PEL Pensamiento Educativo. Revista de Investigación Educacional Latinoamericana
2017, 54(2), 1-7

Pontificia Universidad Católica de Chile
Facultad de Educación

Introducción Sección Especial

Medición Educacional Aplicada

Mark Moulton[1], Hong Jiao[2] y María Verónica Santelices[3]

[1]Educational Data Systems
[2]Departmento de Desarrollo Humano y Metodología Cuantitativa,
Universidad de Maryland, College Park
[3]Pontificia Universidad Católica de Chile

El *International Objective Measurement Workshop* (IOMW) es una conferencia bienal a partir de la cual fueron tomados los cuatro artículos de esta edición especial de PEL. Estos se basaron en las presentaciones realizadas en Washington, DC, en Abril de 2016. IOMW siempre ha fomentado un interés por la filosofía y las posbilidades de lo que se llama "medición objetiva", o "invarianza", especifícamente según lo implementado por el modelo de Rasch. Hoy en día, ese interés es tan intenso como en la década de los 80, cuando comenzó la conferencia. A modo de introducción informal, puede ser útil revisar lo que significa "objetividad", cómo está arraigada en las ciencias físicas, y por qué los autores de estos documentos la consideran como un elemento importante a poseer.

¿Qué significa hacer un análisis cuantitativo de un conjunto de datos?

La respuesta varía bastante en las áreas. En estadísticas, el análisis cuantitativo pretende proporcionar una descripción matemática de un conjunto de datos con énfasis en decidir si las diferencias numéricas observadas son "significativas", lo que quiere decir que no es probable que haya ocurrido por casualidad. Esto implica calcular los medios, las desviaciones estándar, los errores estándar, y las estadísticas relacionadas, que es más o menos el enfoque tomado de la "teoría clásica de los tests".

En la década de los 50, pensando en un conjunto de datos de aptitud lingüística, el matemático danés Georg Rasch se dio cuenta de que una descripción estadística sobre el desempeño de los estudiantes en una prueba en particular *no era* lo que él quería. Supongamos que los estudiantes reciben distintos tipos de pruebas con diferentes ítems. Supongamos que los tipos de pruebas cambian cada año. Supongamos que queremos comprar estudiantes de distintos niveles. Supongamos que faltan datos, y no al azar.

Bajo estas circunstancias, la descripción estadística del rendimiento en una prueba en particular no es suficiente para comparar, en ninguna manera generalizable, a los estudiantes en todas las pruebas. Adicionalmente, Rasch se dio cuenta de que quería hablar con claridad sobre estudiantes *individuales*, y no sobre la población como un conjunto, y no deseaba usar estadísticas que dependieran del rendimiento de otros estudiantes ("calificación ponderada por la campana de Gauss") o de lo ítems que les podrían ser asignados, lo que parece evidentemente injusto. En resumen, Rasch quería una manera de medir la habilidad de los estudiantes que fuera tan simple, reproducible y justa, como medir la estatura de un alumno con un palo de madera o medir cantidades físicas como la fuerza y la masa con un dinamómetro o una balanza.

ISSN:0719-0409 DDI:203.262, Santiago, Chile
doi: 10.7764/PEL.54.2.2017.1

¿Cómo miden los físicos las cosas? Rasch consideró el ejemplo de la fuerza y de la masa como lo definió Newton (Rasch, 1960):

$$f = ma$$

La fuerza ejercida sobre una masa (en una dirección dada) se define para igualar la magnitud de la masa multiplicada por su aceleración. A pesar de que los científicos miden la masa usando una balanza y otros diversos métodos, Rasch se dio cuenta de que la formulación de Newton puede ser interpretada como una definición circular. La fuerza se define en términos de masa; la masa es definida en términos de fuerza. Supongamos, entonces, que los únicos datos que tenemos son las aceleraciones que ocurren cuando varias fuerzas actúan en varias masas. El resultado es una matriz:

Tabla 1
Matriz de aceleraciones como el product de Fuerza y Masa

Fuerza/ Masa	M1	M2	M3
F1	A11	A12	A13
F2	A21	A22	A23
F3	A31	A32	A33

Teniendo en cuenta solo los valores de las aceleraciones y la ecuación de fuerza de Newton, as casi posible calcular valores para las fuerzas y masas individuales. Lo que obtenemos son sus *proporciones*. Por ejemplo, si A11 es la mitad de grande que A21, podremos calcular que F1 es también igual de grande que F2:

