabilidad unidimensional (adicional a la sobreestimación de la confiabilidad debido a la DLI que se describe en la sección anterior). Sin embargo, un análisis

5 Las escalas de PISA están estandarizadas con una desviación estándar de 100; por lo tanto, la diferencia de 25 puntos corresponde a una *d* de Cohen de 0,25.

más detallado de este aspecto escapa del alcance del presente trabajo.

Tabla 3

Ponderación de las sub-dimensiones de las matemáticas en los cuadernillos 2, 8, 12 y 13 en PISA 2012

Cuadernillo	Cambio y relaciones	Cantidad	Espacio y forma	Incertidumbre y datos
2	30,8%	30,8%	7,7%	30,8%
8	15,4%	23,1%	23,1%	38,5%
12	18,2%	36,4%	27,3%	18,2%
13	40,0%	20,0%	26,7%	13,3%

Nota. Las ponderaciones se calcularon en base al puntaje y las clasificaciones del ítem indicadas en el anexo A del Informe técnico PISA 2012 (OCDE, 2012b).

Fusión de los enfoques utilizados hoy: El modelo de sub-dimensión generalizada

Con el propósito de evitar los inconvenientes descritos anteriormente respecto a los enfoques utilizados actualmente en las evaluaciones a gran escala, se desarrolló el GSM. Combina los dos enfoques limitando un modelo TRI multidimensional para producir un puntaje total ponderado adicional. De esta manera, el modelo permite calcular puntajes individuales confiables y al mismo tiempo, evitar los problemas descritos debido al supuesto equivocado de unidimensionalidad. El modelo se desarrolló en dos pasos: En el primer paso, se desarrolló el modelo de sub-dimensión. Esta versión anterior del GSM también permite calcular un puntaje medio ponderado basado en una restricción del modelo multidimensional, sin embargo, incluye una restricción oculta en las varianzas de las sub-dimensiones. Por lo tanto, el modelo de sub-dimensión solo muestra un ajuste igual al modelo multidimensional si las varianzas de las sub-dimensiones son iguales (cf. Brandt, 2008, 2010, 2012b). En el segundo paso, el modelo de sub-dimensión se generalizó para el GSM mediante la incorporación de un tipo de parámetro adicional que considera las diferencias de varianza de las sub-dimensiones (cf. Brandt, 2012a, 2016; Brandt, Duckor y Wilson, 2014). En las siguientes secciones se entrega una definición, se consideran las características del GSM y se proporciona información sobre cómo clasificarlo en comparación con otros modelos existentes. Luego, el artículo concluye con algunas observaciones finales sobre investigaciones actuales y posibles trabajos futuros.

Definición del modelo

La definición de GSM (en su ampliación de crédito parcial) es la siguiente:

$$\log\left(\frac{p_{nij}}{p_{ni(j-1)}}\right) = d_{k(i)}\left(\theta_n + \gamma_{nk(i)}\right) - b_{ij}, \qquad (1)$$

donde p_{nij} es la probabilidad de que la persona n de una respuesta correspondiente para responder a la categoría j del ítem i; p_{ni0} la correspondiente probabilidad de dar una respuesta que corresponda con la categoría (j-1); b_{ij} es la dificultad del paso j del ítem; θ_n es la capacidad de la persona n en la dimensión unidimensional construida (indicada como dimensión principal); $\gamma_{nk(i)}$ es la capacidad específica de la persona para subpruebas para la (sub-) dimensión k (con el ítem i se refiere a la dimensión k) en relación con la capacidad de la dimensión principal; y $d_{k(i)}$ es el parámetro de traducción que traduce las diferentes escalas multidimensionales (o subdimensionales) en una común. Correspondiente a los modelos jerárquicos, se supone que cada ítem se carga exactamente en una sub-dimensión. Para identificar el modelo, se deben aplicar varias restricciones a los parámetros. En primer lugar, la media

de la capacidad se estima en θ y se tiene que limitar γ_k a cero, mientras que las correlaciones entre la dimensión principal y las sub-dimensiones$K$ se deben establecer en cero. Además, la suma de los parámetros específicos de subpruebas de cada persona se tiene que limitar a cero ($\sum_k \gamma_{rk} = 0$; Limitación I), y el cuadrado de los parámetros d_k se limitan a la suma de K con cada d_k limitados además, a ser positivos ($\sum_k d_k^2 = K$; Limitación II).

Las últimas dos limitaciones se generan a partir de las características de un puntaje medio y se puede demostrar que los resultados de la definición dada en la estimación de la capacidad principal es el promedio (ponderado igualmente) de las capacidades específicas.

