Calibración de instrumentos de evaluación – clasificación en matemáticas en la Universidad Jorge Tadeo Lozano[1]

Calibration of assessment instruments – classification in mathematics at the Universidad Jorge Tadeo Lozano

Daniel BOGOYA, Sandra BARRAGÁN, Manuel CONTENTO y Adelina OCAÑA
Universidad de Bogotá Jorge Tadeo Lozano

Recibido: Abril 2013
Aceptado: Junio 2013

Resumen
La Teoría de Respuesta al Ítem, TRI, es una metodología valiosa para el análisis de la calidad de los instrumentos utilizados en evaluaciones de logro académico. En este artículo se presenta una implementación de la teoría mencionada, particularmente del modelo de Rasch, con el propósito de calibrar los ítems y el instrumento que se emplean en la prueba clasificatoria para la asignatura de Matemáticas Básicas en la Universidad Jorge Tadeo Lozano. Se analizaron 509 cadenas de respuestas de estudiantes, obtenidas en la aplicación de junio de 2011, con un conjunto de 45 ítems, a través de ocho casos de estudio que van mostrando progresivamente los pasos de la calibración. Se definieron y utilizaron criterios de validez de los ítems y del instrumento en su conjunto, para seleccionar los grupos de cadenas de respuestas y de ítems que finalmente fueron empleados en la determinación de los parámetros que permitieron luego la clasificación de los estudiantes evaluados mediante la prueba.
Palabas clave: calibración de instrumentos; evaluación; modelo de Rasch; teoría de respuesta al ítem.

Abstract
Item Response Theory, IRT, is a valuable methodological tool for analyzing the quality of the instruments used in the assessment of academic achievement. This article presents the implementation of the above mentioned theory, particularly Rasch model, to calibrate items and the instrument used in the classification test for the Basic Mathematics subject at Universidad Jorge Tadeo Lozano. 509 students chain responses from June 2011 applications were analysed. They make a total of 45 items and eight case studies that show progressively the calibration

[1] El presente artículo se elaboró dentro del proyecto de investigación: Elementos de evaluación en ciencias mediante la matemática y el lenguaje, código 402- 08-11, financiado por la Universidad de Bogotá Jorge Tadeo Lozano.

steps. Criteria to validate individual items and the research tool itself were defined and used, in order to select groups of chain responses and items that were finally used to determine the parameters to classify the students' performance in the test.
Keywords: Instruments calibration; assessment; Rasch model; item response theory.

La Universidad Jorge Tadeo Lozano se ha considerado una universidad formativa y ha plasmado en su modelo pedagógico la pretensión de *"…estar abierta a los distintos estratos sociales y a las distintas proveniencias culturales, y para ello no pone obstáculos en el ingreso y desarrolla programas de apoyo a los estudiantes para evitar la deserción."* (Universidad de Bogotá Jorge Tadeo Lozano, 2011a, pág. 59). En lo relativo a la fundamentación básica de los programas académicos, la Tadeo ha materializado esta intención ofreciendo la asignatura Matemáticas Básicas, como un primer curso que permita la nivelación de los estudiantes con mayores barreras de aprendizaje de los conceptos elementales de las matemáticas. De aquí que en la categoría de promoción y admisión (Swail, Redd, & Perna, 2003, pág. 91), de los programas de retención estudiantil implementados, se ha formulado una prueba[2] de clasificación para este primer curso, que opera desde el primer período lectivo[3] del año 2007 y que deben presentar los estudiantes que ingresan a primer semestre y quienes realizan transferencias internas o externas.

Si un estudiante aprueba el examen, aprueba el curso y es promovido a las siguientes asignaturas previstas en el plan de estudios. En caso contrario, el estudiante se inscribe en la materia para tomar un curso de 16 semanas, con una intensidad horaria semanal presencial de cuatro horas y no presencial de ocho horas. Las estrategias pedagógicas implementadas en este curso gravitan en torno a los dominios conceptuales requeridos para un conocimiento de los códigos elementales de las matemáticas, lo cual facilita la transición a la vida universitaria y brinda herramientas cognitivas para el tránsito hacia la autonomía.[4]

Con el compromiso permanente de una metaevaluación robusta y confiable, se propone estudiar la validez del instrumento que se emplea para evaluar el logro académico de quienes presentan la prueba mencionada. La evaluación es entendida como un acto que reconoce atributos y fortalezas, así como estados de desarrollo y profundidad de la capacidad en un campo particular de los estudiantes cuyo conocimiento es objeto de observación. Los resultados de esta clasificación constituyen

---

[2] Se toma como prueba el sistema de operaciones e instrumentos que sigue un protocolo, previamente diseñado y divulgado, que permite interrogar de forma similar a un grupo de estudiantes, con el fin de establecer su logro académico alrededor de un campo específico.

[3] La Universidad Jorge Tadeo Lozano ofrece tres períodos lectivos por año: primero, de febrero a mayo; segundo, de junio a julio; y tercero, de agosto a noviembre.

[4] "Es autónomo quien se orienta por sus propios valores y asume sus propias responsabilidades", es decir, aquel estudiante que toma sus decisiones en ejercicio de su capacidad para gobernarse y que genera hábitos de estudio para aprender toda la vida (Universidad de Bogotá Jorge Tadeo Lozano, 2011b, pág. 52).

una fuente de información importante para la universidad, pues los profesores pueden afinar el currículo que finalmente se despliega en el aula y orientar adecuadamente a los estudiantes, con prioridad y énfasis en las debilidades detectadas. A su vez este ejercicio ofrece elementos de juicio al estudiante evaluado, porque él mismo puede identificar los aspectos aún pendientes de ser abordados para satisfacer los objetivos del programa oficial de la asignatura.