A11 = F1/M1
A21 = F2/M1
A11/A21 = (F1/M1) / (F2/M1) = F1/F2 = ½

En otras palabras, debido a que las dos aceleraciones se aplican a la misma masa (M1), la M1 no se contabiliza y somos capaces de obtener la proporción de las dos fuerzas solo de las aceleraciones, desconociendo M1. Entonces, obtenemos la misma proporción de fuerza así la masa sea M1, M2, M3, o cualquier otra masa. En otras palabras, la relación de fuerza es "invariante" entre las masas, una propiedad que Rasch llamó "objetividad específica". Según ese mismo razonamiento, es claro que cada proporción de masa es también invariante entre las fuerzas.

Lo que no tenemos son valores absolutos para F1 y F2 y las diversas masas. Esto se aborda estableciendo una convención — definiendo una "masa de referencia". Por ejemplo, podemos definir M1 como la unidad de masa, relacionándola con algo así como el peso de un volumen de agua. Como ya tenemos una medida para A11, eso significa

F1 = M1 * A11 = 1 * A11 = A11

lo que luego podemos usar para calcular valores para otras masas como M2:

M2 = F1 / A12

Por lo tanto, siempre que relacionemos todas las fuerzas a la misma masa de referencia, se pueden calcular valores para las mediciones de fuerza y, a partir de estos, los valores de las masas restantes, y serán comparables en todos los marcos de referencia basados en la masa de referencia. Una definición de masa de referencia fue de hecho propuesta el 7 de abril de 1795, cuando el gramo fue decretado en Francia como "el peso absoluto de un volumen de agua pura igual al cubo de la centésima parte del metro, y a la temperatura de hielo en fusión" ("Décret…", 1795). El estándar ha sido ajustado un par de veces, pero sigue siendo prácticamente el mismo.

(Aunque la definición de la masa física en términos de fuerza fue defendida por Ernst Mach (1919) y fue teóricamente suficiente para los propósitos de Rasch, se encuentra con dificultades prácticas relacionadas con la imposibilidad de medir la aceleración en un instante y desenredar las fuerzas a nivel atómico (Belkind, 2012)).

¿Qué tiene que ver esto con pruebas en los niños? Rasch se dio cuenta de que así como la física puede establecer un sistema de medidas para la fuerza y la masa usando solo aceleraciones, debería ser posible establecer un sistema de medidas para la habilidad del estudiante y la dificultad de los ítems usando solo el desempeño del estudiante en cada ítem. Para obtener valores individuales, configuramos algo análogo a una masa de referencia anclando el promedio de las dificultades de los ítems en una prueba a cero. Para comparar a los estudiantes en las pruebas, definimos "ítems de referencia" a los que todos los alumnos pueden ser asociados a través de ítems en común en los formularios de prueba.

Rasch también se dio cuenta que medir a los estudiantes de esta manera introduce un sistema integrado de control de calidad. Notamos que la relación de dos fuerzas cualquiera y de las aceleraciones asociadas a esas fuerzas, debería ser la misma para cada masa. Por lo tanto, en la medida en que *no* veamos esta equivalencia (la relación de aceleración para una de las masas, por ejemplo, no coinciden con la de las otras masas), sabemos que la medición objetiva (invariante) no se ha producido.

Los datos para esa masa deben ser examinados y probablemente omitidos del conjunto de datos. Lo que un científico no puede hacer es pretender que esas aceleraciones son válidas y usarlas desinhibidamente en sus cálculos. En otras palabras, el modelo de Rasch impone el requisito de que los datos deben ajustarse a las expectativas generadas al aplicar el modelo de medición; de lo contrario no pueden ser utilizados. Esto es lo opuesto al enfoque usualmente adoptado en estadísticas, en donde para protegerse contra el sesgo de selección, los datos son tratados como sacrosantos, y nuevos modelos de parámetros son introducidos y ajustados, si es necesario, para mejorar el ajuste a los datos.