El modelo jerárquico, de nivel superior y de testlet

El problema de calcular puntajes de escala unidimensionales para datos que supuestamente son multidimensionales ha recibido atención substancial y ha generado la formulación de una creciente variedad de modelos TRI para hacer frente a este problema. Por lo tanto, en esta sección, primero se caracterizará brevemente los diferentes grupos existentes de modelos TRI antes de analizar más detalladamente las características especiales del GSM y su relación con estos modelos en las secciones siguientes.

Dependiendo de si una DLI supuesta se origina en la construcción de la prueba, o en el constructo psicológico que se va a medir, por lo general, los modelos reciben el nombre de modelos de testlet, modelos jerárquicos o modelos de nivel superior, respectivamente.

Los modelos testlet (Bradlow, Wainer y Wang, 1999; Wang y Wilson, 2005) suponen que las respuestas de una prueba dependen de un solo constructo psicológico. Además, aunque asumen que, debido a que un testlet se basa en la construcción de pruebas, esta incluye una multidimensionalidad que corresponde al conocimiento de cada persona con el estímulo determinado de un testlet. Desde la perspectiva del rasgo latente unidimensional que se va a medir, esta multidimensionalidad corresponde a la dependencia local del ítem (DLI).

Por otro lado, los modelos jerárquicos o de nivel superior (de la Torre y Song, 2009; Gibbons y Hedeker, 1992; Sheng y Wikle, 2008), por lo general suponen que las respuestas de una prueba dependen de múltiples constructos psicológicos relacionados por una estructura subyacente en común. Es decir, desde una perspectiva unidimensional, la prueba incluye dependencias locales del ítem que no se deben a la construcción de la prueba, sino a la naturaleza del constructo psicológico.

Aunque las razones de la necesidad de la multidimensionalidad del modelo (o DLI) son diferentes para los dos grupos de modelos mencionados, los supuestos estadísticos son muy similares. Yung, Thissen y McLeod (1999) y Li, Bolt y Fu (2006) han demostrado que tanto el modelo de nivel superior como el modelo de testlet son restricciones del modelo jerárquico. Además, Rijman (2010) ha demostrado que el modelo de testlet es formalmente equivalente a un modelo de segundo nivel (es decir, un modelo de nivel superior con un segundo nivel como el más alto). Un supuesto general de los modelos jerárquicos (es decir, al igual que el modelo de testlet y de modelos de nivel superior) es que las correlaciones existentes entre los factores de la subprueba (o testlet) se origina completamente

en el rasgo latente común subyacente que miden (Holzinger y Swineford, 1937). Es decir, limitado en el rasgo latente común, a menudo recibe el nombre de factor g y en el GSM llamado dimensión principal, los factores de la sub-dimensión son independientes (ver Rijmen, 2010; Yung et al., 1999). La medida en que este requisito teórico de modelos jerárquicos y las resultantes restricciones aplicadas para la estimación generan un ajuste significativamente peor del modelo en comparación con el modelo multidimensional común, depende del constructo multidimensional que se mide (o de los estímulos utilizados para los testlets). En el capítulo 3, se muestra que para la prueba de rendimiento académico de matemáticas TIMSS 2003, por ejemplo, la diferencia en la desviación del modelo de testlet y el modelo multidimensional es alrededor del 50% de la diferencia de desviación entre el modelo multidimensional y el unidimensional. Es decir, el modelo de testlet solo puede modelar parcialmente la multidimensionalidad en el conjunto de datos determinado. La capacidad de los modelos de testlet, o modelos jerárquicos en términos más generales, para modelar la multidimensionalidad será diferente para cada conjunto de datos, dependiendo de la estructura de covarianza dada de las sub-dimensiones. Sin embargo, debido a los supuestos restrictivos es muy poco probable que sean capaces de modelar completamente la multidimensionalidad, por lo que el ajuste del modelo resultante será significativamente peor que en el análisis que utiliza los datos TIMMS.

Parámetros estimados

Una característica básica de todos los modelos TRI es el número de parámetros que utilizan. Mientras más parámetros tenga un modelo, de mejor manera podrá modelar un conjunto de datos dados. Sin embargo, un mayor número de parámetros, por lo general significa una interpretación más complicada, mientras que los modelos con menos parámetros, a menudo entregan interpretaciones más claras. Por otro lado, normalmente menos parámetros corresponden a supuestos más sólidos, es decir, más restricciones para la calibración del modelo. Por lo tanto, comparar el número y tipo de los parámetros estimados es una manera útil de analizar las características de modelos diferentes.

