Actualidades en Psicología, 29(119), 2015, 153-165 ISSN 2215-3535
http://revistas.ucr.ac.cr/index.php/actualidades DOI: http://dx.doi.org/10.15517/ap.v29i119.18693

Análisis de la confiabilidad de los resultados de la Prueba de Diagnóstico Matemática en la Universidad Nacional de Costa Rica utilizando el modelo de Rasch

Reliability Analysis Diagnostic Mathematics Test at the National University of Costa Rica

José Andrey Zamora Araya[1]
Universidad Nacional, Costa Rica

Resumen. El objetivo del presente estudio es evidenciar cómo la implementación de la teoría de respuesta a los ítems, en particular el modelo de Rasch, ha logrado mejorar los índices de confiabilidad de la prueba de diagnóstico matemático en la Universidad Nacional de Costa Rica, pasando de un alfa de .51 en el 2010 a uno de .78 en el 2012. El análisis psicométrico permitió brindar recomendación acerca de la construcción de la prueba que redundaron en mejores indicadores y elaboración de ítems para medir los diferentes niveles de habilidad que es el propósito fundamental de una prueba diagnóstica.

Palabras clave. Confiabilidad, Análisis de Rasch, Matemática, Diagnóstico.

Summary. The goal of this study is to show how the implementation of the theory of item response, particularly the Rasch model has improved the reliability indices math test diagnosis at the National University of Costa Rica, from an alpha of 0.51 in 2010 to one of 0, 78 in 2012. The psychometric analysis allowed providing recommendation regarding the construction of the test which resulted in better indicators and preparation of items to measure the different levels of skill that is the fundamental purpose of a diagnostic test.

Keywords. Reliability, Rasch Analysis, Mathematics, Diagnosis.

[1]José Andrey Zamora Araya. Escuela de Matemática de la Universidad Nacional de Costa Rica y Escuela de Estadística de la Universidad de Costa Rica. Dirección Postal: 86-3000, Universidad Nacional de Costa Rica. E-mail: andreyzamora@gmail.com

Introducción

La prueba de diagnóstico en matemática (PDM) en la Universidad Nacional Autónoma de Costa Rica (UNA), surge como una necesidad de obtener información ante los bajos resultados obtenidos por los y las estudiantes en los cursos introductorios de matemática. Inicialmente la PDM se aplica a todos aquellos y aquellas estudiantes de primer ingreso que dentro de su plan de estudios deban llevar al menos un curso de matemática. En particular la prueba la realizan los y las estudiantes que cursan las carreras de economía, ingeniería en sistemas, química, planificación social, ingeniería en topografía y todas aquellas carreras que tienen en su malla curricular el curso de Matemática General. La primera aplicación, que se tenga registro de pruebas similares en la UNA se realizó en el año 2009 en la sede central, no obstante, no se cuenta con ningún informe técnico acerca de su construcción o resultados. Posteriormente, en los años 2010, 2011 y 2012, se aplicó la PDM también en las sedes regionales. Para dichas pruebas se cuentan con informes que dan detalle sobre el tipo de ítem evaluado, porcentaje de aprobación por pregunta y algunas estadísticas descriptivas; como lo muestra la tabla 1. La prueba se ha seguido aplicando hasta el año 2015, sin embargo, a partir del 2013 no se cuenta con registros referentes al análisis psicométrico de la PDM.

Hasta el año 2011, la construcción de la prueba se ha realizado de manera empírica, y se ha tomado como referencia ítems de similares características a los que presentan las pruebas nacionales de bachillerato en las áreas de álgebra, números reales, funciones y trigonometría; pero sin un análisis psicométrico que permita determinar la validez y confiabilidad de sus resultados. Por tanto, se requiere que la PDM sea elaborada y analizada con criterios técnicos, para de esta manera realizar un diagnóstico de fortalezas y debilidades, en cuanto a conocimientos y habilidades matemáticas de los y las estudiantes de primer ingreso en la UNA.

¿La PDM aplicada en la UNA es confiable?, basados en los resultados de la PDM, ¿es posible brindar recomendaciones a los y las estudiantes para mejorar su rendimiento académico en matemática?

Para responder estás preguntas, se propone utilizar los modelos de Rasch para realizar el análisis psicométrico de la PDM para los años 2010, 2011 y 2012. De esta forma se pudo evaluar la confiabilidad de los ítems que conforman la PDM y brindar recomendaciones a cerca de los resultados obtenidos con el fin de mejorar los indicadores de confiabilidad de la PDM en la UNA.

Pruebas de Diagnóstico

Para referirse a las pruebas de diagnóstico, es necesario en primera instancia tratar el tema de dominio educativo. El dominio educativo, por lo general, no se refiere a un verdadero dominio (desempeño sobresaliente) sino a un desempeño satisfactorio a cerca de los contenidos propuestos (Nunnaly & Berstein, 1995). El dominio educativo suele ser muy amplio, por lo general referido a una materia o disciplina (Matemáticas, Lenguaje) en función de los objetivos de un período educativo o de las dimensiones de las mismas (Álgebra, Geometría, Comprensión de Lectura). Además, los límites del dominio no siempre son claros, por una parte debido a la dificultad de la definición del dominio como tal

Tabla 1

Estadísticas descriptivas de las calificaciones obtenidas en la PDM durante los años 2010, 2011 y 2012.

Estadístico	Año 2010	Año 2011	Año 2012
Total de estudiantes	1194	1051	1279
Mínimo	5	2.5	0
Máximo	92.5	82.5	90
Rango	87.5	80	90
Media	31.7	30.7	33.94
Mediana	30	29.3	31.67
Moda	27.5	27.5	30
Desviación estándar	11.4	10.4	12.71
Percentil 25	25	23.9	25
Percentil 50	30	29.3	31.67
Percentil 75	37.5	35.6	38.33

Nota. La escala utilizada en la PDM es de 0 a 100. Fuente: Escuela de Matemática Universidad Nacional

y por otra, porque las pruebas por lo general tienen la finalidad de evaluar a una gran cantidad de sujetos de una población provenientes de diferentes modelos didácticos y curriculares (Jornet & Suárez, 1996).

