Universidad de Costa Rica
Facultad de Educación
Instituto de Investigación en Educación

REVISTA ACTUALIDADES INVESTIGATIVAS EN EDUCACIÓN

http://revista.inie.ucr.ac.cr/ ISSN 1409-4703

Uso del modelo de Rasch para la construcción de tablas de especificaciones: Propuesta metodológica aplicada a una prueba de selección universitaria

Use of the Rasch model to build specification tables: Methodological proposal applied to a university selection test

Volumen 17, Número 1
Enero-Abril
pp. 1-16

Este número se publicó el 1° de enero de 2017
DOI: http://dx.doi.org/10.15517/aie.v17i1.27299

Danny Cerdas Núñez
Eiliana Montero Rojas

Revista indizada en **REDALYC**, **SCIELO**

Revista distribuida en las bases de datos:

LATINDEX, **DOAJ**, **REDIB**, **IRESIE**, **CLASE**, **DIALNET**, **SHERPA/ROMEO**, **QUALIS-CAPES**, **MIAR**

Revista registrada en los directorios:

ULRICH'S, **REDIE**, **RINACE**, **OEI**, **MAESTROTECA**, **PREAL**, **CLACSO**

Uso del modelo de Rasch para la construcción de tablas de especificaciones: Propuesta metodológica aplicada a una prueba de selección universitaria

Use of the Rasch model to build specification tables: Methodological proposal applied to a university selection test

Danny Cerdas Núñez[1]
Eiliana Montero Rojas[2]

Resumen: *Mediante un enfoque cuantitativo y un análisis descriptivo de investigación, este artículo presenta una aplicación menos conocida del modelo Rasch, como herramienta técnica para la construcción y validación de tablas de especificaciones para pruebas estandarizadas. El estudio hace una propuesta para generar, empleando este modelo, y una con mayor rugosidad científica, una tabla de especificaciones en términos de procesos y contenidos para el componente de razonamiento en contextos matemáticos de la prueba de selección para ingreso a la Universidad de Costa Rica, la prueba de aptitud académica. Debe recordarse que la aproximación estándar en pruebas educativas para construir tablas de especificaciones ha partido tradicionalmente de lo que llamamos "juicio de expertos", pero sin buenos asideros de evidencia empírica científica. El interés principal, más allá de esta aplicación particular, consiste en detallar la metodología utilizada para mostrar de qué forma se puede emplear este marco de referencia para construir y validar tablas de especificaciones de una forma científicamente más rigurosa, apoyada por la evidencia empírica provista por el modelo de Rasch, conjuntamente con el conocimiento y valoración de jueces expertos.*

Palabras clave: *modelo Rasch, tabla de especificaciones, pruebas estandarizadas, pruebas psicométricas, validez.*

Abstract: *Through quantitative approach and a descriptive analysis research, this paper presents a less known application of the Rasch Model, as a technical tool for the development and validation of specifications tables for standardized tests. The study proposes a framework to generate a specification table with processes and contents for the math context component of the undergraduate admission test used by the University of Costa Rica. It is worthwhile to remember that the standard approach to construct specifications tables in educational tests has been based on what traditionally is called "expert judgments", with no good hold on scientific approaches. The main interest, beyond this particular application, is to detail the methodology, and show a way that could be implemented to use this frame of reference to develop and validate specification tables in a more rigorous and scientific way, supported for the empirical evidence provided by the Rasch Model, along with the knowledge and assessment of expert judges.*

Key words*: Rasch model, specifications table, standardized testing, psychometric tests, validity.*

[1] *Estadístico del Programa Permanente Prueba de Aptitud Académica del Instituto de Investigaciones Psicológicas, Universidad de Costa Rica. Dirección electrónica: danny.cerdas@ucr.ac.cr*

[2] *Docente e Investigadora de la Escuela de Estadística y del Instituto de Investigaciones Psicológicas, Universidad de Costa Rica. Dirección electrónica: eiliana.montero@ucr.ac.cr*

Artículo recibido*: 2 de marzo, 2016*
Enviado a corrección*: 28 de junio, 2016*
Aprobado*: 10 de octubre, 2016*

1. Introducción

Cuando se construyen instrumentos estandarizados cognitivos, para ser usados en Educación o Psicología, se deben considerar, entre otras cosas, dos elementos fundamentales en el proceso de su elaboración para brindarle al instrumento evidencias de validez: el modelo de medición utilizado para juzgar la calidad psicométrica de los reactivos, y la tabla de especificaciones, la cual resume los procesos y contenidos incluidos en el instrumento, así como los parámetros meta en cuanto a dificultad y discriminación.

Bajo esta perspectiva, uno de los métodos de análisis de calidad técnica de pruebas, adscrito al marco de la Teoría de Respuesta al Ítem, es el modelo de Rasch, creado por el matemático danés Georg Rasch en 1960. Su aplicación (Bond y Fox, 2001; Prieto y Delgado, 2003) permite la construcción de pruebas más adecuadas y eficientes. Las bondades ostentadas por él permiten la elaboración de una tabla que contenga, de manera detallada, los contenidos, procesos y parámetros de dificultad de los ítems. El funcionamiento que adopta una tabla de especificaciones orienta a los constructores de ítems en la creación de nuevos reactivos.