El GSM limita la media de las capacidades de la persona a cero, lo que constituye una restricción estándar en los modelos TRI y es necesario fijar las escalas al continuo latente. Además, es una característica de los puntajes medios calculados a partir de las distribuciones (dimensiones en este documento) en las que varianzas iguales generan una covarianza cero entre los puntajes medios y las distribuciones de los valores de diferencia, es decir, los valores de distribución menos los respectivos puntajes medios. Por lo tanto, limitar a cero la covarianza de la dimensión principal y las sub-dimensiones no limita la estimación del puntaje medio. Esta limitación sigue estando de acuerdo con los modelos jerárquicos, las restantes restricciones del GSM son diferentes. Son necesarias para permitir que las correlaciones entre los parámetros específicos de subpruebas, que están limitados para que sean independientes en los modelos jerárquicos. Las diferencias entre estos supuestos del GSM y el modelo jerárquico se describen en la Figura 1. Aquí, $\mathbf{y}_k$ indica el vector de respuesta correspondiente a la subprueba k, θ la variable latente de la dimensión principal, y θ_k la variable latente de la sub-dimensión k. Además, las características de los dos modelos se contraponen al modelo multidimensional que se muestra y se puede demostrar que el número de parámetros estimados (libremente) en el GSM es igual al del modelo multidimensional.

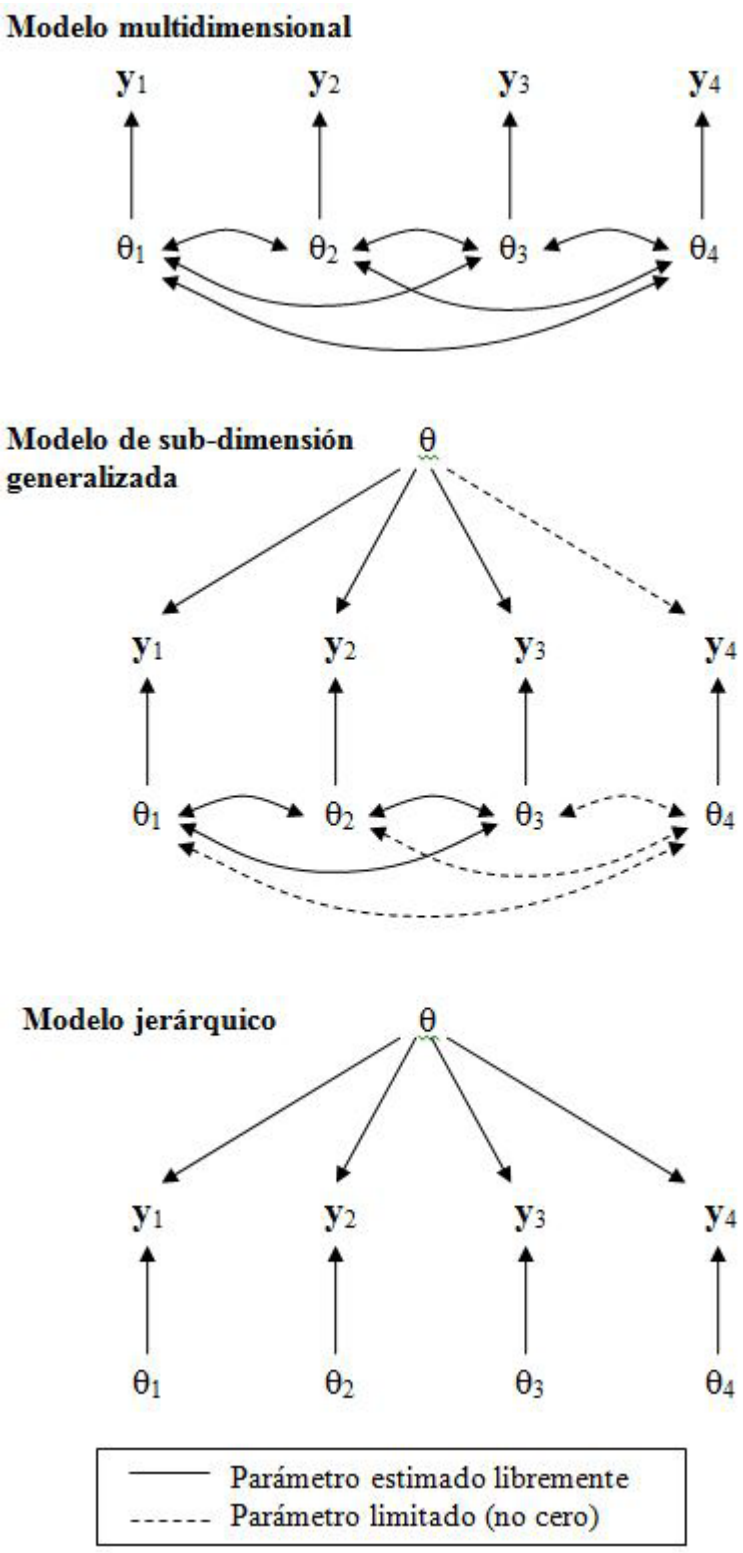

Figura 1. Representaciones gráficas de los parámetros estimados del modelo multidimensional, el modelo de sub-dimensión generalizada y el modelo jerárquico.