El programa que materializa los dominios conceptuales previstos para la asignatura Matemáticas Básicas se publica en el sitio web de la Universidad con dos meses de anticipación a la realización de la prueba, de forma tal que los estudiantes conozcan la temática sobre la que serán evaluados. Considerando que la evaluación es un acto que reclama visibilidad (Bogoya, 2006, pág. 2), además de las características del instrumento también se publican las condiciones administrativas y logísticas previstas para la aplicación.[5]

Durante los años 2007 a 2009, con anterioridad a la aplicación de la prueba principal, se incluyó una versión de demostración para familiarizar a los estudiantes, con los tipos de preguntas y el sistema de evaluación. Para el tercer período lectivo de 2011, se estableció una nueva metodología de procesamiento de los datos, que incorpora la teoría de respuesta al ítem, TRI, con un parámetro (modelo de Rasch), como herramienta de análisis, antes de producir resultados de los estudiantes evaluados. La metodología prevé que después de consolidada la base de datos con las respuestas proporcionadas por los estudiantes, se procede a verificar los parámetros de los ítems y del instrumento en su conjunto, para luego generar y entregar resultados confiables sobre la clasificación de los estudiantes.

Método

Instrumento

Un dispositivo o constructo lógico, confiable y debidamente calibrado, conformado por un conjunto de ítems que aborda plenamente todo el espectro de los dominios conceptuales y cognitivos del campo bajo consideración, cuya finalidad es estimar el nivel de logro académico de un grupo de estudiantes en un campo específico, se denomina un instrumento. El conjunto de parámetros que caracterizan a los ítems seleccionados ilustran la calidad del instrumento. Cuando se emplea el modelo de Rasch, porque el conjunto de datos cumple con los supuestos establecidos, el único parámetro que se necesita para estimar la habilidad de los evaluados es la dificultad de cada ítem (Wright & Stone, 1998, pág. 111).

Para el diseño del instrumento que se utiliza en la prueba se tuvieron en cuenta tanto los dominios conceptuales, que reúnen los campos de las matemáticas y sus relaciones,

[5] Ver: http://www.utadeo.edu.co/es/noticia/novedades/departamento-de-ciencias-basicas/5121/examenes-de-matematicas-basicas-y

como los dominios cognitivos, que se refieren a los procesos mentales que deben desplegarse para la solución de los problemas planteados.

En el contenido programático del curso se consideran los dominios conceptuales: 1) números reales, 2) expresiones algebraicas y las operaciones entre ellas, 3) factorización de expresiones polinómicas y 4) ecuaciones y problemas de aplicación, con ecuaciones de primero y segundo grado en una variable y casos de proporcionalidad, como se encuentra en el plan curricular. Y los tres dominios cognitivos comprendidos son: 1) reconocimiento, 2) aplicación y 3) razonamiento[6]. Las categorías indicadas, que permiten acciones con los cuatro dominios conceptuales considerados, guardan relación con niveles crecientes de complejidad.

Bancos de ítems, bloques y cuadernillos

Los bancos, que constituyen un elemento fundamental de la TRI, deben contener ítems calibrados en la misma escala. La capacidad de un banco de ítems para determinar de manera válida la habilidad de un conjunto de evaluados, dependen de una apropiada cobertura en dominios conceptuales y cognitivos (Muñiz & Hambleton, 1992, pág. 53).

En la Tadeo se ha elaborado un banco de ítems que se alimenta y revisa permanentemente; los ítems que comprenden los dominios conceptuales y cognitivos indicados se encuentran ubicados en la plataforma TestInWeb. Se cuenta con 225 ítems con un enunciado cerrado y cuatro opciones de respuesta, con única respuesta correcta o válida. Los ítems que se emplean para conformar el instrumento se eligen procurando la mayor cobertura posible de los dominios mencionados, así como una distribución de dificultad cercana a la respectiva distribución de habilidad de la población que debe ser evaluada.

Los 45 ítems que conforman el instrumento se organizan en tres bloques completos, cada uno de 15 ítems, con los cuales se estructuran tres cuadernillos[7]; cada estudiante evaluado responde sólo un cuadernillo con 30 ítems. El cuadernillo 1 incluye los bloques 1 y 2; el cuadernillo 2, los bloques 2 y 3; y el cuadernillo 3, los bloques 3 y 1. De este modo, cada cuadernillo tiene un bloque común con otro cuadernillo, para establecer un lazo de unión y permitir la equiparación y comparabilidad de resultados.

Para la aplicación de la prueba, previa publicación de las instrucciones respectivas, en la plataforma TestInWeb se programan los tres cuadernillos con los dos bloques establecidos para cada uno, fecha y hora de presentación y duración de la prueba. Los estudiantes que cumplen las condiciones establecidas para presentar la prueba se incluyen en una base de datos ordenada alfabéticamente; el procedimiento prevé

---

[6] La estructura de los tres dominios cognitivos ha sido tomada del diseño de TIMSS (Trends in International Mathematics and Science Study) (Mullis, Martin, Ruddock, Sullivan, & Preuschoff, 2009, págs. 40-46).

[7] Un cuadernillo es un ejemplar impreso o virtual que comprende el número de bloques previsto en el diseño y que mantiene la misma distribución y presentación en todos los ejemplares de un mismo tipo.

asignar un cuadernillo (1, 2 o 3) a cada estudiante, que debe ser respondido el día de la aplicación. La prueba es presencial, se realiza en línea en salas de cómputo y se programa en dos horarios, uno diurno de 9 a.m. a 12 m. y otro nocturno de 6 a 8 p.m. Dos días después de presentada la prueba, una vez realizado el procesamiento de datos, se informa el resultado obtenido por los estudiantes evaluados.

Procesamiento de datos

Reconociendo que el procesamiento requerido puede realizarse mediante la Teoría Clásica o la Teoría de Respuesta al Ítem, TRI (Muñiz & Hambleton, 1992, págs. 42-47), se decidió como parte del diseño, implementar la TRI con un parámetro - modelo de Rasch - para llevar a cabo el procesamiento. Se tomó esta decisión considerando que mediante la aplicación de las etapas de precalibración y calibración de los datos se obtiene mayor precisión en la estimación del nivel de desempeño de los estudiantes evaluados así como del instrumento aplicado.