Tener especificaciones detalladas de los ítems junto con su dificultad otorga mayor claridad al constructo y aumenta las evidencias de validez, al identificar y eliminar en la construcción de ellos, las fuentes de varianza irrelevante al constructo. Al respecto, Messick (1989) menciona que la mejor protección contra las consecuencias sociales adversas que amenazan la validez de una prueba, en cuanto a su interpretación de puntuaciones y al uso, está en reducir o minimizar, en el proceso de medición, alguna fuente potencial de invalidación de la prueba, especialmente la subrepresentación del constructo y la varianza irrelevante en su medición.

La creación de una tabla de especificaciones resulta innovadora para la prueba de aptitud académica, la cual es el instrumento psicométrico empleado por la Universidad de Costa Rica para la selección de estudiantes que ingresan, con el fin de contar con una medición validada científicamente que permita elegir a aquellos que, en términos generales, tengan mayor probabilidad de éxito académico. Este instrumento se compone de dos partes: ítems de razonamiento en contexto verbal e ítems de razonamiento en contexto matemático. Para realizar la prueba de aptitud académica se pretende que el examinado no tenga que recordar datos específicos, sino que aplique sus conocimientos generales para la solución de los ítems planteados y emplee su capacidad de razonamiento.

El objetivo general de este artículo radica en generar una propuesta metodológica para la construcción de tablas de especificaciones utilizando el modelo de Rasch como herramienta de validación.

2. Marco teórico

El modelo de Rasch es un moderno enfoque psicométrico para la construcción, validación e interpretación de instrumentos de medición relacionados con las ciencias del comportamiento. Su creación atañe a 1960 cuando el matemático danés Georg Rasch propuso un modelo de medida más eficiente y adecuado para la elaboración de pruebas psicométricas (Bond y Fox, 2001; Prieto y Delgado, 2003) y para la construcción de medidas basadas en una relación probabilística entre la dificultad de un ítem y el nivel del examinado en el constructo (Bond y Fox, 2001). Este modelo consiguió resolver muchas de las carencias de la Teoría Clásica de los *test* en cuanto a la elaboración de pruebas psicométricas, pues desde comienzos del siglo XX la construcción y el uso de pruebas se basó principalmente en esta teoría (Prieto y Dias, 2003), pruebas psicométricas concernientes a la medición de una muestra de conducta (Gómez e Hidalgo, 2003).

El modelo de Rasch es un modelo psicométrico probabilístico adscrito al marco general de la teoría de respuesta al ítem, donde solamente se estima el nivel de dificultad del ítem (b), pues se asume que el parámetro de respuesta al azar (c) es cero y que el parámetro de discriminación (a) es constante para todos los ítems (Montero, 2001); debido a la sencillez emanada de su lógica, se convierte en el modelo más popular de la teoría de respuesta al ítem (Muñiz, 1997). Para Prieto y Delgado (2003), el modelo de Rasch permite indagar acerca de las características de los ítems representativos de los distintos niveles en el constructo y genera interpretaciones sustantivas muy útiles en términos de los procesos y contenidos que un examinado puede o no lograr a partir de la puntuación estimada en el constructo medido.

Este modelo se fundamenta en que el nivel del examinado en el constructo y la dificultad del ítem son los que determinan la probabilidad de responderlo correctamente (Bond y Fox, 2001). Para modelar dicha relación, Rasch utilizó la siguiente función logística:

$$\ln\left(\frac{P_{is}}{1 - P_{is}}\right) = \left(\theta_s - \beta_i\right)$$

Donde:

$\mathbf{P_{is}}$ **:** es la probabilidad del examinado "s" de responder correctamente al ítem i en determinado nivel de Θ.

θ_s **:** es la estimación del nivel del examinado "s" en el constructo.

β_i **:** es la dificultad del ítem i.

La ecuación indica que el cociente entre la probabilidad de una respuesta correcta y una respuesta incorrecta es una función de la diferencia ($\theta_s - \beta_i$), entre el nivel del examinado en el constructo y la dificultad del ítem (Prieto y Delgado, 2003).

Para expresar los valores escalares de examinados e ítems, según Embretson y Reise (citado en Prieto y Dias, 2003), pueden emplearse distintas métricas, sin embargo, la más utilizada es la escala logit[3] (Prieto y Dias, 2003). La localización del punto cero en la escala es arbitraria, pero generalmente se sitúa en la dificultad promedio de los ítems (Prieto y Delgado, 2003).

A partir de las funciones matemáticas (Prieto y Delgado, 2003; Bond y Fox, 2001), se desprende que, si un examinado responde un ítem que se encuentra por debajo de su nivel de competencia en el constructo, la probabilidad de responderlo correctamente es mayor a 0,50; mientras que si un examinado responde un ítem que se encuentra por encima de su nivel de competencia en el constructo, la probabilidad de responderlo correctamente es menor a 0,50. En caso de que un examinado responda a un ítem equivalente a su mismo nivel de competencia en el constructo, la probabilidad de responder correctamente ese ítem es de 0,50. Además, entre más alejado se encuentre un ítem de un examinado (en términos de su nivel en el constructo), menor es la probabilidad de responder correctamente ese ítem. Esta es una propiedad única del modelo de Rasch que se denomina *medición conjunta,* y es quizá la ventaja más relevante del modelo, pues permite que los parámetros de los examinados y de los ítems se localicen en el mismo continuo o escala, y se expresen en las mismas unidades, lo cual permite analizar las interacciones entre los examinados y los ítems. Esta es una ventaja que hace al modelo especialmente útil para interpretaciones diagnósticas (Cadavid, Delgado y Prieto, 2007).

