EL MODELO DE RASCH: UNA ALTERNATIVA PARA LA EVALUACIÓN EDUCATIVA EN COLOMBIA

CARLOS A. PARDO ADAMES*

FACULTAD DE PSICOLOGÍA
UNIVERSIDAD CATOLICA DE COLOMBIA

SUBDIRECCIÓN DE ASEGURAMIENTO DE LA CALIDAD
INSTITUTO COLOMBIANO PARA EL FOMENTO DE LA EDUCACIÓN SUPERIOR - ICFES

This paper shows the historical development of the Psychological Testing in Colombia, the State Exam, some of the problems that this discipline has worked in this country, the main problems and weaknesses of the Test Classic Theory and the alternative answers that offers the Item Response Theory (TRI). It presents the models of the TRI, the 3-Parameters and the Rasch, and two computer programs (BILOG-MG and WINSTEPS) that processes information based on these models. Finally, it points out the advantages of the Rasch related to its power as a mathematical model as well as its practical applicability in the educational measurement.
Key words: measurement, assessment, Item Response Theory, Rasch Model, Psychological Testing, State Exam, Test Classic Theory

UNA MIRADA HACIA EL PASADO

La historia de la psicometría en Colombia lleva ya varios años durante los que ha consolidado el ejercicio de la evaluación en diversas ramas de la vida cotidiana como la educación y la salud principalmente. Se ha involucrado tan decididamente en la cultura que no es posible pensar que existan adultos colombianos que de una u otra forma no se afecten por la práctica de la psicometría; es así como se puede encontrar que en algún momento de la vida familiar las conversaciones de sus miembros giran en torno a su aplicación, como cuando hablan acerca de los resultados del Examen de Estado o Examen del ICFES.

De ahí la importancia de que quienes lo utilizan y practican lo hagan con rigor, con criterio, con precisión y , lo más importante, con honestidad, ya que los efectos de la práctica en psicometría duran mucho tiempo en la conciencia de los ciudadanos a quienes afecta.

En 1968 se inicia uno de los eventos educativos de mayor impacto en la realidad

* E-mail:cpardo@hemeroteca.icfes.gov.co

de los colombianos: el Examen de Estado para ingreso a la educación superior, realizado por el Servicio Nacional de Pruebas, y conocido mejor como Examen del ICFES. Podríamos decir que durante todos estos años, el examen se ha constituido en el pilar del ejercicio de la psicometría en el país y el ICFES en un verdadero centro de desarrollo y actualización de su práctica.

El equipo de trabajo ha logrado mantener los más altos criterios de calidad en la práctica psicométrica, lo que supone un desarrollo y progreso constante en las diferentes mediciones y evaluaciones que desarrolla la institución impactando de diversas maneras los procesos educativos que se adelantan en el país y siempre con la conciencia de la búsqueda del mejoramiento de la calidad.

LA PSICOMETRÍA APLICADA EN COLOMBIA

Desde el punto de vista de la psicometría, la práctica de la medición educativa en Colombia ha evolucionado con el tiempo y ha pasado de la Teoría Clásica de las Pruebas a nuevos modelos. El Examen de Estado, se inicia bajo los supuestos de la Teoría Clásica de las Pruebas (TCP). En esa época eran desconocidas otras alternativas debido a que apenas se estaban dando los primeros pasos teóricos y empíricos en ellas y a que no existía divulgación de las mismas excepto en ciertos círculos académicos del exterior.

Para la época y de acuerdo con los desarrollos en torno a procesos de selección de estudiantes en la universidad y en relación con conceptos como los de calidad educativa, eran suficientes los resultados proporcionados por la Teoría Clásica. Es decir, el concepto de calidad educativa estaba íntimamente ligado con el de pruebas de rendimiento y la selección se fundamentaba en la elección de los mejores en un grupo particular. En términos reales se respondía a preguntas como: ¿cuántas respuestas correctas tiene un estudiante? ¿Quién tiene más respuestas correctas?