El modelo multidimensional (cf. Rost, 1996) se define como

$$\log\left(\frac{p_{ni1}}{p_{ni0}}\right) = \mathbf{a}_i\left(\boldsymbol{\theta}_n - b_i\mathbf{1}\right), \quad (2)$$

donde $\boldsymbol{\theta}_n = \left(\theta_{n1},\ldots,\theta_{nK}\right)^T$ es el vector de capacidad de la persona *n* para sus capacidades en las dimensiones *K*; $\mathbf{a}_i = \left(a_1,\ldots,a_K\right)^T$ es un vector con valores solo de 0 y 1, que indican si un ítem se carga o no en una dimensión; $\mathbf{1}$ es el vector de unidad con los elementos de *K*; y las restantes variables como se definen arriba. Sin pérdida de generalidad, suponemos que las varianzas en el modelo multidimensional están limitadas a una media de cero para identificar el modelo. Es decir, se deben estimar los parámetros de *K* para realizar la estimación de las varianzas de las dimensiones. En el modelo de sub-dimensión generalizada también se tienen que estimar las varianzas de *K*: varianzas específicas de la sub-dimensión*K*-1 (una varianza específica de sub-dimensión, por lo general de la sub-dimensión*K,* no se considera, ya que sus parámetros se obtienen mediante la limitación $\sum_k \gamma_{nk} = 0$) y la varianza para la dimensión principal (cf.ecuación 5.1 del capítulo anterior). Es decir, también se tienen que estimar los parámetros de *K* para obtener las varianzas de las dimensiones.

Si consideramos las estimaciones de la covarianza, todos estos parámetros son estimados libremente en el modelo multidimensional, lo que causa que se tengan que estimar $\sum_{k=1}^{K}(k-1)$ parámetros de covarianza. En el GSM, la matriz de covarianza incluye a la dimensión principal, de las que las covarianzas están restringidas a cero y las sub-dimensiones*K*-1 (una sub-dimensión se genera a partir de parámetros limitados; cf. arriba) con covarianzas estimadas libremente; es decir, aquí, $\sum_{k=1}^{K-1}(k-1)$ se deben estimar parámetros de covarianza, que son *K*-1 parámetros de covarianza menos que en el modelo multidimensional. Sin embargo hay exactamente *K*-1 parámetros adicionales que estimar para los parámetros de traducción d_k (cf. ver definición arriba). Debido a que el número de parámetros estimados para las dificultades de los ítems también es igual, entonces el GSM y el modelo multidimensional incluyen igual número de parámetros.

Equivalencia con el modelo multidimensional

La equivalencia del GSM y el modelo multidimensional se entrega en la definición de $\theta_{nk} = d_k\left(\theta_n + \gamma_{nk}\right)$, donde θ_{nk} es igual a la estimación de la (sub-) dimensión *k* correspondiente en el modelo multidimensional. Para mostrar la equivalencia, se demuestra en lo siguiente que la estimación de las medias, varianzas y covarianzas de las distribuciones de los parámetros θ_{nk} son iguales a las del modelo multidimensional.

Sin pérdida de generalidad, una vez más se supone que ambos modelos se calculan utilizando limitaciones en los casos. Es decir, las distribuciones de los parámetros θ_n y γ_{nk} están limitadas a una media de cero. Por lo tanto, las distribuciones *K* de θ_{nk} simplemente están limitadas a una media de cero y sus medias se corresponden con las del modelo multidimensional.

La independencia de la estimación de varianza para las sub-dimensiones del GSM no es tan sencilla. Si se limitan los parámetros específicos de la sub-dimensión a una suma de cero para cada

persona (Limitación I), las estimaciones de los parámetros de capacidad para las diferentes sub-dimensiones son dependientes entre sí; debido a esto, el modelo estándar de sub-dimensión incluye una limitación de varianza implícita para la estimación de las sub-dimensiones mediante la aplicación de dicha limitación. Para neutralizar esta limitación implícita de varianza, es necesaria la introducción adicional de los parámetros d_k de traducción. Estos indican que cada varianza de las sub-dimensiones de K se estima de manera independiente. Sin embargo, debido a que también se estima la varianza de la dimensión principal (que da como resultado un número total de varianzas estimadas de K+1) los parámetros de varianza necesitan una limitación adicional para ser identificado, lo que se logra limitando el cuadrado de los parámetros d_k a la suma de K (Limitación II).