Al examinar los resultados obtenidos cuando se utilizan uno o tres parámetros de la TRI, no se encuentra una diferencia significativa entre la función de información del instrumento y la función de sensibilidad de cada ítem. De igual manera no hay diferencia en los valores de la función de información al utilizar modelos de uno o dos parámetros, lo que está de acuerdo con los estudios realizados por Hambleton (1977) y Friedrich (2004) (Moghadamzadeh, Salehi, & Khodaie, 2011, págs. 1359-1367). El procesamiento proporciona la estimación de la dificultad de cada ítem y la habilidad para cada evaluado (Wright & Stone, 1998, págs. 1-9), separando las características del evaluado de las del instrumento en sí. La dificultad de los ítems se define en la TRI en términos de la probabilidad de la respuesta correcta, no de la dificultad percibida o de la cantidad de esfuerzo requerido para contestarlos (DeMars, 2010, pág. 4). Por otra parte, la TRI permite que las estimaciones de los parámetros sean independientes de la población evaluada y que las estimaciones de las habilidades de los evaluados sean independientes de los ítems que componen la prueba (Hambleton, Swaminathan, & Rogers, 1991, pág. 5).

La TRI ha sido implementada en proyectos internacionales como se referencia en TIMSS (Olson, Martin, & Mullis, 2008, págs. 249,256), en PISA (OECD, 2009, págs. 22, 147), y en estudios a nivel local como los realizados por Abdullah et al (Abdullah, Arsad, Hashim, Aziz, Amin, & Ali, 2012, págs. 119-123) y por Osman et al (Osman, Naam, Jaafar, Badaruzzaman, & Rahmat, 2012, págs. 59-66), en los que se utiliza el modelo de Rasch. En el primer estudio los autores analizan el desempeño de los estudiantes que presentan la prueba final de Microelectrónica, en función del nivel de dificultad de cada ítem. Dicho estudio concluye que estos análisis pueden ayudar a observar la habilidad de cada estudiante, para responder cada ítem. De igual manera, el nivel de dificultad de cada pregunta diseñada puede ser evaluado y mejorado continuamente para futuras pruebas. El segundo estudio muestra que con el modelo de Rasch puede clasificarse y tabularse de manera efectiva a los estudiantes y a los ítems por medio de un mapa de distribución de la habilidad de los estudiantes y de dificultad de los ítems, en una misma escala. El estudio demostró que los resultados obtenidos

por los estudiantes pueden ser medidos utilizando el modelo de Rasch; en este análisis, los estudiantes fueron clasificados de acuerdo a sus logros que reflejaban su habilidad de aprendizaje en este curso. El modelo de Rasch produce un patrón de asociación entre los estudiantes y el nivel de desempeño, lo que lo convierte, según el estudio, en el mejor modelo de evaluación para medir el desempeño en los resultados del curso.

De otra parte, para aplicar la TRI, se requiere que los instrumentos que se utilizan en los procesos de evaluación sean previamente validados, a la luz de la confiabilidad[8], consistencia interna[9] y unidimensionalidad[10]. La validez de *constructo* se demuestra a través de la prevalencia de una dimensión sobre las demás, mientras que la independencia local se comprueba con el valor de correlación inter-ítem[11].

El modelo de Rasch estima la probabilidad p con la que un estudiante evaluado responde correctamente un ítem. Esta probabilidad, frente a ítems con una única opción de respuesta correcta, es una función exponencial de la distancia entre el valor de la habilidad del estudiante y la dificultad del ítem (Ecuación 1).

$$p\{X_{vi} = 1|\beta_v, \delta_i\} = \frac{e^{\beta v - \delta_i}}{1+e^{\beta v - \delta_i}} \quad . \qquad (1)$$

Donde el parámetro β_v representa la habilidad del estudiante v $(v = 1, 2, \dots, n)$; y δ_i denota la dificultad del ítem i $(i = 1, 2, \dots, m)$ (Wright & Stone, 1998, pág. 15).

Una vez realizada la aplicación de la prueba a una cierta población, se consolida un archivo de datos, con un registro para cada cadena de respuestas dada por cada estudiante de la población. El procesamiento se realiza en tres etapas consecutivas: una primera etapa de precalibración, en la cual se detectan y retiran los registros cuyo patrón de respuestas es anómalo, ya sea porque presentan una omisión mayor a 2, un valor negativo de correlación o porque no satisfacen el criterio de ajuste[12] con el modelo; una segunda etapa de calibración, en la que se hace el análisis estadístico con

---

[8] La confiabilidad está referida a la reproducibilidad de la localización relativa de la medida. Así, una alta confiabilidad de ítems (o personas) significa que existe una alta probabilidad de que ítems (o personas) que fueron estimados con alta dificultad (o alta habilidad para personas) realmente tengan una mayor medida que las estimaciones de ítems (o personas) con menor dificultad (o habilidad para personas) (Linacre, 2008, pág. 440).

[9] La consistencia interna es el camino más habitual para estimar la confiabilidad de las medidas y dentro de esta categoría el coeficiente alfa de Cronbach estima el límite inferior del coeficiente de confiabilidad (Linacre, 2008, págs. 235, 441).

[10] La unidimensionalidad es uno de los supuestos más críticos y fundamentales de la teoría de la medición. Principalmente, enfatiza en que los ítems que conforman un instrumento deben contribuir conjuntamente a medir algo en común, es decir que solo un rasgo latente o constructo se encuentra en la base del conjunto de ítems. Para determinar si un instrumento es unidimensional, mediante la metodología del análisis factorial, se tiene en cuenta la proporción de la varianza total explicada por las medidas y la explicada por el contraste con el primer factor extraído (Linacre, 2008, págs. 376-377).

[11] En el procesamiento de los datos se emplea el parámetro de correlación en dos perspectivas: entre parejas de ítems (correlación inter-ítem) y entre un ítem y el conjunto (correlación ítem-prueba).

[12] El valor del parámetro de ajuste se estima tanto para los ítems como para los estudiantes.

la información proveniente de los registros seleccionados en la precalibración, con el fin de retirar los ítems que no satisfagan los criterios establecidos y luego estimar el valor de los parámetros de los ítems admitidos y del instrumento; y una última etapa, de *calificación*, en la que se retoman de nuevo todos los registros iniciales y se asigna un valor de habilidad a cada estudiante evaluado, utilizando la escala de dificultad establecida en la etapa anterior.