En Costa Rica, las pruebas de diagnóstico en el área de matemática a nivel universitario las realizan las principales universidades, siendo la que posee mayor experiencia la Universidad de Costa Rica (UCR), que año tras año, administra una prueba de diagnóstico de conocimientos y destrezas en matemática, conocida como DiMa, que es una prueba que aplica la Escuela de Matemática de la UCR, a estudiantes de primer ingreso, cuyo plan de estudio incluye uno de los siguientes cursos de Matemática: MA0230, MA1001 o MA1210, que son cursos de cálculo diferencial para las carreras de las áreas de: economía, ciencias básicas e ingenierías, ciencias de la salud y agroalimentarias. A partir del 2006, también la realizaron estudiantes de las carreras de computación, matemáticas e informática empresarial, cuyos primeros cursos de matemática son: MA0129, MA0150 y MA0320, respectivamente (Jiménez, 2010).

Las estadísticas de promoción muestran que el nivel de repitencia en estos cursos ronda el 30%. Ante este panorama, la Escuela de Matemática crea la prueba diagnóstico con el fin de alertar al estudiante de sus posibles deficiencias y ofrecerle, a la vez, opciones para remediarlas. Entre las recomendaciones se consideran los talleres de nivelación (se ofrecen en febrero, son intensivos y gratuitos. Se encuentran a cargo de los Centros de Asesoría Estudiantil, CASE) y el curso MA0110 Matemática Básica. El DiMA se aplica con el objetivo de conocer el grado de dominio de los temas de matemática con el que los y las estudiantes ingresan a la Universidad (UCR, 2011).

Por su parte, el Instituto Tecnológico de Costa Rica (ITCR) motivado por los resultados en los cursos de Matemática General, Matemática Básica y Fundamentos de Matemática I, donde el promedio de aprobación es de un 44.43%, 42.44% y un 51.32% en los períodos 2003-2009, 2000 al 2009 y 2005-2009 respectivamente, decide implementar una prueba de conocimientos en matemática que permita obtener con certeza, la cantidad y calidad de la información con el que ingresan los y las estudiantes matriculados/as en los cursos de matemática. El objetivo de la prueba es el de convertirse en un instrumento de predicción del rendimiento académico en los cursos de Matemática General, Matemática Básica para Administración y Fundamentos de Matemática I (Ramírez & Barquero, 2011).

La prueba del ITCR tiene la característica de que está conformada por ítems de desarrollo en su totalidad y se realiza para conocer las fortalezas y debilidades de las y los estudiantes admitidos. La idea es prevenir la reprobación, la repetición consecutiva de cursos, la deserción, y por ende, incrementar los índices de graduación; además se pretende que la información de la prueba sea utilizada para tomar medidas a favor de los y las estudiantes como planes o programas de apoyo en el área psicoeducativa, cursos de nivelación, métodos de estudios, tutorías, entre otros (Ramírez & Barquero, 2011).

En el caso de la Universidad Estatal a Distancia (UNED), desde el tercer cuatrimestre del año 2010 realiza exámenes diagnósticos de manera virtual, apoyado por el proyecto Rendimiento Académico en Matemáticas (RAMA) del Consejo Nacional de Rectores (CONARE), con el propósito de conocer las debilidades y fortalezas de los universitarios que llevan alguna materia de matemáticas en sus carreras profesionales. Por la modalidad educativa que tiene la UNED, se decidió que los exámenes tendrían que ser en línea, apoyados en la plataforma MOODLE. La iniciativa está enfocada en la orientación del estudiante y es por eso que se crearon módulos de aprendizaje. Para ello, se diseñó el sitio web http://euclides.uned.ac.cr/rama/ donde el universitario realiza sus pruebas previo a la matrícula o días después de ella. Posteriormente, de acuerdo con la evaluación, la Universidad guiará al estudiante a mejorar su rendimiento académico, fortaleciendo esas áreas en que mayor dificultad presenta (Kcuno, 2010).

La Universidad Nacional a partir del año 2010 inicia con la aplicación masiva de la PDM como resultado de los malos resultados de los y las estudiantes en los cursos introductorios en especial en el curso de

Matemática General, el principal curso de servicio ofrecido por la Escuela, Entiéndase curso de servicio aquel que brinda la Escuela de Matemática a carreras deferentes al Bachillerato y Licenciatura en Enseñanza de la Matemática. Al iniciar la aplicación de las pruebas de diagnóstico, paralelamente surge la iniciativa de realizar un análisis psicométrico apropiado para este tipo de evaluación y particularmente en el caso de la UNA se decide aplicar los modelos de Rasch, perteneciente a los modelos de la teoría de respuesta a los ítems.

Análisis de confiabilidad. La confiabilidad de una prueba o instrumento se refiere a la consistencia de las calificaciones obtenidas por los mismos individuos en diferentes ocasiones o con distintos conjuntos de reactivos equivalentes (Arginay, 2006). Para Muñiz et al. (1997), la confiabilidad indica el grado con el que las diferencias individuales en las calificaciones de las pruebas, se atribuyen a errores aleatorios de la medición y el grado con el que se asignan a diferencias reales de las características o dominio en consideración.

También se utilizan como sinónimo de confiabilidad el de estabilidad de la medida y el de consistencia interna. La estabilidad de la medida tiene que ver con el hecho de que el atributo psicológico medido con un determinado instrumento, será confiable siempre y cuando al evaluar a los mismos sujetos con el mismo instrumento, las medidas obtenidas en la segunda aplicación sean muy parecidas a las primeras, es decir, son estables a través del tiempo, lo que significa que los errores de medición son mínimos y en consecuencia se tendría una razonable medida de confiabilidad, atribuyéndose las diferencias encontradas entre una medición y otra a los errores aleatorios asociados al proceso de medición y no al instrumento (Muñiz, 1992).