Las propiedades del modelo de Rasch solo pueden obtenerse si los datos se ajustan al modelo (Bond y Fox, 2001; Prieto y Delgado, 2003; Gori y Battauz, 2006); de lo contrario, no será adecuado. Se supone que el modelo es ideal, por lo que los datos son los que deben ajustarse a él y no él a los datos, como suele suceder en otros casos. Consecuentemente, si

[3] Escala logit. Es una escala de intervalo en la que las unidades intervalares entre la localización del examinado y el ítem tienen un valor consistente. (Bond y Fox, 2001)

algunos ítems no se ajustan a este, deben ser eliminados, o si algunos examinados no se ajustan al modelo, no deben ser considerados en los análisis (Bond y Fox, 2001).

De acuerdo con sus funciones matemáticas, la probabilidad de una respuesta correcta a un ítem solo depende del nivel del examinado en el constructo y de la dificultad del ítem, por lo que a partir de las estimaciones se obtiene la diferencia entre lo que el modelo estima y lo observado en los datos. Lo anterior permite que la presencia de respuestas anómalas, por ejemplo, que examinados con un nivel bajo en el constructo respondan correctamente a un ítem difícil, sean identificadas por el modelo.

Como se ha mencionado, por medio del modelo de Rasch, es posible determinar aquellos ítems y a los examinados que presentan problemas de ajuste (Bond y Fox, 2001; Gori y Battauz, 2006). El estadístico INFIT es una de las medidas de bondad de ajuste más utilizadas, basado en los residuos (diferencia entre valores observados y estimados), sensible a patrones de respuesta inesperados que se hallan cerca del nivel de medición para los examinados o para los ítems (Bond y Fox, 2001; Gori y Battauz, 2006). Por un lado, su valor esperado para un buen ajuste es 1 (Bond y Fox, 2001; Prieto y Delgado, 2007). Por otro lado, los índices de ajuste se relacionan con el contexto particular en el que se mide y con características específicas de la prueba, por lo que para una prueba de escogencia múltiple, con altas consecuencias, el rango de ajuste oscila entre 0,8 y 1,2 (Wright y Linacre, 1994, citado en Bond y Fox, 2001).

2.1 Tabla de especificaciones

Una tabla de especificaciones para una prueba puntualiza características y demás aspectos propiamente de la prueba. Su utilidad consiste en que permite describir detalladamente el constructo que interesa medir, definir el perfil de la prueba que se quiere desarrollar, facilitar a los constructores la creación de nuevos ítems, construir pruebas con menos error de medición, entre otros. Para Tristán y Vidal (2006), una tabla de especificaciones es uno de los estándares de calidad que debe cumplir una prueba de alta calidad técnica, como lo es la prueba de aptitud académica de la Universidad de Costa Rica. Acerca de la necesidad e importancia de las tablas de especificaciones, Tristán y Vidal (2006) mencionan una serie de aspectos característicos, entre los cuales se encuentran los siguientes:

- Una tabla de especificaciones debe proporcionar la fundamentación necesaria para identificar detalladamente lo que se quiere medir con la prueba, ya sea conocimientos o competencias.
- El establecimiento del peso de los elementos que conforman la tabla de especificaciones hace factible diseñar ítems que atiendan especialmente sus necesidades.
- Las definiciones en una tabla de especificaciones deben ser suficientemente claras, para que especialistas y constructores relacionen los ítems y el dominio o contenido que representan.
- A partir de una tabla de especificaciones, una persona especialista del área debe identificar la correspondencia de ítems y la posición que estos ocupan en las especificaciones dentro del perfil.

No obstante, para la creación de una tabla de especificaciones, no puede declararse una única definición sobre lo que debe contener dicha tabla, pues esta se establece, en gran medida, con las necesidades y objetivos que cada institución requiera, por lo que, generalmente, cada institución plantea sus propias especificaciones.

La utilización de una tabla de especificaciones se convierte en un requisito fundamental que debe tener una prueba estandarizada. Según Esquivel (2001), en una prueba referida a normas, contar con una tabla de especificaciones la cual proporcione evidencias de que la prueba es una muestra representativa de los contenidos de una disciplina, se convierte en información esencial para el establecimiento de la validez de las interpretaciones de los resultados, así como del uso de estos últimos (Messick, 1995). Asimismo, una tabla de especificaciones debe cumplir una función de guía para la construcción de ítems al delimitar contenidos y procesos; además, debe constituirse en la plantilla para generar las distintas fórmulas de examen a partir de los bancos de ítems. Los procesos y contenidos se refieren, respectivamente, a las tareas o actividades que son necesarias de llevar a cabo por parte de un estudiante para dar respuesta a un problema y a los conocimientos que el estudiantado requiere tener de ciertos temas o contenidos temáticos, tales como álgebra y geometría. (INEE, 2005)

3. Metodología

La creación de la tabla de especificaciones para el área matemática de la prueba de aptitud académica de la Universidad de Costa Rica se desarrolló bajo un enfoque cuantitativo y un análisis descriptivo de investigación, y se generó a través del análisis de ítems del contexto matemático de la fórmula 1 de esta prueba del 2007, mediante el criterio de jueces y por medio de la aplicación del modelo psicométrico de Rasch, para determinar los contenidos y procesos representados en los ítems, así como sus niveles de dificultad.