El procesamiento de los datos se realiza con el programa WINSTEPS 3.73, desarrollado por el profesor John Linacre de la Universidad de Chicago. También se emplea el software ConQuest, desarrollado por el Australian Council for Educational Research, ACER, con el fin de verificar el valor encontrado para los parámetros utilizados. Finalmente, los datos que resultan del procesamiento se exportan a una base de datos para elaborar los reportes correspondientes.

Como resultado del procesamiento, se genera un reporte que se entrega a los profesores encargados de la asignatura Matemáticas Básicas, para orientar los ajustes y la organización de las prácticas de aula más apropiadas, que conduzcan al desarrollo de los dominios conceptuales observados con mayor debilidad.

Primera etapa: precalibración

Al aplicar un instrumento se encuentra una gran diversidad de condiciones y de actitudes por parte de quienes responden las preguntas, lo cual genera a su vez distintos patrones en las cadenas de respuestas. Hay casos de estudiantes que responden al azar, otros de manera sistemática, algunos en forma correcta sólo ciertas preguntas y también quienes abandonan la prueba; estas y otras tipologías de cadenas de respuestas conducen a *patrones anómalos*, que inciden sustancialmente en el valor de los parámetros que se estiman. Por lo anterior, es indispensable disponer de indicadores que hagan posible la detección de tales patrones. La etapa de precalibración (ver Gráfico 1) consiste en procesar todas las cadenas de respuestas de los estudiantes evaluados, con el fin de detectar las cadenas aptas para llevar a cabo la calibración, con base en el número de omisiones y el valor calculado para los parámetros de correlación y de ajuste próximo y ajuste lejano.

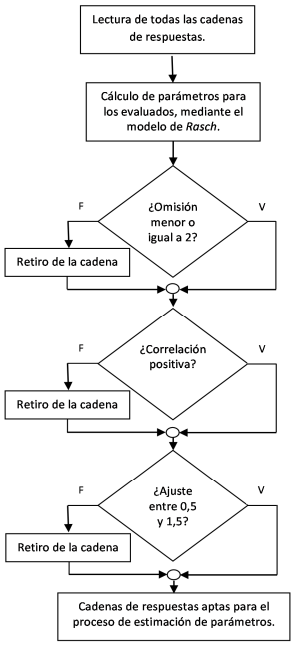

Gráfico 1. Momentos de la precalibración.

Las cadenas de respuestas con alta omisión no ofrecen información relevante sobre la habilidad β del evaluado; los ítems que fueron leídos pero que el estudiante decidió no contestar son considerados como ítems omitidos, lo que implica que no pueden catalogarse como respuestas erróneas. Es así que se incluyen en la base de datos las cadenas de respuesta que tengan en cualquier posición, máximo dos respuestas omitidas. El parámetro de correlación, calculado para cada estudiante, es útil para evaluar el grado de asociación entre la estructura de su cadena de respuestas y la tendencia global del conjunto de cadenas de respuestas de toda la población, conjunto que genera una escala de dificultades de los ítems del instrumento aplicado. Se espera que un estudiante responda en forma correcta, con una probabilidad mayor que 0,50, aquellos ítems cuya dificultad es menor que la habilidad de este estudiante; el mismo estudiante debe responder en forma no correcta, también con una probabilidad mayor que 0,50, los ítems cuya dificultad es mayor que la habilidad referida. La correlación calculada para un estudiante tendrá un valor negativo cuando su cadena de respuestas infringe la estructura esperada, es decir, cuando dicha cadena presenta una proporción de respuestas correctas mayor que 0,50 en la zona de ítems con dificultad mayor que la habilidad, a expensas de una proporción de respuestas correctas menor que 0,50 en la zona de ítems con dificultad menor que la habilidad (Wright & Stone, 1998, págs. 12 - 13).

La correlación indicada puede clarificarse utilizando el escalograma de *Guttman* (ver Tabla I, para el cuadernillo 1), que corresponde a una matriz de $n \times m$ (n y m definidas en la Ecuación 1), en cuyas columnas se localizan los ítems, ordenados de acuerdo con el valor de la dificultad[13] (la primera columna con el ítem de menor dificultad y la última con el ítem de mayor dificultad), mientras que en las filas se disponen las cadenas de respuestas de los estudiantes evaluados, ordenadas según el valor de la habilidad (la primera fila con la cadena del estudiante de mayor habilidad y la última con la cadena del estudiante de menor habilidad).

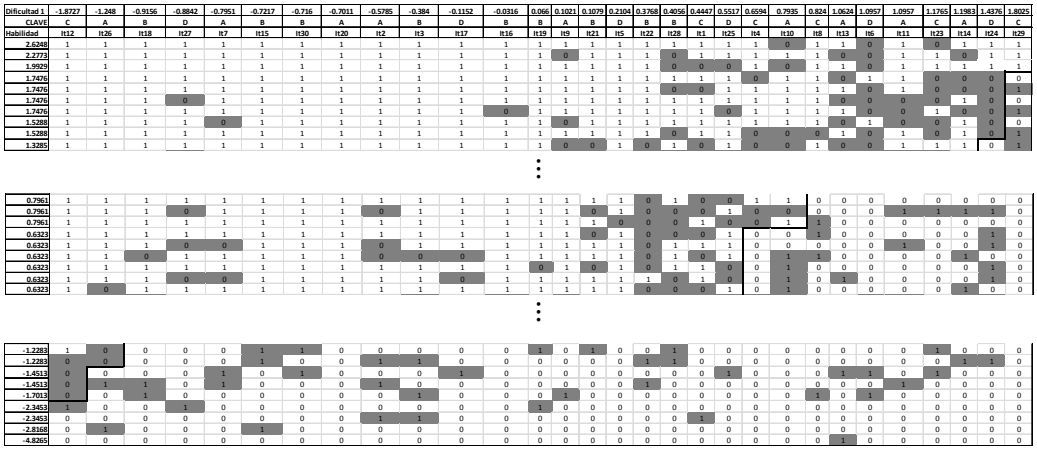

Tabla I. Escalograma de *Guttman* para el primer cuadernillo con 174 estudiantes.

En cada posición de la matriz se asigna el valor igual a uno, si la respuesta es correcta, y cero en caso contrario, para llegar a un arreglo de dos triángulos separados: uno localizado en la zona superior izquierda, con alta densidad de valores iguales a uno; y el otro en la zona inferior derecha, con alta prevalencia de valores iguales a cero (Linacre, 2008, pág. 278).