La consistencia interna de un instrumento se refiere al hecho de que los reactivos que lo constituyen son consistentes entre sí, es decir, midan lo mismo o evalúen el mismo atributo psicológico propuesto. Esto significa que los sujetos de manera individual puntearán alto en aquellos reactivos que tienden a medir dicho atributo y puntearán bajo en aquellos que no lo miden, siendo así consistentes los reactivos entre sí en la evaluación del atributo por evaluar (Aragón, 2004).

Una de las formas de aproximarse a la confiabilidad, en la teoría clásica de los test, es por medio del llamado coeficiente alfa propuesto por Cronbach en 1951, que es un índice usado para medir la consistencia interna de una escala, es decir, evalúa la magnitud en que los ítems de un instrumento están correlacionados. También se puede interpretar este coeficiente como la medida en la cual un constructo o rasgo latente está presente en cada ítem (Celina & Campo, 2005).

En este estudio se ha decidido adoptar los modelos de Rasch, perteneciente a los modelos de la teoría de respuesta a los ítems (TRI), para evaluar el grado de confiabilidad de la PDM aplicada en la UNA.

Confiabilidad desde la teoría de respuesta a los ítems

La TRI constituye un enfoque para la medición psicológica y educativa que ha dado lugar a un significativo avance en la tecnología para la construcción y análisis de los test. Como parte de sus características y conceptos se pueden citar las funciones de información de los ítems y del test, errores típicos de medida distintos para cada nivel de la variable medida o el establecimiento de bancos de ítem con parámetros estrictamente definidos. Una de las grandes ventajas es que estos últimos posibilitan la construcción de test adaptados al nivel del examinado, permitiendo así exploraciones exhaustivas y rigurosas en función de las características de los sujetos. Un modelo TRI, se puede definir como:

> Un modelo TRI es una conceptualización, que partiendo de ciertos conceptos básicos de medición y usando las herramientas de la estadística y la matemática, busca encontrar un descripción teórica para explicar el comportamiento de datos empíricos derivados de la aplicación de un instrumento psicométrico. Los parámetros estimados por el modelo permiten entonces evaluar la calidad técnica de cada uno de los ítems por separado y del instrumento como un todo y a la vez estimar el nivel que cada examinado presenta en el constructo (o habilidad) de interés. En un modelo de TRI se asume que hay una variable latente o constructo θ, no observable directamente y que se desea estimar

> para cada examinado a partir de las respuestas suministradas por este en el instrumento de medición. Además, se asume que para cada ítem o pregunta el comportamiento de las respuestas dadas por los examinados puede ser modelado mediante una función matemática que se denomina curva característica del ítem o CCI. Otros conceptos fundamentales en TRI son la función de información del test y el error estándar de medición (Montero, 2000, p. 220).

Uno de los supuestos de los modelos TRI formula la existencia de una relación funcional entre los valores de la variable que miden los ítems y la probabilidad de dar una respuesta correcta, dicha función se conoce como Curva Característica del Ítem (CCI). La variable por medir suele ser un rasgo que no es directamente observable, como puede ser el nivel de habilidad o aptitud. En los modelos más simples de la TRI este rasgo latente se considera unidimensional, es decir, se representa como una variable que toma valores en la recta real, los cuales determinan totalmente la probabilidad de elegir cada una de las posibles respuestas para el ítem. Por ejemplo, para un ítem que mide algún tipo de habilidad, la probabilidad de respuesta correcta de dos sujetos será la misma si y solo si dichos sujetos son igualmente hábiles (Nunnally & Bernstein, 1995).

Existen varios modelos para medir un rasgo latente, que se fundamentan en la TRI como son: El modelo logístico de un parámetro (conocido como Modelo de Rasch), el modelo de dos parámetros y el modelo de tres parámetros. Cada uno de ellos tiene sus características, ventajas y desventajas. Por su simplicidad y aplicabilidad en el ámbito educativo para la elaboración de los análisis del estudio se elige trabajar con el modelo de Rasch.

Modelo de Rasch

El modelo de Rasch, fue propuesto por el matemático Georg Rasch y es solamente uno de una familia completa de modelos descrita por Rasch en su texto de 1960. El modelo de medida permite solventar muchas de las deficiencias de la teoría clásica de los test (TCT) y construir pruebas más adecuadas y eficientes, por lo que es muy apropiado su uso en el ámbito de la evaluación psicológica y educativa. El modelo es una formulación matemática que enlaza la probabilidad del resultado a las características de la persona y el ítem, cuando un solo individuo intenta resolver un ítem. Rasch es uno de los modelos de la familia de rasgo latente para la medición de logro y se puede decir que es uno de los más simples de esta familia (Choppin, 1983). Puede ser escrito de la siguiente forma:

$$P\left(X_{vi}=1\right)=\frac{\theta_V}{\theta_v+b_i} \qquad (1)$$

Donde:

$P(X_{v\,i} = 1)$ es la probabilidad de que la persona v responda correctamente al ítem i, y el valor sería cero en cualquier otro caso

θ_v: Es un parámetro que describe la habilidad de la persona v

b_i: Es un parámetro que describe la dificultad del ítem i.