En retrospectiva, la analogía con la fuerza y la masa es irresistible. Los estudiantes son en verdad como fuerzas tratando de empujar una pregunta de prueba (masa) desde un estado de no resuelto a resuelto.
La diferencia principal es que los datos no consisten en números métricos uniformes y precisos como las aceleraciones, sino que consisten en números no métricos discontinuos e imprecisos como 0 y 1 ("incorrecto" y "correcto"), haciendo necesario estimar y trabajar con probabilidades. Entonces, en lugar de decir:

$$a = \frac{f}{m}$$

decimos

$$\frac{p[success]}{p[failure]} = \frac{ability[person\ n]}{difficulty[item\ i]}$$

En donde p significa "probabilidad". Las dos formulaciones son matemáticamente muy similares.

Debido a que a los humanos les resulta más fácil imaginar cantidades en un intervalo en vez de una escala de proporción, tomamos el logaritmo (natural) de ambos lados de la ecuación para hacerlo aditivo en lugar de multiplicativo:

$$log\left(\frac{p[success]}{p[failure]}\right) = log\left(\frac{ability[person\ n]}{difficulty[item\ i]}\right)$$
$$log\left(\frac{p[success]}{p[failure]}\right) = log(ability[person\ n]) - log(difficulty[item\ i]$$

Para usar una nomenclatura común de Rasch, definimos:

$$\beta_n = log(ability[person\ n])$$
$$\delta_i = log(difficulty[item\ i])$$
$$p_{ni} = probability\ of\ success\ of\ person\ n\ on\ item\ i$$

So,

$$\beta_n - \delta_i \equiv log\frac{p_{ni}}{(1-p_{ni})}$$

La diferencia entre la habilidad de una persona *n* y la dificultad de un ítem *i* se define como el logaritmo de la probabilidad de éxito sobre la tasa de fallo, también denominada "logaritmo de probabilidades" de éxito. Tengan en cuenta que esta es una *definición* — estamos definiendo la diferencia entre la habilidad de una persona y la dificultad de un ítem. Esto equivale a decir: la diferencia entre una persona y un ítem se define como cero cuando la probabilidad de éxito es de 0.50.

Con un poco de álgebra, la fórmula se puede reordenar para calcular la probabilidad de que la persona *n* tenga éxito en el ítem *i*.

$$p_{ni} = \frac{e^{(\beta_n - \delta_i)}}{1 + e^{(\beta_n - \delta_i)}}$$

que es la forma canónica del modelo de Rasch para datos dicotómicos. Esta establece que la probabilidad de éxito de una persona en un ítem es una función de la diferencia entre la habilidad de esa persona y la dificultad del ítem. Para analizar un conjunto de datos de una prueba, ajustamos las personas betas e ítems deltas hasta que las probabilidades que implican por cada celda en la matriz se ajusten mejor a los datos observados, un proceso llamado "estimación de máxima probabilidad". Más allá de tales detalles, el modelo de Rasch es solo una manera probabilística de implementar la ecuación de fuerza de Newton.

Hay beneficios significativos en esta forma de pensar acerca de la medición:

- Las suposiciones estadísticas son mínimas
- Funciona para un conjunto de datos pequeños (por ejemplo, del tamaño de un aula)
- Los datos faltantes no son un problema
- Los estudiantes son comparables en todos los formularios, incluso si uno es más difícil
- Los ítems son comparables en todos los grupos de estudiantes, incluso si un grupo es mucho más avanzado
- Identifica ítems marcados que no pertenecen en la prueba, por el motivo que sea
- Identifica estudiantes que no fueron medidos adecuadamente por la prueba, por el motivo que sea
- Aclara si los ítems contienen una construcción coherente
- Ajustar al modelo implica que las mediciones de los estudiantes no están influenciadas por el desempeño de otros estudiantes o por los ítems que fueron seleccionados para la prueba
- Y eso significa que *la prueba es justa*

Las principales advertencias son:

- Unidimensionalidad. Todos los ítems deberían obtener la misma construcción subyacente que incorpora un dominio del contenido bien especificado. El modelo de Rasch ha sido generalizado para incluir modelos multidimensionales, pero aun así estos deben usar ítems que reflejen un dominio del contenido definido claramente o la combinación de dominios de contenido.

- Independencia del ítem local. Cada ítem debería ser estadísticamente independiente de los demás.