Las cadenas de respuestas con correlación negativa se reflejan en el escalograma de *Guttman* cuando prevalecen valores iguales a uno en el triángulo de la zona inferior derecha y valores iguales a cero en el triángulo de la zona superior izquierda. Si el valor del parámetro de correlación es negativo, se considera que el patrón de respuestas está en disonancia con la tendencia mostrada por el conjunto. Para el procesamiento de los datos sólo se consideran las cadenas de respuestas cuyo valor del parámetro de correlación es positivo.

Luego de aplicar el criterio de correlación se examinan los valores de los parámetros de ajuste, tanto próximo como lejano; ambos indicadores reflejan la adecuada correspondencia entre las respuestas dadas por un estudiante a los ítems del instrumento, y los resultados esperados con el modelo de Rasch. Para el procesamiento de los datos sólo se utilizan las cadenas de respuestas cuyo valor del parámetro de ajuste, tanto próximo como lejano, está comprendido entre 0,5 y 1,5 (Linacre, 2008,

[13] La unidad utilizada para la dificultad y la habilidad es el logito.

pág. 249), considerando un valor ideal de uno, que indica un ajuste perfecto entre la cadena de respuestas y el modelo.

La etapa de precalibración concluye con la definición del conjunto de cadenas de respuestas que presentan máximo dos respuestas omitidas, valores positivos para el parámetro de correlación y que cumplen con el criterio previsto para los parámetros de ajuste, pues aportan la mayor cantidad de información y conducen a una mayor consistencia interna.

Segunda etapa: calibración

El objetivo que persigue la calibración es verificar el cabal desempeño estadístico de los ítems y del instrumento en su conjunto, así como fijar un valor para los parámetros de dificultad, error de estimación, correlación ítem-prueba, ajuste próximo y lejano y discriminación, de cada uno de los ítems que finalmente son admitidos. Aquellos ítems cuyos parámetros presentan valores que no satisfacen los criterios establecidos se excluyen del procesamiento. La Tabla II muestra los valores ideales y los criterios definidos para admitir los ítems que se consideran en la calibración.

Parámetro	Valor ideal	Valor para aprobación
Correlación ítem-prueba	Entre 0,10 y 0,60	Entre 0,10 y 0,60
Ajuste próximo y lejano	1,00	Mayor que 0,60 y menor que 1,40
Discriminación	1,00	Mayor que 0,20 y menor que 1,80
Error de la estimación	Menor que 0,01	Menor que 0,18[14]

Tabla II. Valores ideales y criterios para admitir ítems.

El ejercicio de calibración comprende un proceso iterativo donde se estiman los valores de los parámetros de los ítems, con base en las cadenas de respuestas que satisfacen los criterios indicados; después se retira el ítem con indicador estadístico más débil y luego se repite el proceso hasta satisfacer plenamente los criterios de validez establecidos tanto para ítems como para el instrumento.

En relación con los parámetros del instrumento, se consideran la confiabilidad; consistencia interna, tomada como el valor del coeficiente *Alfa de Cronbach;* dimensionalidad, mediante la relación entre el porcentaje de varianza explicada por las medidas y el porcentaje de varianza que explica el contraste con el primer componente, la cual conduce a mostrar la validez de constructo; e independencia local, a través de la correlación inter-ítem de cada uno de los subconjuntos diferentes de dos ítems. La Tabla III presenta el criterio para los indicadores estadísticos deseables en un instrumento.

[14] Se define este valor de error de estimación debido al tamaño relativamente pequeño de la población de estudiantes evaluados. A mayor número de registros, menor valor del error.

Parámetro	Valor de aprobación
Confiabilidad	Mayor que 0,60
Consistencia interna	Mayor que 0,60
Relación de varianzas	Mayor que 4,00
Correlación inter-ítem	Menor que 0,50

Tabla III. Criterio para los indicadores estadísticos del instrumento.

La etapa de calibración termina con la obtención de un conjunto de valores satisfactorios para los parámetros, tanto de los ítems que finalmente fueron admitidos como del instrumento en su conjunto, y con una escala donde se estiman los valores de dificultad de los ítems. (Ver Tablas VI, VII y VIII).

Tercera etapa: calificación

Se procede ahora a calcular un valor de habilidad para cada uno de los estudiantes evaluados, operando de nuevo con la totalidad de los registros provenientes de la aplicación del instrumento y fijando el valor del parámetro de dificultad de los ítems que fue hallado en la etapa de calibración. A cada estudiante se le asigna un valor de habilidad, que indica su localización ordinal relativa dentro de la escala (de habilidad) y que refleja su desempeño frente a los ítems admitidos y utilizados en la calibración. La escala de habilidad constituye una función de distribución con sus respectivos parámetros de posición y dispersión: como parámetro de posición suele utilizarse la media, con valor igual a cero; y como parámetro de dispersión, la desviación estándar, con valor igual a uno. El valor de habilidad puede convertirse en un puntaje, que también obedece a una escala y una función de distribución (Wright & Stone, 1998, pág. 111), con determinados valores para los parámetros de posición y dispersión, de acuerdo con la transformación indicada por la Ecuación 2[15].

$$x_v = (\beta_v - \beta) * \left(\frac{\sigma}{\sigma_o}\right) + \mu \qquad (2)$$

Donde x_v y β_v corresponden al valor del puntaje y la habilidad estimada del estudiante v $(v = 1, 2, \dots, n)$; β representa el promedio de habilidad de la población de estudiantes; σ y σ_o significan la desviación estándar prevista para la distribución de puntajes y la distribución de habilidad de la población de estudiantes; y μ indica la media prevista para la escala de puntajes.

La calificación también comprende la asignación de un nivel de desempeño para cada uno de los estudiantes, de acuerdo con la siguiente regla: un estudiante v

[15] Asumiendo que la distribución de las habilidades estimadas es logística (con promedio β y desviación σ_0) y que la asignación del puntaje se hace mediante una transformación cuya imagen también es logística (con promedio μ y desviación σ), se busca que las estandarizaciones sobre la habilidad y el puntaje correspondan, es decir, $\frac{\beta_v - \beta}{\sigma_0} = \frac{x_v - \mu}{\sigma}$ lo que implica (2).

pertenece al nivel q, si su habilidad β_v es superior o igual a la habilidad β_q que corresponde con una probabilidad de respuesta correcta de al menos 0,60 para el ítem de menor dificultad del referido nivel q.