Según Wright y Stone (1998) en esta formulación θ y b pueden variar de cero a más infinito, pero usualmente se realiza una transformación para simplificar el análisis matemático de la función. En la transformación más usada se hace un ajuste con la constante e, que es la base de los logaritmos naturales (e =2.72) y en su formulación más conocida el modelo describe la predicción de la probabilidad de una respuesta al ítem (resolverlo correctamente, estar de acuerdo, etc.) a partir de la diferencia en el atributo entre el nivel de la persona (θ_v) y el nivel del ítem (b_i), cuya representación matemática está dada por la fórmula:

$$\mathrm{P_{vi}}\left(\theta\right)=\frac{\mathrm{e}^{\left(\theta_v-\mathrm{b_i}\right)}}{1+\mathrm{e}^{\left(\theta_v-\mathrm{b_i}\right)}} \qquad (2)$$

Donde θ y b pueden tomar cualquier valor real y la medición de la habilidad de la persona y la dificultad del ítem están en la misma escala llamada logit. La expresión ($\theta_v - b_i$) indica el resultado probable de la interacción persona – ítem. Precisamente son estas características

las que hacen del modelo de Rasch una metodología de análisis para ítems de pruebas educativas muy valiosa y relativamente fácil de interpretar en los contextos escolares, razón por la cual se adoptó para fines de la presente investigación.

Método

Participantes

Los participantes del estudio son aquellos y aquellas estudiantes de nuevo ingreso a la UNA que realizaron la PDM durante los años 2010, 2011 y 2012 que fueron un total de 1194, 1050 y 1279 respectivamente. Las bases de datos fueron suministradas por el departamento de registro de la UNA, sin embargo, a excepción de la información referente a 2010; los archivos sólo contienen la boleta de identificación y las respuestas de los y las estudiantes a los reactivos, por lo que no se cuenta con información referente a otras variables como sexo, zona de residencia o colegio de procedencia. Los datos referentes al año 2010 se presentan en la tabla 2, cabe resaltar que el total de estudiantes que aplicaron la prueba ese año fue de 1194, pero la información relacionada con las variables sociodemográficas de 51 de ellos no se encontraban en la base de datos.

Tabla 2

Principales variables sociodemográficas de los y las estudiantes de primer ingreso que efectuaron la PDM en la UNA para el año 2010.

Variables		
Sexo		
Hombre	605	52.9%
Mujer	538	47.1%
Total	1143	100.0%
Zona		
Urbano	730	63.9%
Rural	413	36.1%
Total	1143	100.0%
Colegio de procedencia		
Público	929	81.4%
Privado	156	13.7%
Subvencionado	56	4.9%
Sin ubicar	2	0.17%
Total	1143	100.0%

Instrumentos

El instrumento aplicado fue la PDM para los años en análisis. Dicha prueba es elaborada por la Escuela de Matemática de la UNA y consta de preguntas relativas a contenidos abarcados en la educación secundaria en los tópicos de aritmética, números reales, álgebra, funciones y trigonometría. La PDM se construye con el propósito de obtener información acerca del nivel de habilidades y conocimientos matemáticos que posee el estudiante de nuevo ingreso a la UNA.

Cabe resaltar está es la primera experiencia que tiene la Escuela de Matemática de la UNA en la elaboración de pruebas diagnósticas y por tanto no se contaba con ningún banco de ítems que ayudara a la elaboración de la PDM y tampoco se contó con una aplicación piloto. Para el ensamblaje de la prueba solo se tomaron como criterios que el ítem fuera de una redacción similar a la de una prueba de bachillerato y que abarcaran los contenidos de números reales, aritmética, álgebra, funciones y trigonometría, luego los miembros de la comisión resuelven y discuten los ítems para su posible incorporación a la prueba.

Una comisión de académicos quienes redactan y revisan los ítems de la prueba son los encargados de llevar a cabo el proceso de elaboración y aplicación de las pruebas en colaboración con el proyecto éxito académico. La PDM se aplica a inicios del mes de febrero de cada año, una vez que se ha culminado con el proceso de matrícula de los y las estudiantes de nuevo ingreso y en coordinación con el departamento de orientación y el proyecto éxito académico se convoca a la población estudiantil para la aplicación de la prueba.

Procedimientos y estrategia de análisis

Se verificaran los principales supuestos del análisis

de Rasch como lo son la independencia local de los ítems y la unidimensionalidad de la prueba. En el caso del primero supuesto, los ítems fueron construidos de manera tal que no tuvieran dependencia entre sí y se aplicaron protocolos de cuido de exámenes a la hora de aplicar la prueba para minimizar la posibilidad de copia entre los y las estudiantes. Estos protocolos se estandarizaron para aplicarlos en todas las sedes de la UNA donde se realiza la PDM.

En el caso de la unidimensionalidad de la prueba se aplicó un análisis de componentes principales para determinar el porcentaje de varianza total explicada por el primer componente y de esta manera tener una medida relacionada con la unidimensionalidad de los datos. No obstante, como lo señala Muñiz (1997) la unidimensionalidad perfecta es rara y "la unidimensionalidad se convierte en una cuestión de grado, cuanta más varianza explique el primer factor más unidimensionalidad existirá" (p.26).

Por otra parte, los estadísticos de ajuste para la calibración de ítems más usados en el modelo de Rasch son los valores MNSQ de INFIT y OUTFIT. El INFIT MNSQ es un estadístico de ajuste calculado a partir de las medias cuadráticas sin estandarizar, cuyo valor esperado es 1. De acuerdo con Smith, Schumaker y Bush, 1995 (como se citó en Prieto & Delgado, 2003) se considera que los valores superiores a 1,3 o inferiores a 0,7 indican desajuste en muestras con menos de 500 casos, 1,2 en muestras de tamaño medio (entre 500 y 1000 casos) y 1,1 en muestras con más de 1000 casos para tipos de pruebas de escogencia única.

El estadístico OUTFIT es el promedio de los residuales estandarizados derivados tanto de los examinados como de los ítems. Su valor se interpreta como una media cuadrática no ponderada sensible a los comportamientos extremos no esperados en los patrones de respuesta. Este estadístico de ajuste es sensible a valores extremos y a comportamientos no esperados que afectan respuestas a ítems que se encuentran lejos del nivel de habilidad del sustentante.