En resumen, al aplicar una simple analogía física al comportamiento humano en una situación controlada (una prueba), es posible, en principio, realizar mediciones científicas precisas del estado interior y rasgos mentales de una persona, y otras cosas adicionales. Estas mediciones tienen el potencial de ser igual de válidas que, y son en efecto matemáticamente indistinguibles de, la medición física de la fuerza y la masa. Como un instrumento de medición, el modelo es aplicable a cualquier conjunto de datos, y en cualquier campo en donde un tipo de "fuerza" sea imaginada en encuentro con algún tipo de "resistencia" para producir un resultado dentro de un dominio bien especificado. De hecho, en la medida en que una medición reproducible es un prerrequisito de la ciencia, no queda claro cómo un campo científico puede progresar sin una gama comparable de técnicas de medición. Es esta intuición la que ha mantenido activa la comunidad de psicometría de Rasch por los últimos 50 años y la que destaca en los cuatro artículos de esta edición especial.

En su artículo, *Desarrollando un cuestionario exploratorio de la sensación de pertenencia relacionado al aprendizaje de idiomas utilizando la teoría de medición educacional de Rasch*, Knisley & Wind muestran cómo la teoría de medición puede aplicarse a algo tan sutil como el sentido de pertenencia en una comunidad. Particularmente, exploran la pregunta sobre el "sentido de pertenencia" y es esta un constructo coherente que puede ser medido. El modelo de Rasch es a menudo usado de esta manera – no solo para confirmar e implementar construcciones ya conocidas (por ejemplo, la habilidad lingüística), sino que para identificar y pulir construcciones nuevas y poco entendidas. En este caso, un procedimiento para medir la "pertenencia" hace posible explorar una variedad de preguntas interesantes, como por ejemplo, si un sentimiento fuerte de pertenencia facilita el logro académico en estudiantes universitarios.

En *Evaluación de calidad de calificaciones en exámenes escritos a través del modelo multifocal de lente y la teoría de medición de Rasch***,** Wang y Engelhard desarrollan una analogía entre las ópticas y las evaluaciones mediadas por el evaluador – la idea de que los evaluadores son como "lentes" que revelan un patrón de rendimiento del estudiante y lo distorsionan de maneras sistemáticas. Basándose en el trabajo sobre "modelos de lentes" de Brunswick en la década de los 50, muestran que "el modelo de Rasch de múltiples facetas" puede ser utilizado en vez del análisis tradicional de regresión para modelar los efectos del evaluador. Los modelos de facetas basados en la idea de que un fenómeno observado puede ser, y casi con toda certeza es, el efecto de *más de dos causas*, explotando la propiedad de que los modelos de Rasch, debido a su simple naturaleza aditiva, pueden usarse para separar y desenredar cada capa de causalidad.

Volviendo a la física, la aceleración observada de una nave espacial puede ser modelada como el resultado de: a) el empuje de los motores de cohete (faceta 1); b) masa inercial del cohete (faceta 2); y c) fuerza gravitacional del planeta (faceta 3). De manera similar, el rendimiento en una evaluación de escritura puede ser modelada como el resultado de: a) la competencia escrita del estudiante (faceta 1); b) la dificultad del dominio de la escritura (faceta 2); y (c) la exigencia del evaluador (faceta 3). Al separar estas fuerzas, Wang y Engelhard pueden comparar el desempeño de los evaluadores expertos y operacionales para predecir la imparcialidad de la prueba cuando entra en funcionamiento. Estas muestran cómo utilizar el modelo de múltiples facetas de Rasch para implementar el concepto de lentes de Brunswick linealiza las estimaciones de la competencia escrita en medidas de intervalo, admite análisis de datos de escala de calificación (le regresión requiere datos de intervalo) y permite una comparación de manzanas-a-manzanas entre dos clases de evaluador.

Se mencionó anteriormente que el modelo básico de Rasch requiere unidimensionalidad pero que extensiones multidimensionales del modelo han sido desarrolladas. Tal vez la más prominente de estas sea el modelo vagamente llamado Multidimensional Random Coefficient Multinomial Logit (MRCML) (Adams, Wilson, & Wang, 1997). *Calibración concurrente unidimensional y multidimensional dentro de la teoría de respuesta del Ítem* de Steffen Brandt, hace una contribución original a este cuerpo de trabajo al abordar un problema simple, aunque vergonzoso, al que se enfrentan los examinadores en todas partes.