La correspondencia de las covarianzas en el GSM y el modelo multidimensional se muestra en la ecuación 3. Para cualquiera de las dos distribuciones de θ_{n1} y θ_{n2}, es cierto que

$$\begin{aligned}
\mathrm{Cov}\left(\theta_{n1},\theta_{n2}\right) &= \mathrm{Cov}\left(d_1\left(\theta_n+\gamma_{n1}\right),d_2\left(\theta_n+\gamma_{n2}\right)\right)\\
&= d_1 d_2 \mathrm{Cov}\left(\theta_n+\gamma_{n1},\theta_n+\gamma_{n2}\right)\\
&= d_1 d_2\left(\mathrm{Cov}\left(\theta_n,\theta_n\right)+\mathrm{Cov}\left(\theta_n,\gamma_{n2}\right)+\mathrm{Cov}\left(\gamma_{n1},\theta_n\right)+\mathrm{Cov}\left(\gamma_{n1},\gamma_{n2}\right)\right)\\
&= d_1 d_2\left(\mathrm{Var}\left(\theta_n\right)+\mathrm{Cov}\left(\gamma_{n1},\gamma_{n2}\right)\right)\\
&\qquad\qquad (\text{since } \mathrm{Cov}\left(\gamma_{n1},\theta_n\right)=\mathrm{Cov}\left(\gamma_{n2},\theta_n\right)=0)
\end{aligned} \tag{3}$$

Es decir, la estructura de la covarianza existente entre las dimensiones del modelo multidimensional se puede recuperar completamente mediante el modelo de sub-dimensión generalizada, aunque los parámetros subyacentes θ_{nk} se dividan en los parámetros γ_{nk}, θ_n, y d_k, y solo se estime la estructura de covarianza de los parámetros γ_{nk}.

Correspondencia con el modelo unidimensional

Si no existe la supuesta estructura multidimensional, pero en realidad las sub-dimensiones miden el mismo constructo, las varianzas de los parámetros específicos de la sub-dimensiónγ_{nk} se reducen a cero y, por lo tanto, es verdadero que todas las varianzas de la sub-dimensión son iguales (es decir, cero). Con el fin de generar la Limitación II del GSM, luego los parámetros d_k son todos iguales a 1 y la varianza de la dimensión principal es igual a la varianza de las estimaciones de capacidad del modelo unidimensional. Debido a la definición de la Limitación II, generalmente se señala que la varianza de la dimensión principal es la media de los componentes de varianza unidimensional de las sub-dimensiones, es decir, la media del total de las varianzas de la sub-dimensión, menos la media de la varianzas específicas de la sub-dimensión (ver ecuación 4).

$$\mathrm{Var}\ (\theta_n) = \mathrm{Var}\ (\theta_n)\frac{\sum_k d_k^2}{K} \qquad \text{(due to Constraint II)}$$

$$= \frac{\sum_k d_k^2 \mathrm{Var}\ (\theta_n)}{K} + \frac{\sum_k \mathrm{Var}_k\left(d_k \gamma_{nk}\right)}{K} - \frac{\sum_k \mathrm{Var}_k\left(d_k \gamma_{nk}\right)}{K}$$

$$= \frac{\sum_k \mathrm{Var}_k\left(d_k\left(\theta_n + \gamma_{nk}\right)\right)}{K} - \frac{\sum_k \mathrm{Var}_k\left(d_k \gamma_{nk}\right)}{K}$$

$$= \frac{\sum_k \mathrm{Var}_k\left(\theta_{nk}\right)}{K} - \frac{\sum_k \mathrm{Var}_k\left(d_k \gamma_{nk}\right)}{K}$$

$$= M\left(\mathrm{Var}_k\left(\theta_{nk}\right)\right) - M\left(\mathrm{Var}_k\left(d_k \gamma_{nk}\right)\right) \qquad (4)$$

Además, como se muestra en la ecuación 5, la capacidad de cada persona en la dimensión principal es la media de sus capacidades en las sub-dimensiones consideradas en la escala común a la cual se traducen. Como se señaló anteriormente, esta traducción se obtiene mediante los parámetros d_k. Una multiplicación utilizando el recíproco de d_k, por lo tanto, traduce las estimaciones de capacidad de la sub-dimensión en la que sus varianzas son equivalentes. Es decir, mientras θ_{nk} es equivalente a la estimación de la capacidad multidimensional, $\frac{1}{d_k}\theta_{nk}$ se traduce la estimación de capacidad correspondiente a la escala en la que son equivalentes las varianzas.