Debe notarse que ambos INFIT y OUTFIT se obtienen de la suma de cuadrados de la diferencia entre la expectativa del modelo y los residuales (o diferencias observadas) para cada ítem y para cada examinado (Bond & Fox, 2001). El valor esperado de estos estadísticos es 1 y se considera que los valores superiores a 1.5 para personas y superiores a 1.3 para ítems indican un desajuste moderadamente alto (Wright & Linacre, 1998 citado en Prieto et al., 2007). Por su parte, el OUTFIT es un indicador muy sensible a los valores extremos (basta una respuesta muy inesperada para que adopte un valor muy elevado), en cambio el INFIT es más robusto: los valores altos se deben a patrones de respuesta aberrantes, por ello en el análisis se usará el criterio de INFIT mayor a 1.3 tanto para ítems y 1.5 para personas. (Wright y Linacre, 1998 citado en Prieto et al., 2007).

Un valor MNSQ de 1 + x explica 100 x % más variación entre los datos observados y los patrones de respuesta esperados si el modelo y los datos observados fueran compatibles. Por lo tanto, valores superiores a 1.30 indican 30% más de variación entre lo que el modelo predice y los patrones de respuesta observados de hecho, por ello no se recomiendan valores superiores a 1.3. Igualmente valores menores a 1.00, como pueden ser 0.80 indican 20% menos de variación que la esperada bajo el modelo. (Bond & Fox, 2001).

Resultados

Para la verificación del supuesto de unidimensionalidad se ejecutó un análisis de componentes principales a los tres conjuntos de datos PDM 2010, PDM 2011 y PDM 2012 y los porcentajes de varianza explicada para el primer factor fueron respectivamente 7.41% ; 7.9% y 9.4%. A pesar de que dichos valores son relativamente bajos, como lo señala Muñiz (1997) la unidimensionalidad es una cuestión de grado, por lo que no podemos descartar la presencia de unidimensionalidad en la prueba.

La figura 1 presenta el gráfico de sedimentación para la PDM de 2010. Entre los criterios para determinar el número de factores se encuentra el Catell, (como se citó en Cea D'Ancona, 2002) sugiere que se consideren todos aquellos factores situados antes del punto en el

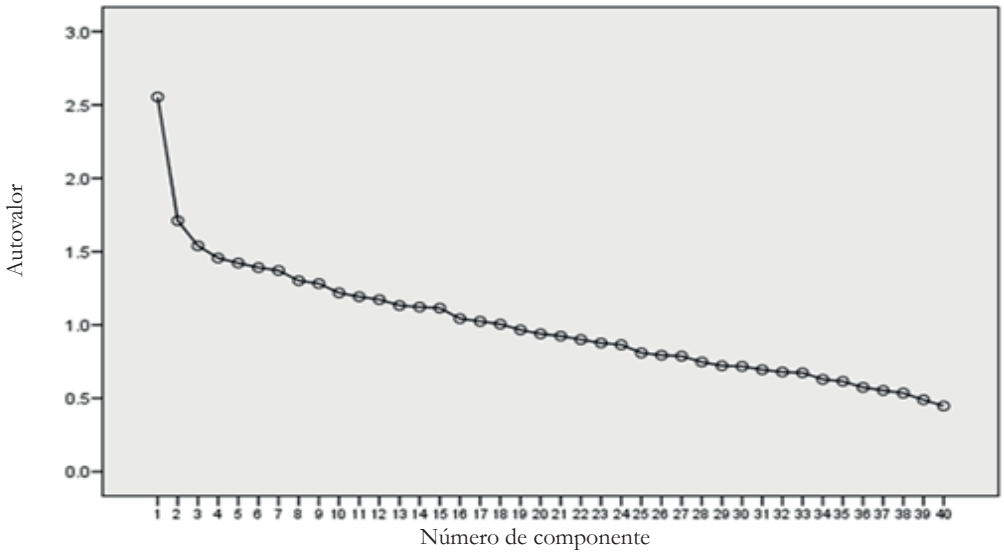

Figura 1. Gráfico de sedimentación para la PDM 2010

que se presenta un cambio importante en la trayectoria de caída de la pendiente, en este caso, el gráfico sugiere la existencia de un componente predominante que agrupa los ítems, aunque no permite garantizar la unidimensionalidad de la prueba.

En cuanto al análisis de los estadísticos de ajuste tanto para el caso de los ítems como para el de personas los valores de INFIT y OUTFIT se encuentran dentro de los valores considerados aceptables, por lo que no existe necesidad de eliminar reactivos o sujetos, como puede apreciarse en las tablas 3 y 4.

Como se mencionó anteriormente, aunque se

Tabla 3

UNA: Valores de INFIT y OUTFIT para ítems de la PDM período 2010-2012.

Año	2010		2011		2012	
INFIT ítems	Número de ítems	Porcentaje	Número de ítems	Porcentaje	Número de ítems	Porcentaje
Menores a 0.70	0	0.0%	0	0.0%	0	0.0%
Entre 0.70 y 0.90	0	0.0%	1	2.5%	6	10.0%
Entre 0.9 y 0.99	19	47.5%	18	45.0%	23	38.3%
Entre 1 y 1.3	21	52.5%	21	52.5%	31	51.7%
Total	40	100.0%	40	100.0%	60	100.0%
OUTFIT ítems						
Menores a 0.70	0	0.0%	0	0.0%	0	0.0%
Entre 0.70 y 0.90	1	2.5%	3	7.5%	7	11.7%
Entre 0.9 y 0.99	16	40.0%	18	45.0%	21	35.0%
Entre 1 y 1.3	23	57.5%	19	47.5%	32	53.3%
Total	40	100.0%	40	100.0%	60	100.0%

Tabla 4

UNA: Valores de INFIT y OUTFIT para personas de la PDM período 2010-2012.