Resultados

Partiendo de la población total de 509 estudiantes que respondieron los ítems del instrumento, se trabajaron ocho casos de estudio (ver Tabla IV) siguiendo distintos criterios para seleccionar las cadenas de respuestas procesadas y cuyos valores de los parámetros estimados para el instrumento se muestran en la Tabla V: confiabilidad, separación, consistencia interna (coeficiente *Alfa de Cronbach*), varianza explicada por las medidas y en el contraste con el primer componente y el intervalo de variación de la correlación inter-ítem entre subgrupos de dos ítems. En los casos 1 a 6 se consideraron los 45 ítems propuestos, mientras que en los casos 7 y 8 se retiró el ítem 6, debido a que no satisfizo los criterios de validez establecidos.

Caso de estudio	Registros iniciales	Número de ítems	Criterio			Número de registros procesados
			Omisión	Correlación	Ajuste	
1	509	45				509
2	509	45	≤ 2	Positiva		455
3	509	45	≤ 2	Positiva	Entre 0,4 y 1,6	452
4	509	45	≤ 2	Positiva	Entre 0,5 y 1,5	452
5	509	45	≤ 2	Positiva	Entre 0,6 y 1,4	437
6	509	45	≤ 2	Positiva	Entre 0,7 y 1,3	404
7	509	44				509
8	509	44	≤ 2	Positiva	Entre 0,5 y 1,5	451

Tabla IV. Casos de estudio según criterio de selección de cadenas de respuestas

Caso de estudio	Confiabilidad	Separación	Coeficiente *Alfa de Cronbach*	Porcentaje de varianza explicada		Intervalo de variación de la correlación entre dos ítems
				Por las Medidas	Contraste con el primer componente	
1	0,79	1,97	0,66	23,2	3,5	-0,02 a 0,18
2	0,78	1,88	0,64	24,2	3,3	-0,19 a 0,17
3	0,78	1,88	0,63	24,2	3,3	-0,19 a 0,16
4	0,78	1,88	0,63	24,2	3,3	-0,19 a 0,16
5	0,75	1,74	0,60	23,3	3,4	-0,19 a 0,15
6	0,73	1,64	0,56	22,3	3,4	-0,21 a 0,15
7	0,80	1,98	0,67	23,6	3,5	-0,19 a 0,18
8	0,78	1,89	0,64	24,3	3,3	-0,19 a 0,16

Tabla V. Valores estimados de parámetros para el instrumento en los ocho casos de estudio.

Manteniendo los 45 ítems propuestos, la confiabilidad calculada con las 509 cadenas de respuestas fue igual a 0,79, valor que disminuye hasta 0,73 cuando se utilizan 404 cadenas de respuestas con omisión menor o igual a dos, correlación positiva y con valor del parámetro de ajuste entre 0,7 y 1,3. Para los mismos casos de análisis, el valor del coeficiente *Alfa de Cronbach* decae, desde 0,66 hasta 0,56, y la relación entre los porcentajes de varianza explicada las medidas y el contraste con el primer componente, desde 6,63 hasta 6,56.

Concluido el análisis se decidió utilizar los parámetros resultantes de procesar el conjunto de 451 cadenas de respuestas con omisión menor o igual a dos, correlación positiva, valor del parámetro de ajuste entre 0,50 y 1,50, excluyendo el ítem 6, porque en este caso de estudio 8 se evidencia una combinación aceptable de valores para los parámetros: confiabilidad (0,78); separación (1,89); consistencia interna (0,64); relación de porcentajes de varianza explicada por las medidas y por el contraste con el primer componente ($24{,}3/3{,}3 = 7{,}36$), confirmando así la validez de constructo[16] (Linacre, 2008, pág. 376); y correlación inter-ítem, cuyos valores fluctúan desde -0,19 (para los ítems 1 y 5) hasta 0,16 (para los ítems 17 y 18), demostrando la condición de independencia local.

La dificultad de los 44 ítems admitidos varía desde -2,11 logitos (para el ítem 12, que resultó ser el más fácil, con 86,18% de respuestas correctas), hasta 2,07 logitos (para el ítem 29, el más difícil, con 14,81% de respuestas correctas), mientras que la

[16] Al procesar las cadenas de respuestas seleccionadas en el caso de estudio 8, considerando por separado cada uno de los dominios conceptuales, la relación entre varianzas para los dominios referidos oscila entre 7,05 y 7,42. De otra parte, si se procesan todos los ítems, como un solo constructo, se obtiene una relación entre varianzas igual a 7,36. Lo anterior confirma la prevalencia de una dimensión en el conjunto de los 45 ítems que conforman el instrumento y el cumplimiento del supuesto de unidimensionalidad.

habilidad de los 451 estudiantes cuyas cadenas de respuestas dieron origen a la calibración oscila entre -2,93 logitos (para el estudiante de menor desempeño, con sólo una respuesta correcta) y 3,74 logitos (para el estudiante más destacado, con 29 respuestas correctas para los 30 ítems respondidos). Las Tablas VI, VII y VIII presentan los valores encontrados para los parámetros de los ítems admitidos en la calibración, correspondientes al caso de estudio 8, agrupados para los bloques 1, 2 y 3, respectivamente.

Ítem	Respuestas correctas (%)	Dificultad	Error	Ajuste próximo	Ajuste lejano	Correlación	Discriminación
01	40,13	0,43	0,13	1,13	1,11	0,24	0,61
02	63,49	-0,67	0,13	1,05	1,06	0,29	0,84
03	59,21	-0,46	0,13	0,88	0,83	0,49	1,45
04	34,21	0,72	0,13	1,20	1,23	0,16	0,51
05	45,72	0,17	0,12	0,87	0,82	0,50	1,52
07	68,42	-0,92	0,13	0,90	0,85	0,44	1,24
08	32,24	0,83	0,13	0,89	0,89	0,47	1,24
09	48,68	0,03	0,12	1,06	1,10	0,29	0,71
10	32,89	0,79	0,13	1,02	0,99	0,34	0,97
11	28,29	1,04	0,14	1,08	1,15	0,26	0,84
12	86,18	-2,11	0,17	0,99	0,84	0,31	1,03
13	26,97	1,12	0,14	0,97	0,94	0,38	1,06
14	24,01	1,30	0,14	1,07	1,13	0,26	0,88
15	67,11	-0,85	0,13	0,93	0,87	0,42	1,19

Tabla VI. Parámetros estimados para 14 ítems admitidos del bloque 1 del instrumento.