Se trabajó con un grupo de 4 jueces de ítems, quienes, a partir de su conocimiento y experiencia tanto de la población como del tipo de prueba, describieron los contenidos y procesos representados en los ítems y efectuaron una estimación de su dificultad con base en los respectivos procesos y contenidos. Sucesivamente, se procedió a aplicar el modelo psicométrico de Rasch al conjunto de datos (mediante el programa psicométrico Winsteps), con la finalidad de apreciar el escalamiento conjunto de examinados e ítems y de obtener las estimaciones de dificultad, las cuales fueron usadas en el proceso de validación empírica de la tabla de especificaciones generada a partir de las valoraciones de los jueces. Posteriormente, se contrastaron las estimaciones de dificultad del modelo psicométrico de Rasch con el trabajo realizado por los jueces, concretamente la estimación de los niveles de dificultad (no a la descripción de contenidos y procesos involucrados en cada ítem).

El grupo de reactivos utilizados para los juzgamientos y análisis reunió un total de 29 ítems, divididos en dos grupos, en dos folletos iguales con 15 ítems y dos folletos iguales con 14 ítems; los cuales se asignaron aleatoriamente entre los cuatro jueces participantes. Cada juez trabajó de manera independiente con su respectivo folleto de ítems, resolviendo y analizando, según su criterio, los contenidos y procesos representados en cada ítem, de manera que se identifiquen aquellas variables que expliquen su dificultad. Además, cada juez realizó una estimación a priori de la dificultad del ítem, considerando 5 posibles niveles de dificultad, donde el valor 1 correspondía a la menor dificultad y, ascendentemente, el valor 5 correspondía a la mayor dificultad.

Para la estimación de dificultad del ítem, se procedió a contrastar el juzgamiento individual de cada juez, con la clasificación derivada de los análisis con el modelo de Rasch; examinando así la concordancia entre ambos valores y estableciendo el nivel predictivo de los jueces.

Resulta fundamental puntualizar que al referirse a validación empírica de la tabla de especificaciones, esto compete a la validez de la apreciación de dificultad que estiman los

jueces para cada ítem, lo que no necesariamente garantiza la validez de los procesos y contenidos identificados por cada juez. En la medida en la que se encuentre consistencia entre las estimaciones de dificultad del modelo y la de jueces, se va a tener mayor confianza en las apreciaciones sustantivas de cada uno en cuanto a procesos y contenidos representados en el ítem.

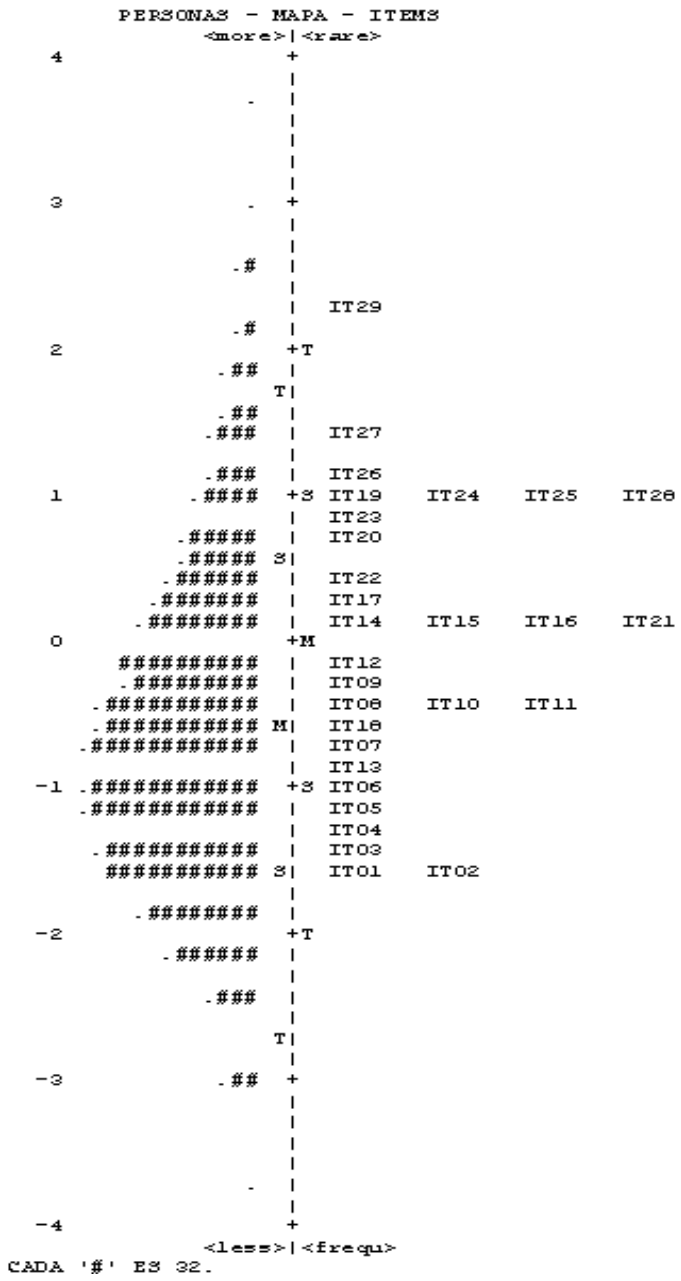
4. Análisis y resultados

Se inició con la verificación de la unidimensionalidad de los datos, la cual dice que un único constructo es suficiente para explicar los resultados de examinados y las relaciones entre ítems (Martínez, 2005). Mediante el programa SPSS 17.0, se efectuó un análisis factorial exploratorio bajo el método de factorización componentes principales y se obtuvo que la varianza total explicada presentaba un gran factor que explicaba casi el 18 % de los datos de la matriz de correlación observada; de igual forma, en el gráfico de sedimentación se apreció la existencia de un factor predominante.