$$\theta_n = \theta_n + \frac{\sum_k \gamma_{nk}}{K} \qquad \text{(due to Constraint I)}$$

$$= \frac{K \cdot \theta_n}{K} + \frac{\sum_k \gamma_{nk}}{K}$$

$$= \frac{\sum_k \left(\theta_n + \gamma_{nk}\right)}{K}$$

$$= \mathrm{M}\left((\theta_{n1} + \gamma_{n1}), \ldots, (\theta_n + \gamma_{nK})\right)$$

$$= \mathrm{M}\left(\frac{1}{d_1}\theta_{n1}, \ldots, \frac{1}{d_K}\theta_{nK}\right) \qquad (5)$$

Relación para el modelo jerárquico

La restricción de los parámetros de traducción d_k mediante la Limitación II del GSM deja claro que no se puede interpretar como coeficientes de regresión. Esto contrasta con los modelos jerárquicos, donde los parámetros correspondientes son equivalentes a las cargas o coeficientes de regresión de las sub-dimensiones de la dimensión principal(cf. de la Torre y Song, 2009). Sin embargo, la estimación de estos coeficientes en los modelos jerárquicos se basa en el supuesto de que las capacidades específicas de la subprueba son independientes y por lo tanto, corresponden a la correlación de la capacidad de la subprueba y la capacidad general de la prueba solo si sostiene este supuesto de independencia. Por otro lado, el GSM permite una correlación de las capacidades específicas de la sub-dimensión. De acuerdo con la definición de Holzinger y Swineford (1937), el GSM, por lo tanto, corresponde a un modelo jerárquico modificado que denota un modelo jerárquico con factores específicos superpuestos.

Además de las diferencias en la estructura de covarianza supuesta, la diferencia entre el modelo jerárquico y el GSM también es representada por los diferentes niveles en los que se aplican las limitaciones de los modelos. Aunque la principal limitación del modelo jerárquico genera una característica en el nivel de la prueba, o del rasgo latente (la independencia de la distribución de los factores específicos), la principal limitación del GSM genera una característica en el nivel de la persona individual, es decir, que la suma (traducida) de las estimaciones de capacidad específica de cada persona es cero.

Conclusión

Al considerar el modelo multidimensional, anteriormente se ha demostrado que el GSM produce estimaciones de parámetro equivalentes y que las estimaciones de parámetro unidimensionales se construyen como medias de los parámetros multidimensionales traducidos. Por lo tanto, la aplicación del GSM produce importantes ventajas para calcular calificaciones unidimensionales en pruebas multidimensionales. En primer lugar, la estimación unidimensional resultante está libre de cualquier impacto negativo de DLI debido a las sub-dimensiones; en segundo lugar, la estimación se lleva a cabo en el marco común de TRI, lo que permite calcular de manera confiable las estimaciones individuales para los puntajes totales; y en tercer lugar, se define como puntaje medio la interpretación de la estimación de manera clara y transparente, lo que es particularmente importante en las pruebas de gran importancia.

Otra ventaja del GSM se basa en su característica para generar directamente estimaciones estandarizadas y las posteriores distribuciones de los puntajes de diferencia, es decir, las diferencias en los logros de las sub-dimensiones. A menudo, los investigadores no demuestran interés en las capacidades absolutas en las sub-dimensiones, sino en las diferencias entre estas. Esto podría ser con el propósito de investigar si los estudiantes presentan diferentes capacidades de sub-dimensión, distinguir tipos de estudiantes a través de sus perfiles de capacidad, o para considerar las tendencias en estudios longitudinales (donde las sub-dimensiones representan mediciones en diferentes puntos temporales). Brandt, Duckor y Wilson (2014), por ejemplo, presentaron un enfoque basado en los puntajes de diferencia del GSM con el fin de investigar la dimensionalidad de un test determinado.

Perspectivas para investigaciones actuales y futuras

La definición que se entrega se relaciona con el modelo de Rasch (Rasch, 1980). Sin embargo, será de gran interés su extensión a un modelo logístico de 2 parámetros correspondiente (Birnbaum, 1968) es sencilla y la aplicación de una variante logística de 2 parámetros del GSM a los datos NAEP para comparar los resultados de los puntajes unidimensionales del GSM con aquellos basados en puntajes medios de las estimaciones de valor plausible multidimensional. Hasta la fecha, si bien el análisis de datos a gran escala usando el GSM ha sido difícil, se debe a que solo era posible realizar una calibración del modelo usando WinBUGS (Lunn, Thomas, Best y Spiegelhalter, 2000; cf. capítulo 5)[6]. Además de la complejidad de definir correctamente los modelos, las estimaciones en WinBUGS requieren demasiado tiempo, incluso para pequeñas muestras de datos. Sin embargo, desde que se realizó una actualización del paquete TAM del software R, (Kiefer, Robitzsch y Wu, 2015) el GSM se puede calibrar con TAM y se puede estimar de manera tan eficiente como cualquier modelo TRI multidimensional. Además, TAM también permite la calibración del GSM, incluidos modelos de regresión. Por lo tanto, ahora las futuras aplicaciones del GSM a datos de evaluaciones a gran escala no presentan problemas.