Ítem	Respuestas correctas (%)	Dificultad	Error	Ajuste próximo	Ajuste lejano	Correlación	Discriminación
16	52,86	-0,12	0,13	0,98	0,99	0,39	1,05
17	54,55	-0,20	0,13	0,95	0,95	0,42	1,18
18	70,37	-0,98	0,14	0,98	0,87	0,38	1,09
19	50,17	0,01	0,13	0,91	0,89	0,47	1,33
20	68,01	-0,86	0,13	0,96	0,97	0,38	1,10
21	49,49	0,04	0,13	0,92	0,91	0,46	1,29
22	42,09	0,39	0,13	1,16	1,16	0,23	0,51
23	25,25	1,30	0,14	1,11	1,23	0,24	0,81
24	20,88	1,59	0,15	1,14	1,24	0,21	0,82

25	38,72	0,56	0,13	0,98	0,97	0,40	1,05
26	79,12	-1,51	0,15	0,91	0,81	0,40	1,13
27	74,07	-1,19	0,14	1,08	1,19	0,23	0,83
28	43,77	0,31	0,13	1,07	1,08	0,32	0,78
29	14,81	2,07	0,17	0,94	1,00	0,36	1,03
30	69,36	-0,93	0,13	0,87	0,77	0,49	1,33

Tabla VII. Parámetros estimados para 15 ítems admitidos del bloque 2 del instrumento.

Ítem	Respuestas correctas (%)	Dificultad	Error	Ajuste próximo	Ajuste lejano	Correlación	Discriminación
31	33,89	0,75	0,13	1,09	1,07	0,31	0,82
32	66,78	-0,85	0,13	0,91	0,84	0,45	1,24
33	43,85	0,26	0,13	0,85	0,80	0,54	1,51
34	49,17	0,00	0,13	1,10	1,15	0,29	0,62
35	38,54	0,52	0,13	0,82	0,80	0,56	1,48
36	25,25	1,25	0,14	1,22	1,28	0,16	0,67
37	42,52	0,32	0,13	1,22	1,32	0,17	0,26
38	58,14	-0,42	0,13	0,90	0,86	0,47	1,35
39	37,54	0,57	0,13	0,95	0,95	0,44	1,13
40	62,46	-0,63	0,13	0,99	0,93	0,39	1,07
41	49,50	-0,01	0,13	1,10	1,21	0,27	0,57
42	20,93	1,53	0,15	0,95	0,91	0,41	1,06
43	58,80	-0,45	0,13	1,05	1,11	0,32	0,80
44	61,79	-0,60	0,13	0,95	0,91	0,42	1,17
45	39,87	0,45	0,13	0,89	0,91	0,49	1,29

Tabla VIII. Parámetros estimados para 15 ítems admitidos del bloque 3 del instrumento.

Discusión y conclusión

El instrumento diseñado, elaborado y aplicado en la Universidad de Bogotá Jorge Tadeo Lozano, para la clasificación en la asignatura Matemáticas Básicas, estaba conformado por 45 ítems que se agruparon en tres bloques de 15 ítems cada uno; cada estudiante responde un subconjunto de 30 ítems organizados en un cuadernillo que comprende dos de los tres bloques previstos. La técnica de bloques permitió una cobertura mayor de dominios conceptuales y cognitivos y el diseño en espiral permitió comparar los resultados alcanzados por estudiantes que respondieron cuadernillos distintos; Los Ítems que conformaron el cuadernillo 1 tienen dificultad promedio 0,06

y desviación estándar 1,00, para el cuadernillo 2 el promedio fue 1,10 y desviación estándar 0,87 y finalmente para el cuadernillo 3, el promedio fue 0,14 y la desviación estándar 0,83.

Entre los ocho casos de estudio que fueron construidos y analizados, se optó por el caso 8, retirando el ítem 6 y procesando las 451 cadenas de respuestas cuyo valor del parámetro de ajuste se localizó entre 0,50 y 1,50, mostrando un desempeño estadístico satisfactorio para el instrumento a pesar del reducido grupo de estudiantes a quienes se aplicó: confiabilidad igual a 0,78; separación igual a 1,89; consistencia interna igual a 0,64; relación entre la varianza explicada por las medidas y la varianza explicada por el contraste con el primer componente igual a 7,36; e intervalo de variación de la correlación inter-ítem entre -0,19 y 0,16.

Por su parte, también se estimaron los indicadores del desempeño estadístico de los 44 ítems finalmente admitidos: la dificultad que oscila entre -2,11 y 2,07 logitos, con un error de estimación que varía desde 0,12 hasta 0,17; el ajuste, entre 0,77 y 1,32; la correlación ítem-prueba, entre 0,16 y 0,56; y el parámetro de discriminación, entre 0,26 y 1,52. La dificultad presentó un rango de variación un tanto más estrecho de lo esperado evidenciando posiblemente redundancia en algunas de las preguntas; el error de estimación fue grande en magnitud, ocasionado por el relativo bajo número de estudiantes evaluados. Aunque se reporta el valor de la discriminación (segundo parámetro de la TRI) y se utiliza como criterio de admisión de los ítems, en el modelo de Rasch empleado para el procesamiento de los datos sólo se incluye el parámetro de dificultad.