En la primera etapa de análisis, es decir, en el momento inicial de análisis, se aplicó el modelo de Rasch al conjunto total de datos (7125 examinados y 29 ítems) para identificar examinados e ítems que se ajustaran con la expectativa del modelo. En la segunda etapa de análisis, se aplicó el modelo de Rasch a un grupo de 5670 examinados y a los mismos 29 ítems, debido a que fueron los examinados e ítems que en la primera etapa de análisis arrojaron índices aceptables de ajuste. Por lo tanto, en la segunda etapa se generaron resultados provenientes de datos que sí cumplían con la expectativa del modelo utilizado.

En el gráfico 1, se muestra un mapa de la distribución conjunta de examinados e ítems. A la izquierda se distribuyen los examinados según su nivel de habilidad y a la derecha se distribuyen los ítems según su nivel de dificultad.

Gráfico 1
Distribución conjunta de examinados e ítems



Fuente: Elaboración propia a partir de las salidas del paquete Winsteps (2011).

Del mapa se aprecia que el promedio del nivel de habilidad de los examinados es sensiblemente inferior a la dificultad promedio de los ítems, por lo que el conjunto de ítems resultó un poco difícil para la población examinada, lo cual es precisamente la intención de la prueba, discriminar en niveles altos del constructo, porque interesa la elección de los mejores

candidatos para ingresar a la universidad (los que posean mayor nivel). Además, se muestra un grupo de examinados que presenta la habilidad para responder correctamente todos los ítems, pues se encuentran escalonados por encima de la distribución de los 29 ítems. En contraparte, se aprecia un nutrido grupo de examinados con alta probabilidad de no responder correctamente ningún ítem.

En la Tabla 1, se observan algunas características psicométricas de los ítems, donde todos los ítems se ajustan al modelo (siendo 0,85 el valor menor y 1,11 el valor mayor).

Tabla 1
Estadísticas del grupo de ítems

ENTRY NUMBER	TOTAL SCORE	COUNT	MEASURE	MODEL S.E.	INFIT MNSQ	INFIT ZSTD	OUTFIT MNSQ	OUTFIT ZSTD	PTMEA CORR.	EXACT MATCH OBS%	EXACT MATCH EXP%	ITEM
29	517	5670	2.26	.05	.99	-.2	1.40	5.4	.28	91.6	91.1	IT29
27	967	5670	1.42	.04	1.01	.6	1.01	.4	.37	83.6	84.2	IT27
26	1138	5670	1.18	.04	.98	-1.1	1.01	.3	.41	82.4	81.8	IT26
28	1225	5670	1.07	.04	1.08	3.9	1.20	5.5	.33	79.6	80.7	IT28
25	1227	5670	1.07	.04	1.01	.7	1.05	1.6	.39	80.7	80.7	IT25
19	1268	5670	1.02	.04	1.04	2.3	1.17	4.7	.35	79.8	80.1	IT19
24	1322	5670	.95	.03	1.10	5.0	1.29	8.2	.31	78.4	79.4	IT24
23	1405	5670	.85	.03	1.02	1.1	1.01	.4	.40	77.9	78.5	IT23
20	1513	5670	.73	.03	1.08	4.6	1.17	5.7	.35	75.9	77.2	IT20
22	1652	5670	.58	.03	.91	-6.1	.90	-4.5	.50	76.5	73.8	IT22
17	1972	5670	.26	.03	1.06	4.1	1.09	4.1	.38	71.3	72.7	IT17
14	2049	5670	.18	.03	.85	-9.9	.79	-9.9	.56	77.6	72.1	IT14
16	2085	5670	.15	.03	1.01	1.1	1.03	1.7	.42	72.3	71.9	IT16
21	2095	5670	.14	.03	1.00	-.3	1.02	1.2	.44	72.2	71.8	IT21
15	2146	5670	.09	.03	.98	-1.3	.99	-.6	.45	71.7	71.4	IT15
12	2362	5670	-.10	.03	.90	-8.9	.88	-6.7	.52	74.9	70.1	IT12
9	2553	5670	-.28	.03	.94	-5.4	.92	-4.6	.49	72.3	69.3	IT09
11	2680	5670	-.39	.03	1.04	3.5	1.06	3.3	.40	67.1	68.9	IT11
8	2691	5670	-.40	.03	1.06	5.2	1.07	3.8	.39	66.3	68.8	IT08
10	2777	5670	-.47	.03	.92	-6.9	.90	-6.1	.50	71.7	68.7	IT10
18	2873	5670	-.56	.03	.99	-1.1	.98	-1.2	.45	69.0	68.6	IT18
7	3076	5670	-.73	.03	1.00	.2	1.01	.8	.43	68.9	68.6	IT07
13	3285	5670	-.92	.03	.98	-1.5	.96	-1.8	.44	69.8	69.1	IT13
6	3297	5670	-.93	.03	1.11	9.6	1.17	8.1	.33	64.6	69.1	IT06
5	3508	5670	-1.12	.03	.95	-4.2	.94	-3.0	.46	72.8	70.1	IT05
4	3716	5670	-1.31	.03	1.01	1.0	1.04	1.6	.39	71.3	71.5	IT04
3	3829	5670	-1.42	.03	1.00	-.2	.99	-.5	.40	72.8	72.4	IT03
2	3927	5670	-1.52	.03	.96	-3.0	.95	-2.0	.43	74.4	73.3	IT02
1	4028	5670	-1.62	.03	.90	-7.1	.85	-5.3	.47	77.4	74.3	IT01
MEAN	2323.6	5670.0	.00	.03	1.00	-.5	1.03	.4		74.7	74.1	
S.D.	972.8	.0	.97	.00	.06	4.5	.13	4.4		5.7	5.6	

Fuente: Elaboración propia a partir de las salidas del paquete Winsteps (2011).