A los diseñadores de pruebas se les pide a menudo que diseñen exámenes que reporten un puntaje total para cada estudiante, además de un conjunto de puntajes de subescala de "diagnóstico". La solicitud *suena* inocua, pero presenta un serio dilema. Si diseñamos una prueba para que sea *unidimensional*, el puntaje total en la dimensión principal tiene sentido y encaja con el modelo de Rasch, pero los puntajes de sub-dimensión no tienen sentido ya que son, en efecto, copias menos precisas de la dimensión principal, de cualquier manera que las nombremos. Por otra parte, si diseñamos las pruebas para que sean multidimensionales, obtenemos una visión más completa de las habilidades de los estudiantes y puntajes de sub-dimensiones significativas (además de un mejor manejo de la "dependencia local de ítems", lo que significa que algunos ítems no son completamente independientes entre sí en términos estadísticos), pero no hay una puntuación total claramente definida de la dimensión principal que obedezca al modelo de Rasch, solo una media ponderada conjunta de estadísticas de sub-dimensión. Brandt descubrió una manera de obtener *ambos* tipos de puntuaciones – un puntaje general único y puntajes de sub-dimensión – de una prueba usando lo que él llama el modelo de sub-dimensión generalizada (generalized subdimension model GSM). Además de la obvia utilidad del modelo para lidiar con un dilema problemático de diseño de prueba, existen otras ventajas: a) el puntaje total resultante está libre de dependencia de artefactos de ítems locales; b) la estimación del puntaje total se realiza dentro del marco de estimación de la teoría de respuesta de ítem y no de manera externa, permitiendo el cálculo confiable de puntaje de una persona individual y errores estándar; c) el puntaje total tiene un claro significado intuitivo como una media de las habilidades de sub-dimensión del estudiante.

En *Examinando formas de hacer evaluación: Un acercamiento desde la teoría de respuesta del ítem para educadores de profesores*, Duckor, Draney & Wilson abordan el problema de la medición de la "alfabetización de la evaluación" de los profesores – un aspecto de la profesión docente que suena inofensivo pero que conduce a un laberinto de metas de políticas de estado y federales y teoría de pedagogía y medición. Los autores planean hipótesis sobre al menos tres "progresiones de aprendizaje" docente que son parte de la alfabetización evaluadora: a) comprensión de objetivos de aprendizaje; b) comprensión de herramientas de evaluación; y c) comprensión de interpretación de datos. Dentro de cada progresión de aprendizaje, ellos buscan detectar un continuo de tareas que distinguen niveles de desempeño de la comprensión del profesor como una parte de un espacio de construcción general de la alfabetización de la evaluación en la sala de clases (Classroom Assessment Literacy (CAL)).

El instrumento CAL fue administrado a una muestra de profesores y los resultados analizados para encontrar evidencia de confiabilidad y validez, y para determinar si la estructura conceptual de las progresiones tiene sentido. El artículo es un ejemplo detallado del trabajo involucrado en la conceptualización, construcción, perfeccionamiento y síntesis de construcciones que son compatibles con el modelo de Rasch y, como tal, reúne temas planteados en los tres artículos anteriores.

Mediante dichos estudios, esperamos ampliar el alcance y la practicidad de la visión de la medición científica de Rasch. Para relacionarse más con este tipo de investigación, los invitamos a asistir a la próxima conferencia de IOMW en la ciudad de Nueva York en abril del 2018 (www.iomw.org).

Referencias

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

Belkind, O. (2012). Physical Systems: Conceptual Pathways between Flat Space-time and Matter. Springer (Chapter 5.3) .

Décret relatif aux poids et aux mesures du 18 germinal an 3 (7 avril 1795) [Decree of 18 Germinal, year III (April 7, 1795) regarding weights and measures]. Grandes lois de la République (in French). Digithèque de matériaux juridiques et politiques, Université de Perpignan. Recuperado el 3 de noviembre de 2011.

Mach, E. (1919). "Science of Mechanics" .

Rasch, G. (1960). Probabilistic models for some intelligence and attainment tests (Reprint, with Foreword and Afterword by B. D. Wright, Chicago: University of Chicago Press, 1980). Copenhagen, Dinamarca: Danmarks Paedogogiske Institut.