Además de una aplicación más amplia del GSM para explorar más en detalle las diferencias estadísticas de las estimaciones unidimensionales ponderadas y no sesgadas del GSM junto con las estimaciones unidimensionales TRI estándar, es de esperar que el GSM también se pueda agregar a la discusión y determinación de la dimensionalidad de conjuntos de datos. La oportunidad de seleccionar un modelo TRI que proporcione un puntaje unidimensional, pero que no asuma unidimensionalidad, podría ayudar a los desarrolladores de pruebas a aceptar más fácilmente la multidimensionalidad dada y permitir la construcción de pruebas más multidimensionales. Además, los puntajes de diferencia confiable generados por el GSM podrían ayudar a guiar la discusión respecto de la dimensionalidad desde una decisión basada en el ajuste del modelo hasta una decisión basada en la utilidad. Por lo general, la dimensionalidad de una prueba se define mediante comparaciones del ajuste del modelo, aunque, a menudo, los resultados son contradictorios y dependen del criterio de ajuste elegido. En un enfoque basado en la utilidad se plantea una pregunta muy clara: ¿existe un número relevante de personas que realmente difieran en las dimensiones dadas para que sea útil una interpretación independiente de las dimensiones? Por lo tanto, basándose en la utilidad, el enfoque también genera una relación directa con la validez de la dimensionalidad asumida (American Educational Research Association et al., 2014; cf. Brandt et al., 2014).

El artículo original fue recibido el 27 de diciembre de 2016
El artículo revisado fue recibido el 22 de octubre de 2017
El artículo fue aceptado el 27 de octubre de 2017

6 A diferencia del modelo de sub-dimensión, no es posible calibrar el GSM usando ConQuest.

Referencias

Adams, R. J., Wilson, M., & Wang, W. C. (1997). The multidimensional random coefficients multinomial logit model. Applied Psychological Measurement, 21, 1–23.

Allen, N. L., & Carlson, J. E. (1987). Scaling Procedures. In A. E. Beaton (Ed.), The NAEP 1983-1984 technical report. Princeton, NJ: Educational Testing Service.

Allen, N. L., Carlson, J. E., & Donoghue, J. R. (2001). Overview of part II: the analysis of 1998 NAEP data. In N. L. Allen, J. R. Donoghue, & T. L. Schoeps (Eds.), The NAEP 1998 Technical Report (pp. 143–160). Washington, D. C.: National Center for Education Statistics.

American Educational Research Association, American Psychological Association, & National Council on Measurement in Education. (2014). Standards for Educational and Psychological Testing. Washington, DC: American Educational Research Association.

Beaton, A. E. (1987). Implementing the new design: The NAEP 1983-84 technical report. Princeton, NJ: Educational Testing Service.

Birnbaum, A. (1968). Some latent trait models and their use in inferring an examinee's ability. In F. M. Lord & M. R. Novick (Eds.), Statistical theories of mental test scores. Reading, MA: Addison-Wesley.

Bradlow, E. T., Wainer, H., & Wang, X. (1999). A Bayesian random effects model for testlets. Psychometrika, 64, 153–168.

Brandt, S. (2006). Exploring bundle dependencies for the embedded attitudinal items in PISA 2006. Presented at the 13th meeting of the International Objective Measurement Workshop (IOMW), Berkeley, CA.

Brandt, S. (2008). Estimation of a Rasch model including subdimensions. In M. von Davier & D. Hastedt (Eds.), IERI Monograph Series. Issues and Methodologies in Large-Scale Assessments (Vol. 1, pp. 51–70). Princeton, NJ: IEA-ETS Research Institute.

Brandt, S. (2010). Estimating tests including subtests. Journal of Applied Measurement, 11, 352–367.

Brandt, S. (2012a). Definition and Classification of a Generalized Subdimension Model. Presented at the 2012 annual conference of the National Council on Measurement in Education (NCME), Vancouver, BC.

Brandt, S. (2012b). Robustness of multidimensional analyses against local item dependence. Psychological Test and Assessment Modeling, 54, 36–53.

Brandt, S. (2016). Unidimensional Interpretation of Multidimensional Tests (Doctoral dissertation). University of Kiel, Kiel. Recuperado de http://macau.uni-kiel.de/servlets/MCRFileNodeServlet/dissertation_derivate_00006439/Brandt_Dissertation_v1.00.pdf

Brandt, S., & Duckor, B. (2013). Increasing unidimensional measurement precision using a multidimensional item response model approach. Psychological Test and Assessment Modeling, 55, 148–161.