Respecto a los análisis con los jueces expertos, cada juez trabajó por separado realizando su propia descripción de procesos y contenidos y su propia estimación de dificultad; posteriormente, se confrontaron los trabajos realizados por cada par de jueces.

Además, en este punto, se contrastan las expectativas de los jueces, en relación con el nivel de dificultad de los ítems.

Con el análisis minucioso realizado por cada juez, se desprendieron los contenidos y procesos que cada ítem representa; cabe indicar que, dada la forma en la que algunos de los ítems fueron creados, existe la posibilidad de que puedan ser resueltos mediante diferentes estrategias, lo cual puede implicar la anotación de varios contenidos y procesos para un mismo ítem. No obstante, es importante aclarar que si uno de los jueces presenta un nivel superior de concordancia con el modelo de Rasch, con respecto al otro juez que analizó su mismo grupo de ítems, se debe tener mayor confianza en su juzgamiento (debido a que sus estimaciones se asemejan a lo mostrado por los datos empíricos del modelo), por lo que su criterio es el que debe prevalecer en la tabla de especificaciones; pero si no existe una diferencia considerable entre jueces, se anotan todos aquellos procesos que ambos consideran como representativos del ítem.

No hubo grandes diferencias entre las estimaciones de dos jueces hacia un mismo ítem. Los jueces realizaron la estimación de dificultad de los ítems clasificando cada ítem en uno de los cinco niveles preestablecidos. Los jueces 1 y 2 presentaron una diferencia máxima de un punto entre sus estimaciones; mientras que los jueces 3 y 4 presentaron diferencias tanto de un punto como de dos puntos en sus estimaciones. Lo anterior resulta positivo, ya que en caso de que algún ítem presente una estimación extrema de dificultad por parte de los jueces, se debería analizar detalladamente su continuidad en los análisis.

Una vez obtenidas las estimaciones de dificultad, se procedió al establecimiento de las 5 categorías del nivel de dificultad de los ítems según el modelo de Rasch. Para dicha clasificación, se empleó el criterio de los jueces con respecto a la cantidad de ítems que se estimó en cada una de las 5 categorías de dificultad; de manera que se obtuvo el porcentaje de ítems que cada juez clasificó en cada categoría, para luego obtener el porcentaje promedio por categoría y, finalmente, determinar la cantidad de ítems que cada porcentaje promedio representa para el total de 29 ítems. En la Tabla 2 se exponen los porcentajes y la cantidad de ítems.

Tabla 2
Frecuencia y porcentaje de ítems clasificados en categorías de dificultad según jueces

		Categorías de dificultad					
		Nivel 1	Nivel 2	Nivel 3	Nivel 4	Nivel 5	Total ítems
Juez 1	N	0	4	8	3	0	15
	%	0,00	0,27	0,53	0,20	0,00	1,00
Juez 2	N	0	6	7	2	0	15
	%	0,00	0,40	0,47	0,13	0,00	1,00
Juez 3	N	2	4	2	4	2	14
	%	0,14	0,29	0,14	0,29	0,14	1,00
Juez 4	N	3	4	4	3	0	14
	%	0,21	0,29	0,29	0,21	0,00	1,00

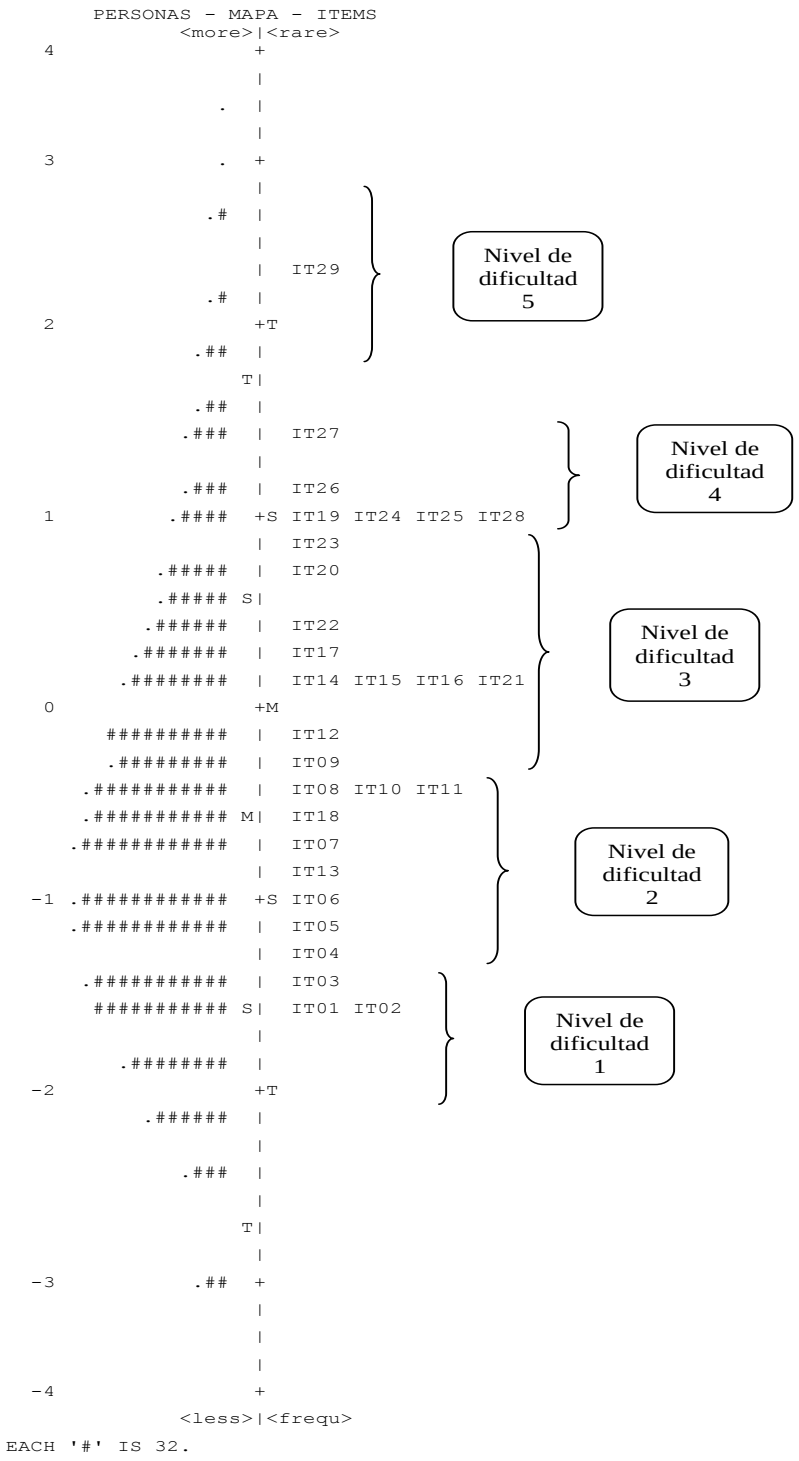
Fuente: Elaboración propia de los autores (2011).