Esta visión se mantuvo hasta principios de la década del 90 cuando adquieren relevancia discusiones educativas relacionadas con la calidad de la educación como una preocupación sentida en el contexto nacional. Los cambios en el concepto de calidad educativa llevan a replantear los referentes de la evaluación y se retoman propuestas en este campo, poco implementadas en Colombia. Aparece la necesidad de diseñar instrumentos cuyo marco de interpretación sea con referencia a criterio, es decir de interpretaciones con referencia a constructos claramente elaborados que soportan la validez de los resultados. En este caso, el concepto de calidad se relaciona más con mediciones a través de pruebas de desempeño o de competencia. En términos reales se responde a preguntas como: ¿qué tan bien se realiza algo? ¿por qué...? ¿Cómo...?

A partir de 1995 se inicia el proyecto de reconceptualización del examen de estado para ponerlo a tono con los planteamientos de la Ley General de Educación de 1994 y la discusión pedagógica nacional. Debido a que esta reconceptualización lleva a transformaciones fundamentales en las pruebas tanto a nivel de los referentes conceptuales que maneja como en los formatos y propósito de la evaluación, es necesario replantear el uso de los modelos psicométricos utilizados hasta el momento.

LOS PROBLEMAS Y LAS SOLUCIONES

En relación con los cambios realizados en el Examen de Estado para ingreso a la educación superior, la Teoría Clásica tiene diversos problemas, siendo los más importantes los que se mencionan a continuación y que impidieron utilizarla para el procesamiento de información.

Tal vez uno de los problemas más importantes que afronta la TCP, es la imposibilidad de análisis de interpretaciones de resultados con referencia a criterio. Este hecho se debe, especialmente, a que los valores de sus estadísticas dependen de la muestra de examinados. El nuevo Examen de Estado dio un vuelco total al diseñar pruebas cuyos resultados se interpretarían con referencia a criterio con el propósito de producir resultados con significado pedagógico - educativo.

Para superar esta dificultad, la solución es utilizar un modelo que estime los estadísticos de los ítems independientemente de la muestra de personas que aborda la prueba. Es decir que sin importar qué grupo de personas responde a una prueba, las calibraciones que realice deben ser invariables. Adicionalmente se requiere que presente los resultados en una verdadera escala de medición en la que se puedan establecer ciertos puntos con significado pedagógico - educativo. Además, que tenga procesos para estimar la confiabilidad y la validez de los items bajo los supuestos de pruebas con referencia a criterio.

Otro problema importante hace referencia a la inflexibilidad del modelo para establecer comparaciones de resultados en distintas aplicaciones. Desde el punto de vista de los propósitos del examen es vital que una institución de educación superior pueda utilizar resultados de diversos años para sus procesos de selección de estudiantes, de tal manera que la interpretación particular que hagan de un resultado sea igual año tras año. En otras palabras, no importa cuando se haya presentado el examen, un resultado particular debe tener las mismas consecuencias en los procesos posteriores. La TCP soluciona este problema con base en el concepto de pruebas paralelas, las cuales suponen estructuras de prueba rígidas, lo mismo que la idéntica distribución de estadísticos de item y de prueba. Para que esto se cumpla, los estadísticos deben cambiar con los cambios de la población haciendo imposible reconocer cambios reales en la calidad de la educación.

La segunda exigencia que se le haría a un modelo psicométrico para resolver este segundo problema es que pudiera ser utilizado para procesos de equating (comparabilidad) de resultados, a partir de cierta flexibilidad en las estructuras de prueba y en la distribución de estadísticos de item o de prueba. Es decir que la comparabilidad se fundamente en la validez y no tanto en los valores estadísticos dependientes de grupos poblacionales.

Una medición basada en competencias implica abordar la medida a través de diversas posibilidades de acción de los evaluados, esto es que es posible necesitar para ciertas acciones, la utilización de formatos de item diferentes al de selección de respuesta con única o varias respuestas correctas. La TCP no tiene problemas con preguntas cuya respuesta sea calificada de manera dicótoma, es decir correcta o incorrecta (uno o cero). Pero le resulta imposible trabajar con otros formatos de ítem.