Brandt, S., Duckor, B., & Wilson, M. (2014). A utility-based validation study for the dimensionality of the performance assessment for california teachers. Presented at the 2014 annual conference of the American Educational Research Association (AERA), Philadelphia, PA.

Cohen, J. (1988). Statistical power analysis for the behavioral sciences (2nd ed.). Hillsdale, NJ: Erlbaum.

de la Torre, J., & Song, H. (2009). Simultaneous estimation of overall and domain abilities: A higher-order IRT model approach. Applied Psychological Measurement, 33, 620–639.

Donahue, P. L., & Schoeps, T. L. (2001). Assessment frameworks and instruments for the 1998 national and state reading assessments. In N. L. Allen, J. R. Donoghue, & T. L. Schoeps (Eds.), The NAEP 1998 Technical Report (pp. 255–268). Washington, D. C.: National Center for Education Statistics.

Gibbons, R. D., & Hedeker, D. (1992). Full-information item bi-factor analysis. Psychometrika, 57, 423–436.

Holzinger, K. J., & Swineford, F. (1937). The bi-factor method. Psychometrika, 2, 41–54.

Kiefer, T., Robitzsch, A., & Wu, M. (2015). TAM: test analysis modules (Version 1.3) [R]. Recuperado de http://cran.r-project.org/package=TAM

Li, Y., Bolt, D. M., & Fu, J. (2006). A comparison of alternative models for testlets. Applied Psychological Measurement, 30, 3–21.

Lunn, D. J., Thomas, A., Best, N. G., & Spiegelhalter, D. J. (2000). WinBugs - A Bayesian modelling framework: Concepts, structure, and extensibilty. Statistics and Computing, 10, 325–337.

Martin, M. O., & Mullis, I. V. S. (2012). Methods and procedures in TIMSS and PIRLS 2011. Chestnut Hill, MA: TIMSS & PIRLS International Study Center, Boston College.

Monseur, C., Baye, A., Lafontaine, D., & Quittre, V. (2011). PISA test format assessment and the local independence assumption. IERI Monograph series–Issues and Methodologies in Large Scale Assessments, 4, 131–158.

Moosbrugger, H., & Kelava, A. (2007). Testtheorie und Fragebogenkonstruktion [Test theory and questionnaire construction]. Heidelberg: Springer Medizin Verlag.

OECD. (2005). PISA 2003 technical report. Paris: OECD.

OECD. (2012a). PISA 2012 assessment and analytical framework: mathematics, reading, science, problem solving and financial literacy. OECD Publishing.

OECD. (2012b). PISA 2012 technical report. Paris: OECD.

OECD. (2014). PISA 2012 results: what students know and can do (Vol. 1, revised edition). Paris: Organisation for Economic Co-operation and Development.

OECD, O. for E. (2004). The PISA 2003 assessment framework: mathematics, reading, science and problem solving knowledge and skills. OECD Publishing.

Rasch, G. (1980). Probabilistic models for some intelligence and attainment tests. Chicago: University of Chicago Press.

Rijmen, F. (2010). Formal relations and an empirical comparison among the bi-factor, the testlet, and a second-order multidimensional IRT model. Journal of Educational Measurement, 47, 361–372.

Rost, J. (1996). Lehrbuch Testtheorie, Testkonstruktion [Textbook test theory, test construction]. Bern; Göttingen; Toronto; Seattle: Verlag Hans Huber.

Sheng, Y., & Wikle, C. K. (2008). Bayesian multidimensional IRT models with a hierarchical structure. Educational and Psychological Measurement, 68, 413–430.

Tuerlinckx, F., & De Boeck, P. (2001). The effect of ignoring item interactions on the estimated discrimination parameters in item response theory. Psychological Methods, 6, 181–195.

Von Davier, M., Gonzalez, E., & Mislevy, R. (2009). What are plausible values and why are they useful. IERI Monograph Series, 2, 9–36.

Wainer, H., Bradlow, E. T., & Wang, X. (2007). Testlet response theory and its applications. Cambridge: Cambridge University Press.

Wang, W.-C., & Wilson, M. (2005). The Rasch testlet model. Applied Psychological Measurement, 29, 126–149.

Wu, M. L., Adams, R. J., & Wilson, M. R. (1998). ACER ConQuest: generalized item response modeling software. Melbourne, Australia: Australian Council for Educational Research.

Yen, W. M. (1984). Effects of local item dependence on the fit and equating performance of the three-parameter logistic model. Applied Psychological Measurement, 8, 125–145.

Yen, W. M. (1993). Scaling performance assessments - strategies for managing local item dependence. Journal of Educational Measurement, 30, 187–213.

Yung, Y.-F., Thissen, D., & McLeod, L. D. (1999). On the relationship between the higher-order factor model and the hierarchical factor model, 64, 113–128.