Año	2010		2011		2012	
	Número de personas	Porcentaje	Número de personas	Porcentaje	Número de personas	Porcentaje
INFIT personas	0	0.0%	0	0,0%	0	0,0%
Menores a 0.70	0	0.0%	0	0.0%	0	0.0%
Entre 0.70 y 0.90	205	17.2%	195	18.6%	125	9.8%
Entre 0.9 y 0.99	415	34.8%	347	33.0%	524	41.0%
Entre 1 y 1.5	574	48.1%	509	48.4%	630	49.3%
Total	1194	100.0%	1051	100.0%	1279	100.0%
OUTFIT personas						
Menores a 0.70	21	1.7%	12	1.1%	2	0.1%
Entre 0.70 y 0.90	291	24.4%	294	28.0%	320	25.1%
Entre 0.9 y 0.99	311	26.0%	275	26.2%	406	31.7%
Entre 1 y 1.5	559	46.8%	455	43.3%	532	41.6%
más de 1.5	12	1.0%	15	1.4%	19	1.5%
Total	1194	100.0%	1051	100.0%	1279	100.0%

presentan los estadísticos de ajuste INFIT y OUTFIT se tomará como referencia los valores de INFIT. En todos los períodos analizados los valores presentan una magnitud apropiada y no superan el valor de 1.3 para el caso de los ítems y de 1.5 para el caso de las personas.

En cuanto a los índices de confiabilidad, para el caso de los reactivos se ha mantenido constante a lo largo del período de análisis (0.99) evidenciando una gran consistencia en las estimaciones del parámetro de dificultad. Contrariamente, en el caso de la confiabilidad de las personas el índice presenta un comportamiento irregular con valores bajos para los años 2010 y 2011, teniendo un repunte para el año 2012 como se muestra en la tabla 5.

Tabla 5

UNA: Valores de los índices de confiabilidad para la PDM período 2010-2012, según los resultados arrojados por el modelo de Rasch.

	Año		
Índice	2010	2011	2012
Confiabilidad Personas	0.51	0.61	0.78
Confiabilidad ítems	0.99	0.99	0.99

Es necesario recordar que para la aplicación de las primeras pruebas 2010 y 2011 la Escuela de Matemática de la UNA no contaba con una metodología de análisis para la PDM, pues fue precisamente en el año 2011 donde se inició con el análisis psicométrico de Rasch para las pruebas ya aplicadas. Como resultado de los análisis para estos dos primeros años se brindaron recomendaciones como aumentar la cantidad de ítems y de esta forma incluir reactivos de temáticas más afines a los temas evaluados en la educación secundaria como lo son preguntas referidas a la aritmética. Este cambio también se vio motivado, pues al analizar la PDM 2010 y la PDM 2011 se observa que el grado de dificultad de la prueba es muy alto para el nivel de habilidad que poseen los sujetos.

Por ejemplo, para el caso de la PDM 2010 se puede observar en la figura 2 como el nivel promedio de habilidad de los y las estudiantes está a poco menos de dos desviaciones estándar del nivel de dificultad

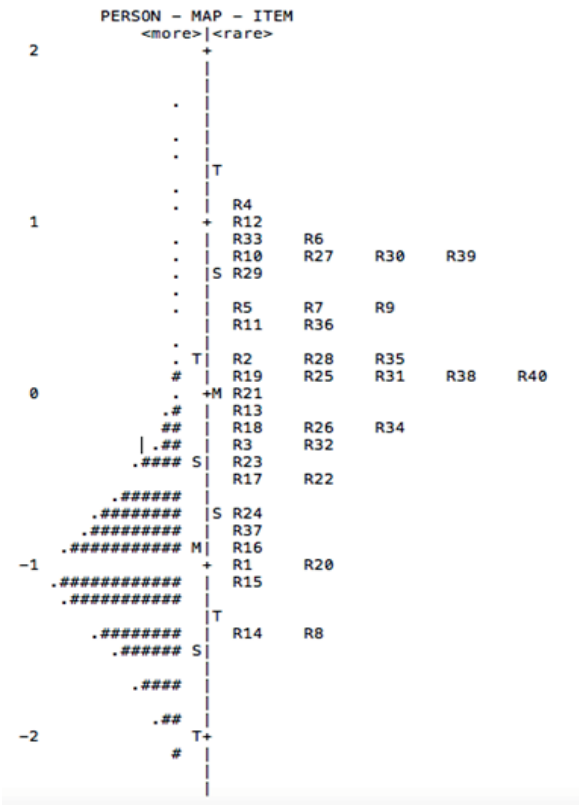

Figura 2. Mapa de personas versus ítems para la PDM 2010

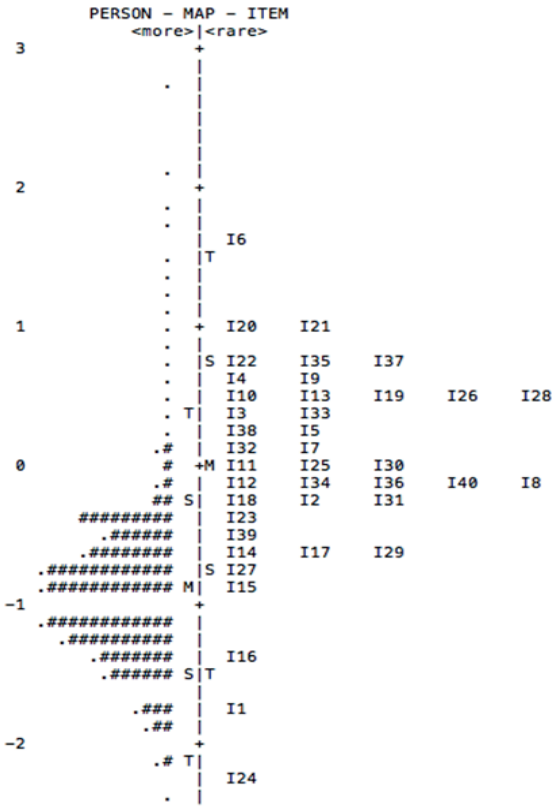

Figura 3. Mapa de personas versus ítems para la PDM 2011

promedio de la prueba y para la PDM 2011 la diferencia es cercana 1.5 desviaciones estándar como se puede apreciar en la figura 3. El alto nivel de dificultad en comparación con el nivel de habilidad podría explicar en parte, el bajo nivel de confiabilidad que arrojan las pruebas para los años 2010 y 2012, pues no existen suficientes ítems para evaluar con exactitud los niveles bajos de habilidad.