Esta tercera exigencia hace referencia a la capacidad del modelo psicométrico para abordar la medición a través de formatos

de item de respuesta polítoma o de crédito parcial (es decir que se da cierto crédito o valor a cada opción de respuesta, no hay opciones erradas).

LAS POSIBILIDADES

Por todo lo planteado, y como sucede actualmente en todos los ámbitos de la psicometría, en el ICFES se descartó la opción de usar la TCP para abordar, desde la psicometría, los planteamientos del nuevo Examen de Estado. Sólo habría dos posibilidades en donde encontrar las respuestas a los problemas planteados y a otros de no menor importancia. Esas dos posibilidades son la Teoría Respuesta al Item y el Modelo de Rasch. Ambas aproximaciones presentan alternativas de solución a los problemas planteados, así que fué necesario comparar los procesos y alternativas ofrecidas por los modelos para seleccionar aquel que mejor diera solución a los problemas planteados en particular y que respondiera mejor a preguntas sobre medición mucho más generales.

Tradicionalmente se ha considerado el modelo de Rasch como perteneciente a la Teoría Respuesta al Item (TRI), pero precisiones realizadas por el Instituto de Medición Objetiva, establecen que la TRI o los modelos basados en ella hacen una descripción de los datos, mientras el modelo de Rasch especifica cómo interactúan pruebas, personas, items, calificadores, etc., para construir mediciones lineales a partir de observaciones dicótomas. Esto plantea una discusión de fondo en relación con la medición, que valdría la pena retomar con intensidad en otro momento.

Por el momento, se compararán las alternativas de solución a los problemas antes planteados que proponen el modelo de Rasch y modelo que ha proporcionado mejores resultados desde una perspectiva de la TRI, es decir el modelo de tres parámetros, propuesto por Lord.

LOS DOS MODELOS

Un modelo de medición establece una relación entre la ejecución de una persona en una prueba y una variable, objeto de la medición, o sea su habilidad, según Hambleton y Swaminathan (1996) "es un axioma en la Teoría Respuesta al Item que aquello que soporta la ejecución de un examinado en una prueba es su habilidad. El término habilidad (o habilidad latente como se le llama a veces) es una etiqueta que se usa para designar la característica que mide una prueba. Esta característica medida puede ser definida ampliamente para incluir habilidades cognitivas, rendimiento, competencias básicas, características de la personalidad, etc.". En términos generales la ejecución de una persona en una prueba son sus respuestas a los ítems, en este sentido, la Teoría clásica de las pruebas representa la relación en la ecuación, Puntuación Observada = Puntuación verdadera + Error. Queda claro que la TCP se fundamenta en las puntuaciones de las personas a una prueba[1].

[1] Información amplia sobre este tema se puede encontrar en NOVICK, M. Y JACKSON, P. 1974. Statistical Methods for Educational and Psychological Research. McGraw Hill.; MEHRENS, W. Y LEHMANN, I. 1982. Medición y Evaluación de la Educación y en la Psicología, CECSA, México.; LORD, F. Y NOVICK, M. 1968. Statistical Theories of Mental Test Scores, Addison Wesley.

Como dicen Embretson y Hershberger (1999), la Teoría Respuesta al Item se aplica a las respuestas individuales a cada ítem. De ahí, se estima la probabilidad de que una persona responda correcta o incorrectamente teniendo como base la habilidad de la persona y de los parámetros[2] del ítem.

Como dicen Hambleton y Swaminathan (1996) cualquier teoría psicométrica basada en las respuestas a ítems, supone que la ejecución de una persona en una prueba puede explicarse al definir las características del evaluado, al estimar las puntuaciones del evaluado y al utilizar las puntuaciones para explicar las ejecuciones del evaluado en los ítems o en las pruebas. Un modelo de respuesta a ítems especifica una relación entre las ejecuciones de una persona en los ítems de una prueba y las habilidades que se asume soportan esa ejecución. Estas relaciones se describen por una función matemática, por lo que se dice que los modelos de respuesta a items son modelos matemáticos.