A partir de los valores presentados en la Tabla 2, se obtienen las proporciones y porcentajes por categoría de dificultad, dividiendo el total de ítems clasificados en cada categoría entre el total de ítems posibles que pudo presentar dicha categoría en los juzgamientos. De esta manera, se obtuvo los siguientes valores: 8,62 para el primer nivel, 31,03 para el segundo, 36,21 para el tercero, 20,69 para el cuarto y 3,45 para el quinto nivel.

De acuerdo con los valores expresados en la tabla anterior, se calculó el número de ítems que representa el porcentaje de cada categoría con respecto al total de 29 ítems empleados. Los resultados fueron los siguientes: 3 ítems para el primer nivel, 9 ítems para el segundo, 10 ítems para el tercero, 6 ítems para el cuarto y 1 ítem para el quinto nivel.

El mapa del escalamiento conjunto de examinados e ítems mostrado en el Gráfico 2 identifica los ítems pertenecientes a cada una de las categorías de dificultad a partir de las estimaciones de los jueces.

Gráfico 2
Distribución conjunta de examinados e ítems para la determinación de los niveles de dificultad de los ítems



Fuente: Elaboración propia a partir de las salidas del paquete Winsteps (2011).

Según el gráfico anterior, únicamente el ítem 29 se clasificó en el nivel 5 (el más alto de dificultad), mientras que los ítems 01, 02 y 03 se clasificaron en el nivel 1 (el de menor dificultad). El resto de ítems se distribuyó en las restantes 3 categorías de dificultad. Subsiguientemente, se procedió a obtener un grado de concordancia entre lo estimado por cada uno de los jueces con lo clasificado desde la perspectiva del modelo de Rasch. Si los valores obtenidos por los jueces y por el modelo son los mismos, la concordancia se clasificó con un valor de 1; mientras que si no hubo concordancia, se clasificó con un valor de 0. De los 15 ítems que analizaron los jueces 1 y 2, el juez 1 concuerda en 6 ítems con la estimación del modelo de Rasch, mientras que el juez 2 concuerda en 5 de los ítems. No obstante, si se considera una diferencia de un punto, por arriba o por abajo, con respecto a la clasificación emanada del modelo de Rasch, tanto el juez 1 como el juez 2 presenta un 80% de acierto, lo cual es un porcentaje muy bueno. De esta forma, de los 14 ítems que analizaron los jueces 3 y 4, el juez 3 coincidió en 2 ítems con lo estimado mediante el modelo de Rasch, mientras que el juez 4 coincidió en 3 de los ítems. Sin embargo, al considerar una diferencia de un punto, el juez 3 presenta un 57,1 % de acierto y el juez 4 presenta un 71,4 % de acierto. Considerando lo anterior, por una parte, se delibera que los jueces 1 y 2 presentan una estimación cercana a la clasificación emanada del análisis con el modelo de Rasch; por lo tanto, se cree pertinente considerar el criterio de ambos con respecto a las estimaciones de contenidos y procesos que se representan en los ítems. Por otra parte, los jueces 3 y 4 presentan niveles de acuerdo más bajos; no obstante, el juez 4 presenta un mejor desempeño en cuanto a concordancia con los resultados del modelo de Rasch, puesto que si se considera una diferencia de un punto con respecto a la estimación de este, el juez 4 presenta un 71,4 % de acierto, mientras que el juez 3 presenta solo un 57,1 % de acierto, lo cual, para efectos de esta investigación, se considera un valor relativamente bajo. Por ello, se cree pertinente considerar solamente el criterio del juez 4 para la descripción de los contenidos y procesos que se representan en la tabla de especificaciones, notablemente el juez 4 posee mejores predicciones que el juez 3. Lo anterior, bajo la idea de que se debe tener mayor confianza o credibilidad en aquel juez que presente estimaciones más acertadas, ya que el nivel de dificultad que cada juez estima para un ítem debe explicar los contenidos y procesos que representan dicho ítem.