Por ejemplo en la figura 2 se puede apreciar que los ítems con mayor probabilidad de ser contestados correctamente por la población que realizó la prueba son el R14 y R8 y R15, en contrasta los que tienen menor probabilidad de contestarse correctamente son el R4, R12 y R33. La mayoría de las personas presentan niveles de habilidad menores en comparación con el nivel de dificultad de los ítems (en los mapas de las figuras 2 y 3 el símbolo # representa una cantidad de 10 personas y el • un conjunto de personas entre 1 y 9).

Algo similar ocurre en la figura 3 donde los ítems con menor probabilidad de acierto son I6, I20 y I21 y los que poseen mayor probabilidad de ser acertados son el I24, I1 y I16. Para ambos años, los ítems que presentaron mayores problemas en su resolución se asocian a las áreas de trigonometría y funciones, en contraste, ítems referentes a números reales y ecuaciones algebraicas fueron los que presentaron mayores niveles de acierto.

Nótese como tanto en la figura 2 como en la 3, existen muchos individuos en niveles bajos de la escala, pero pocos ítems que permitan evaluar su desempeño, contrariamente, en la parte alta de la escala existen muchos ítems, pero pocas personas con una alta probabilidad de contestarlos correctamente.

Por ello, una de las principales recomendaciones para la PDM 2012 fue la inclusión de más ítems que permitieran medir con mayor precisión los niveles bajos de habilidad lo que significó un aumento en el número

de preguntas con respecto a las versiones anteriores de la prueba. Como se muestra en la tabla 5, dicha sugerencia permitió elevar el nivel de confiabilidad de las personas a un valor cercano al 80%, lo cual es deseable si los resultados de la PDM se toman como base para la toma de decisiones tendientes a mejorar el rendimiento académico de los y las estudiantes de primer ingreso.

Discusión

El modelo de Rasch, en el caso de la PDM de la UNA, ha mostrado ser una herramienta útil en el análisis de las propiedades psicométricas de los ítems que la componen. Iniciando con los supuestos básicos del modelo como lo son la unidimensionalidad de la prueba y la independencia local de los ítems, ambos se cumplen satisfactoriamente, a pesar del relativo bajo porcentaje de varianza explicada, la unidimensionalidad es cuestión de grado, por lo que el porcentaje no necesariamente indica ausencia de unidimensionalidad (Muñiz, 1997).

A pesar de lo anterior, al aplicar el modelo de Rasch se muestra un buen ajuste de los datos al modelo, pues tanto los ítems como los sujetos mostraron valores dentro de los rangos esperables en el INFIT para todos los años. En cuanto a los índices de confiabilidad arrojados por el modelo, la confiabilidad de los ítems fue alta 0.99 en todos los casos; no obstante, el índice para el caso de las personas fue muy bajo al inicio (0.51 en el 2010), pero conforme se fueron tomando en cuenta las recomendaciones sugeridas por el análisis de la información como elaborar más reactivos tendientes a medir niveles de habilidad más bajos el índice de confiabilidad mejoró notablemente hasta alcanzar un valor de 0.78 para el año 2012.

Precisamente los resultados de las pruebas del 2010 y 2011, en particular la del 2010, evidenciaron que los inconvenientes de la PDM no se debían a la elaboración de los ítems, sino, por una parte al nivel de dificultad de dicha prueba que resultó muy elevado para la población estudiantil que responde la prueba y por otra a la baja confiabilidad de la prueba. Una posible explicación para los bajos niveles de confiabilidad de los primeros años es la falta de preguntas que permitieran medir con mayor grado de exactitud los niveles de habilidad bajos, donde se encuentra buena parte del estudiantado que rinde la prueba, contrario a los reactivos que miden niveles altos de la prueba que son más abundantes, pero donde muy pocos estudiantes se ubican.

Ahora bien, en una prueba de diagnóstico lo ideal es contar con ítems en todos los niveles de la habilidad, para poder tener una precisión adecuada en niveles bajos, intermedios y altos del constructo y de esta manera medir con mayor certeza el nivel de conocimiento matemático de los estudiantes, para posteriormente brindar recomendaciones.

El modelo de Rasch resulta de gran utilidad a este propósito al poder situar en una misma escala el nivel de habilidad de las personas y el nivel de dificultad de los ítems, permite realizar este tipo de comparaciones. Gracias a ello se propuso a la comisión que elabora la PDM la inclusión de más ítems que midieran niveles bajos de habilidad y la implementación de este tipo de medidas dio como resultado mejores indicadores de confiabilidad.

Por otra parte, el hecho de contar con ítems concentrados en niveles altos de habilidad, permitió determinar aquellas temáticas que presetan una mayor dificultad para los y las estudiantes, además de brindar una idea de cuáles serían los tópicos más apropiados para incluir los reactivos necesarios para medir los niveles bajos. Es así que se tomó la decisión, a partir de la prueba del 2012 de aumentar el número de ítems de 40 a 60 e incorporar preguntas relacionadas con aritmética más propias de niveles como sétimo y octavo año de secundaria, pues tradicionalmente la comisión prestaba más atención a la elaboración de preguntas similares a las pruebas de bachillerato cuyos contenidos se abarcan en los niveles superiores del colegio, principalmente décimo y undécimo año.