En este sentido podremos encontrar diferentes modelos matemáticos en la Teoría Respuesta a Ítems, cada uno de los cuales se fundamenta en supuestos específicos acerca de los datos. Como se mencionó anteriormente, estos modelos matemáticos describen la probabilidad de respuestas específicas a un ítem.

Las principales características de los modelos de respuesta a ítems son descritas por Hambleton y Swaminathan (1996):

- Las estimaciones de los parámetros de los items son independientes del grupo de examinados que responden una prueba.
- Las estimaciones de las habilidades de los examinados son independientes de los ítems particulares utilizados en una prueba.
- Se conoce la precisión de las estimaciones de la habilidad y de la dificultad.

EL MODELO DE TRES PARÁMETROS

Uno de los modelos más utilizados es el Modelo Logístico de tres Parámetros que se obtiene al adicionar un parámetro al modelo de dos parámetros (Brinbaum, 1968).

La forma matemática del modelo de tres parámetros es (Hambleton y Swaminathan, 1996; Zimowski, Muraki, Mislevy y Bock, 1996):

$$P_i(\theta) = c_i + (1 - c_i)\frac{e^{Da_i(\theta - b_i)}}{1 + e^{Da_i(\theta - b_i)}}$$

(i = 1, 2,..., n)

Donde:

$P_i(\theta)$

La probabilidad que un examinado con habilidad 2 responda correctamente el ítem i.

b_i

Parámetro de dificultad del ítem.

a_i

Parámetro de discriminación del ítem

c_i

Representa la probabilidad de los exami-

[2] Parámetro, según la Real Academia Española (1998), es una "variable que, en una familia de elementos, sirve para identificar cada uno de ellos mediante su valor numérico". En el caso de la TRI, los parámetros los explicita cada modelo.

nados con habilidades bajas, respondan correctamente el ítem.

D
Es una constante.

EL MODELO DE RASCH

El otro modelo utilizado para el análisis de información dentro de la perspectiva de este documento, es el Modelo de Rasch[3]. Este modelo parte de dos supuestos fundamentales:

- La probabilidad de responder correctamente a una pregunta, es mayor para una persona con mayor habilidad
- La probabilidad de responder correctamente, es mayor para una pregunta fácil que una difícil.

La forma matemática del modelo de Rasch corresponde a la ecuación (FISCHER y MOLENAAR, 1995):

$$P\left(X_{vi} = x_{vi} / \theta_v, \beta_i\right) = \frac{\exp\left[x_{vi}\left(\theta_v - \beta_i\right)\right]}{1 + \exp\left(\theta_v - \beta_i\right)}$$

Donde:

$P\left(X_{vi} = x_{vi} / \theta_v, \beta_i\right)$
La probabilidad que un examinado con habilidad 2 responda correctamente el ítem i con dificultad ∃.

exp
Antilogaritmo natural

2
Habilidad del examinado

∃
Dificultad del item.

Estos son los dos modelos que se sometieron a estudio con datos reales de mediciones educativas realizadas por el ICFES en su Examen de Estado, de tal manera que se pudieran probar y comprobar las bondades de cada uno de ellos para abordar satisfactoriamente las problemáticas planteadas anteriormente. Dada la complejidad de estos modelos, se hace necesario utilizar software especializado para el análisis de datos con ellos, tema que se tratará a continuación.

LOS PROGRAMAS DE COMPUTADOR

Debido a la complejidad de la TRI y sus modelos, no es fácil encontrar programas de computador que asuman los supuestos de cada uno. Para el presente estudio se utilizaron los programas de mayor desarrollo y diseñados por quienes han trabajado más estrechamente y por más tiempo los modelos.