Finalmente, para modelar la tabla de especificaciones según los resultados obtenidos, se procedió a reunir, organizar y detallar en una sola tabla, los contenidos y procesos que representaban el grupo de ítems y la dificultad observada de cada uno. Mediante filas se

enumeran los ítems, mientras que en las columnas se introduce primeramente el nivel de dificultad del ítem (el obtenido mediante el modelo de Rasch, al ser el que realmente presentó el ítem en la población examinada) y, posteriormente, la lista de todos los procesos y la lista de todos los contenidos representados en el grupo de ítems. El valor de la dificultad del ítem se especifica numéricamente para cada uno, mientras que mediante una X se señala(n) cuál(es) contenido(s) y cuál(es) proceso(s) son representativos para cada ítem.

5. Conclusiones

Con respecto al contenido de la tabla de especificaciones, se definieron contenidos y procesos representativos del conjunto de ítems, contrastando las estimaciones de dificultad de los jueces con los resultados del modelo de Rasch, lo cual permitió obtener un grado aceptable de coincidencia. A partir de esta validación (de la apreciación de dificultad que estiman los jueces para cada ítem), se propone utilizar estos contenidos y procesos como una herramienta para la creación y juzgamiento de nuevos ítems.

Asimismo, se considera que para obtener mayores niveles de coincidencia, es recomendable que para estudios posteriores se empleen jueces que realmente posean un alto conocimiento tanto de los procesos de los ítems como de la población, de manera que la contribución sea cada vez más enriquecedora.

La creación de una tabla de especificaciones puede llevarse a cabo de diferentes maneras, sin embargo, un estudio que cuente con validación empírica (validez de la apreciación de dificultad que estiman los jueces para cada ítem) de un modelo psicométrico, como el modelo de Rasch, brinda mayor cuantía al procedimiento utilizado para la creación de una tabla de especificaciones.

La metodología empleada contribuye al conocimiento del constructo en medición, al exigir al investigador una definición explícita de procesos y contenidos que explican la dificultad del ítem.

La tabla de especificaciones construida constituye una primera aproximación para el Programa Permanente Prueba de Aptitud Académica en cuanto a estudios de este tipo; asimismo, al brindar evidencias empíricas de su validez, cumple un papel innovador en la prueba de aptitud académica. Dado lo anterior, se considera que, más allá de los resultados, su mayor importancia radica en el aporte metodológico, procedimental y de interpretación necesario para construir y validar empíricamente, con el modelo psicométrico de Rasch, tablas de especificaciones de su tipo.

6. Referencias

Bond, Trevor and Fox, Christine. (2001). *Applying the Rasch model: fundamental measurement in the human Sciences*. Mahwah, New Jersey: LEA.

Cadavid, Natalia, Delgado, Ana y Prieto, Gerardo. (2007). Construcción de una escala de depresión con el modelo de Rasch. *Psicothema,* 1*9*(3), 515-521.

Esquivel, Juan. (2001). *¿Cómo evaluar los aprendizajes en América Latina? El diseño de las pruebas para medir el logro académico: ¿Referencia a Normas o a Criterios?* Lima: PREAL.

Gómez, Juana e Hidalgo, María Dolores. (2003). Desarrollos recientes en psicometría. *Avances en Medición, 1*(1), 17-36.

Gori, Enrico and Battauz, Michela. (2006). The Rasch approach to "objetive measurement" in the presence of subjective evaluation from "Judges". *Revista Educación XXI*, (9), 107-133.

Instituto Nacional para la Evaluación de la Educación (INEE). (2005). *PISA para docentes: La evaluación como oportunidad de aprendizaje*. México: INEE.

Martínez, Rosario. (2005). *Psicometría: Teoría de los Tests Psicológicos y Educativos*. Madrid: Editorial Síntesis, S.A.

Messick, Samuel. (1989). Meaning and Values in Test Validation: The Science and Ethics of Assessment. *Educational Researcher, 18*(2), 5-11.

Messick, Samuel. (1995). Standards of validity and the validity of standards in performance assessment. *Educational Measurement: Issues and practice*, *14*(4), 5-8.

Montero, Eiliana. (2001). La teoría de respuesta a los ítems: una moderna alternativa para el análisis psicométrico de instrumentos de medición. *Revista de matemática: Teoría y Aplicaciones, 7*(1-2), 217-228.

Muñiz, José. (1997). *Introducción a la teoría de respuesta a los ítems.* Madrid: Ediciones Pirámide, S.A.

Prieto, Gerardo y Delgado, Ana. (2003). Análisis de un test mediante el modelo de Rasch. *Psicothema, 15*(1), 94-100.

Prieto, Gerardo and Delgado, Ana. (2007). Measuring Math Anxiety (in Spanish) with the Rasch Rating Scale Model. *Journal of Applied Measurement, 8*(2), 149-160.

Prieto, Gerardo y Dias, Angela. (2003). Uso del modelo de Rasch para poner en la misma escala las puntuaciones de distintos tests. *Actualidades en Psicología, 19*(106), 5-23.

Tristán, Agustín y Vidal, Rafael. (2006). *Estándares de calidad para pruebas objetivas*. Bogotá: Cooperativa Editorial Magisterio.