Hasta el momento y pese a los esfuerzos realizados por las autoridades de la Escuela de Matemática, la PDM no se ha utilizado como criterio sustantivo para tomar decisiones respecto del rendimiento académico de los y las estudiantes; sin embargo, se pretende que la

prueba sea un parámetro para brindar recomendaciones en cuanto a la necesidad de asistencia a tutorías, talleres, grupos de estudio y cursos de reforzamiento para la población estudiantil que evidencie un bajo nivel de conocimientos matemáticos y para aquellos y aquellas que muestren un alto desempeño en la prueba, la posibilidad de rendir pruebas de suficiencia.

Otro posible uso es la construcción de una escala de niveles de desempeño, en la que puedan identificarse para cada nivel las áreas de conocimiento que muestran un bajo dominio y a partir de esa información brindar ayuda al estudiantado en las áreas que más lo requieran.

Finalmente, aunque son muchos los factores que intervienen en la debida construcción de una prueba diagnóstica más allá del cuidado en la selección y redacción de los reactivos que la componen, la oportunidad de contar con metodologías apropiadas para el análisis de la información de pruebas como lo es el modelo de Rasch o los modelos de dos o tres parámetros, donde además de la dificultad de los ítems se incorporan el parámetro de discriminación y acierto al azar, respectivamente. No obstante, el modelo de Rasch tiene la ventaja que es más fácil de interpretar y permite mejorar con el paso del tiempo la calidad técnica de las pruebas en beneficio de la UNA y del estudiantado en general.

Referencias

Aragón, L.E. Fundamentos psicométricos en la evaluación psicológica. *Revista Electrónica de Psicología Iztacala, 7*(4), 23-43. Recuperado de http://www.ojs.unam.mx/index.php/repi/article/view/21668

Arginay, J. C. (2006). Técnicas psicométricas. Cuestiones de validez y confiabilidad. S*ubjetividad y Procesos Cognitivos, 8*. Recuperado de http://dspace.uces.edu.ar:8180/dspace/handle/123456789/765

Bond, T:G., & Fox, C.M. (2001). *Applying the Rasch model: Fundamental measurement in the human sciences.* New Jersey, USA: Lawrence Erlbaum Associates, Publisher.

Cea D. Ancona, M. (2002). *Análisis multivariable.* Madrid: Editorial Síntesis, S.A.

Celina, H., & Campo, A. (2005). Aproximación al uso del coeficiente alfa de Cronbach. *Revista colombiana de psiquiatría, 34*(4). Recuperado de http://redalyc.uaemex.mx/pdf/806/80634409.pdf

Choppin, B. (1983). *The Rasch model for item analysis.* Recuperado de http://www.cse.ucla.edu/products/reports/r219.pdf

Jiménez, K. (2010). *Validación de la prueba de diagnóstico de conocimientos y destrezas en matemáticas del estudiante al ingresar a la universidad de la escuela de matemática de la UCR.* Tesis de maestría en Estadística, Universidad de Costa Rica.

Jornet, J.M., & Suárez, J.M. (1996). Pruebas estandarizadas y evaluación del rendimiento: Usos y características métricas. *Revista de Investigación Educativa, 14*(2), 141-163. Recuperado de http://www.uv.es/gem/archivos/RIE14.PDF

Kcuno, R. (2010). Examen de diagnóstico en matemática pretende mejorar rendimiento académico. *Acontecer.* Recuperado de http://web.uned.ac.cr/acontecer/index.php/a-diario/tecnologia/104-examen-diagnostico-de-matematicas-pretendemejorar-rendimiento-academico.html.

Montero, E. (2000). 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). Recuperado de http://revista.emate.ucr.ac.cr/index.php/revista/article/viewFile/101/92

Muñiz, J., Paz, M.D., Prieto,G., Delgado, A., Barbero, M., Arce, C.,…, Maydeu, A.(1997). *Psicometría.* Madrid, España: Universitas.

Muñiz, J., & Hambleton, R. (1992). Medio siglo de teoría de respuesta a los ítems. *Anuario de Psicología*, *52*. 41-66. Recuperado de http://www.raco.cat/index.php/AnuarioPsicologia/article/view/64681/88708

Nunnally, J., & Bernstein, I. (1995). *Teoría psicométrica.* México, D.F: McGraW-Hill.

Prieto G., & Delgado A. R. (2003). Análisis de un test

mediante el modelo de Rasch. *Psicothema, 15*(1). 94-100. Recuperado de http://www.psicothema.com/pdf/1029.pdf

Prieto, G., Velasco, A.D., Arias, R., Anido, M., Nuñez, A. N., & Có, P. (2007). Análisis de la dificultad de un banco de ítems de visualización espacial. *Ciencias Psicológicas, 1*(1), 71-79. Recuperado de http://pepsic.bvsalud.org/pdf/cpsi/v1n1/v1n1a07.pdf

Ramírez, G., & Barquero, J.A. (2011). Análisis de las pruebas de diagnóstico en matemática del instituto tecnológico de Costa Rica. Revista digital Matemática, *Educación e Internet*, *11*(2), 1-10 Recuperado de http://www.tec-digital.itcr.ac.cr/revistamatematica/ARTICULOS_V11_N2_2011/GRAMIREZJBARQUERO_V11N2_2011/GR_JB__V11N2_2011.pdf

Universidad de Costa Rica (2011). *Examen de diagnóstico.* Recuperado de http://diagnostico.emate.ucr.ac.cr

Wright, B.D., & Stone, M.H. (1998). *Diseño de mejores pruebas utilizando la técnica de Rasch*. México: CENEVAL.

Recibido: 20 de mayo de 2015
Aceptado: 16 de setiembre de 2015