Para el modelo de Tres Parámetros, se utilizó el programa BILOG - MG (1996), desarrollado por Michele Zimowski del National Option Research Center, Eiji Muraki del Educational Testing Service, Robert Mislevy del Educational Testing Service y Darrell Bock de la Universidad de Chicago.

Para el Modelo de Rasch, se utilizó el programa WINSTEPS (1997 - 2001), desarrollado por John Linacre de la Universidad de Chicago y Benjamín Wright, también de la Universidad de Chicago.

[3] Este modelo fue propuesto inicialmente por el matemático danés Georg Rasch en 1960.

El programa BILOG-MG (Scientific Software International 1996), asume respuestas binarias a ítems y emplea métodos bayesianos y de máxima verosimilitud para las estimaciones. Se puede seleccionar modelos de 1, 2 y 3 parámetros para las estimaciones. Adicionalmente ofrece varias posibilidades de análisis ya sea para subpruebas, para múltiples grupos, para respuestas omitidas, entre otros.

El programa WINSTEPS (MESA press, 1991 - 2001), está diseñado para construir mediciones a partir de conjuntos de respuestas de personas a ítems ya sea dicótomas o polítomas. Para las estimaciones utiliza los algoritmos PROX (aproximación normal) para estimaciones gruesas y luego el UCON (máxima verosimilitud incondicional, máxima verosimilitud conjunta) para estimaciones precisas. Ofrece varias posibilidades de tratamiento de datos, archivos de salida y estadísticas de análisis.

LAS PRUEBAS Y SUS MÉTODOS

Cada uno de los problemas planteados fue abordado tratando de encontrar la mejor solución a partir de la utilización de un modelo psicométrico u otro. Los métodos utilizados fueron aquellos que pusieran a prueba los modelos de tal manera que se pudieran reconocer las virtudes y debilidades de los mismos. Cada uno de los problemas fué resuelto (o se intentó) utilizando los dos modelos (excepto en el tercer problema).

Primer Problema:

Independencia en la calibración de parámetros. Para abordar este primer problema se utilizó una base de datos con las repuestas de 36394 estudiantes de grado 11 que respondieron la prueba de biología del examen de estado de agosto de 1998. La prueba de biología consta de 50 ítems de selección múltiple con única respuesta y por lo tanto su calificación es dicótoma. La población de estudiantes fué divida en dos grupos de acuerdo con su puntaje bruto[4]: Un grupo con las respuestas de las personas que tuvieron un puntaje bruto entre 0 y 25 puntos, al que denominaremos de ahora en adelante Grupo Bajo y el otro con puntaje bruto entre 26 y 50, al que denominaremos Grupo Alto. La base de datos que contiene la totalidad de estudiantes se denomina Grupo Total. En el Grupo Bajo quedaron 17454 estudiantes y en el Grupo Alto 18940.

La primera prueba trata de comprobar si las calibraciones de los parámetros de los ítems, que hacen los modelos, son independientes de las habilidades[5] de las personas y viceversa. Es decir que se puedan calcular los estadísticos correspondientes a los ítems de manera que no varíen al cambiar las poblaciones de donde se obtienen los datos y que al calcular las habilidades de las personas, éstas no varíen cuando se cambian los itmes. En otras palabras se requiere comprobar que al calcular la escala de evaluación, esta no dependa del grupo de personas que abordan el examen.

[4] Se denomina Puntaje Bruto al número de puntos obtenidos a partir de las respuestas en una prueba. En pruebas con preguntas dicótomas, el puntaje bruto equivale al número de respuestas correctas.

[5] En el caso del nuevo examen de estado la "habilidad" son las competencias evaluadas.

Para el modelo de Rasch, se decidió realizar la prueba con poblaciones extremas, es decir, calibrar una prueba con dos poblaciones de habilidades radicalmente diferentes.