La habilidad estimada para los estudiantes evaluados osciló entre 3,7451 y -2,9332 y el error estándar de las estimaciones de las medidas de habilidad entre 0,3943 y 1,0262, aclarando que cuando se usa TRI, la precisión de las estimaciones no es igual para cada conjunto de ítems y de personas, lo cual se puede apreciar en el Gráfico 2, en donde se observa que el instrumento de Matemáticas Básicas que se aplicó en la Universidad Jorge Tadeo Lozano estima con mayor precisión en el centro de la escala de habilidad. Para incrementar la precisión en la estimación de la habilidad, en el sector de mayor habilidad (alrededor de cuatro logitos), es necesario incluir ítems de mayor dificultad.

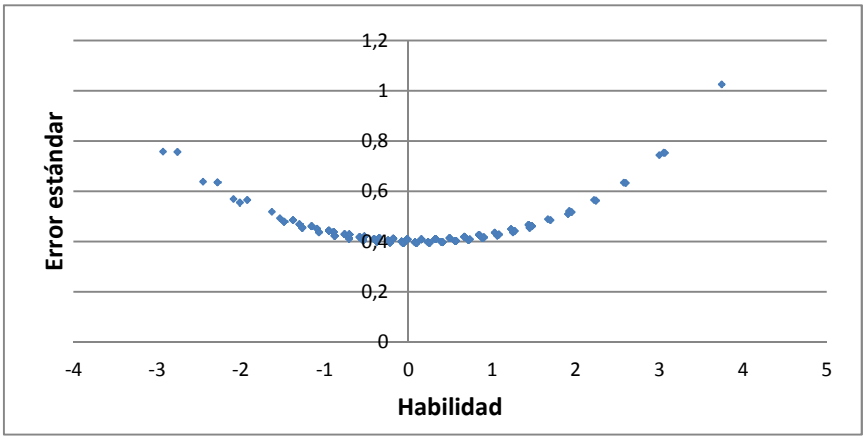

Gráfico 2. Errores estándar en la estimación de las medidas según el nivel de habilidad.

Finalmente se concluye que el modelo de Rasch de la teoría de respuesta al ítem con un parámetro, el de la dificultad, es una herramienta valiosa para procesar los datos del examen clasificatorio de Matemáticas Básicas porque permitió el análisis conjunto de los parámetros de los ítems y del instrumento, y a partir de la estimación de estos parámetros se proporcionó un diagnóstico individual para los evaluados que ofrece información sobre el nivel de conocimiento del campo para ellos mismos y para la institución, redundando en la optimización de los recursos y de los esfuerzos académicos.

Referencias bibliográficas

ABDULLAH, H., ARSAD, N., HASHIM, F. H., AZIZ, N. A., AMIN, N., & ALI, S. H. (2012). Evaluation of Students' Achievement in the Final Exam Questions for Microelectronic (KKKL3054) using the Rasch Model. *Procedia Social and Behavioral Sciences*, 119-123.

BOGOYA, D. (2006). Evaluación Educativa en Colombia. *Seminario Internacional de Evaluación* (págs. N1-N27). Cartagena: ICFES.

DeMARS, C. (2010). *Item Response Theory (Understanding Statistics: Measurement).* New York: Oxford University Press, USA.

HAMBLETON, R., SWAMINATHAN, H., & Rogers, J. (1991). *Fundamentals of item response theory.* Newbury Park: SAGE Publications, Inc.

LINACRE, J. M. (2008). *A User's Guide to Winsteps.* Chicago: John M. Linacre.

MOGHADAMZADEH, A., SALEHI, K., & KHODAIE, E. (2011). A comparison the Information Function of the Item and Test in One, Two and Three Parametric

Model of the Item Response Theory (IRT). *Procedia, Social and Behavioral Sciences*, 1359-1367.

MulLis, I., MARTIN, M., RUDDOCK, G., SULLIVAN, C., & PREUSCHOFF, C. (2009). *Timss 2011. Assessment Frameworks.* Boston: Lynch School of Education, Boston College.

MUÑIZ, J., & HAMBLETON, R. K. (1992). Medio siglo de Teoría de Respuesta a los Ítems. *Anuario de Psicología*(52), 41-66.

OECD. (2009). *PISA 2006 Technical Report.* París: OECD.

OLSON, J., Martin, M., & MULLIS, I. (2008). *TIMSS 2007 Technical Report.* Boston: TIMSS & PIRLS International Study Center, Boston College.

OSMAN, S., NAAM, S., JAAFAR, O., BADARUZZAMAN, W. H., & RAHMAT, R. A. (2012). Application of Rasch Model in Measuring Students' Performance In Civil Engineering Design II Course. *Procedia Social and Behavioral Sciences*, 59-66.

SWAIL, W., REDD, K., & PERNA, L. (2003). *Retaining minority students in higher education: A framework for success.* San Francisco: ASHE-ERIC Higher Education Report: Jossey- Bass.

UNIVERSIDAD DE BOGOTÁ JORGE TADEO LOZANO. (2011a). *Modelo Pedagógico.* Bogotá: Universidad Jorge Tadeo Lozano.

UNIVERSIDAD DE BOGOTÁ JORGE TADEO LOZANO. (2011b). *Proyecto educativo Institucional. PEI.* Bogotá: Universidad de Bogotá Jorge Tadeo Lozano.

WRIGHT, B., & STONE, M. (1998). *Diseño de mejores pruebas utilizando la técnica de Rasch.* México: Ceneval.

Correspondencia con los autores

Daniel BOGOYA
Consultor independiente
Kilómetro 6 Vía La Calera – Arboretto
Teléfono: 057 3214914724
e-mail: dbogoya@yahoo.com

Sandra Patricia BARRAGÁN M.
Universidad de Bogotá Jorge Tadeo Lozano
Calle 22 No. 3-30 Módulo 15 Oficina 201
Teléfono: 057 2427030 Extensión: 1701
e-mail: sandra.barragan@utadeo.edu.co

Manuel Ricardo CONTENTO R.
Universidad de Bogotá Jorge Tadeo Lozano
Calle 22 No. 3-30 Módulo 15 Oficina 201
Teléfono: 057 2427030 Extensión: 1701
e-mail: manuel.contento@utadeo.edu.co

Adelina OCAÑA G.
Universidad de Bogotá Jorge Tadeo Lozano
Calle 22 No. 3-30 Módulo 15 Oficina 201
Teléfono: 057 2427030 Extensión: 1701
e-mail: adelina.ocana@utadeo.edu.co