Se procesaron los resultados de Grupo Alto y de Grupo Bajo, por separado. Se trata de comprobar que las calibraciones que haga el modelo es independiente de las habilidades del grupo de personas analizadas[6]. En caso de que sean independientes, las calibraciones deben dar el mismo resultado, es decir las escalas de resultados construidas para cada grupo poblacional, deben ser muy semejantes, casi iguales. Es decir que se construye una escala de resultados a partir de los datos de los chicos del Grupo Alto y se construye otra escala de resultados a partir de la información de los chicos del Grupo Bajo.

Para poner a prueba el modelo de 3 parámetros, se procesó la información del Grupo Alto y la información del Grupo Total y se comparó el cálculo de habilidad en los dos grupos para las mismas personas[7]. En otras palabras, se calculó la calificación de un persona, cuando ella forma parte de los estudiantes del Grupo Alto y también se le califica su resultado en la prueba cuando forma parte del Grupo Total. En este caso se califica a la misma persona, pero en compañía de grupos distintos. Se esperaría que, por ser la misma persona, su calificación sea igual o muy semejante.

Los resultados de uno y otro modelo aparecen en las Tablas 1 y 2. Como se puede observar la calibración que hace el modelo de Rasch para los dos grupos poblacionales genera resultados muy semejantes. Adicionalmente, se observa que el modelo estima la habilidad correspondiente a un puntaje bruto particular. En el Grupo Bajo estima las habilidades para resultados entre 26 y 50 y para el Grupo Alto estima las habilidades para resultados entre 0 y 25 puntos. Es decir que aunque no hay alguna persona con un puntaje particular el modelo estima cuanto debería sacar alguien con ese puntaje, por eso cuando calcula las habilidades de las personas del Grupo Bajo no se limita sólo a generar la escala de calificación para los puntajes entre 0 y 25, sino que también genera las habilidades para quienes tienen resultados entre 26 y 50 puntos.

Con el modelo de 3 parámetros no se puede realizar esta prueba ya que no puede estimar los datos de poblaciones ausentes. Es decir que no puede calcular la calificación para un puntaje bruto particular si no hay una persona con ese puntaje bruto. Es decir que si se toma a las personas del Grupo Alto, sólo puede estimar los resultados para puntajes entre 26 y 50. Es por esta razón que se selecciona otro método igualmente riguroso pero con poblaciones distintas a las utilizadas para el modelo de Rasch.

---

[6] Por ejemplo, una calibración a partir de la TCP implicaría la construcción de una escala de deciles; si se aplica a los resultados de estos grupos, claramente observaríamos que produce diferentes escalas para cada grupo, llevándonos a confirmar que la calibración no es independiente del grupo de personas que aborda la prueba. Lo mismo sucede si se construye una escala estándar.

[7] Debido a que el modelo de tres parámetros no tiene una estadística suficiente para sus cálculos, no puede generar una tabla de equivalencia entre el puntaje bruto y la habilidad de las personas. Esto implica que para verificar que la calibración de las habilidades sea independiente de la muestra, se debe calibrar la habilidad de una misma persona en grupos distintos.

Tabla 1. Estimación de las habilidades para cada puntaje bruto a partir de los datos de los estudiantes del Grupo Bajo y del Grupo Alto.

	Grupo Bajo	Grupo Alto
0	-4.92	-5.12
1	-4.21	-4.39
2	-3.48	-3.64
3	-3.04	-3.18
4	-2.72	-2.84
5	-2.46	-2.57
6	-2.24	-2.34
7	-2.05	-2.13
8	-1.87	-1.95
9	-1.72	-1.79
10	-1.57	-1.63
11	-1.44	-1.49
12	-1.31	-1.36
13	-1.19	-1.23
14	-1.08	-1.11
15	-0.97	-0.99
16	-0.86	-0.88
17	-0.76	-0.77
18	-0.66	-0.67
19	-0.56	-0.56
20	-0.46	-0.46
21	-0.37	-0.36
22	-0.27	-0.26
23	-0.18	-0.17
24	-0.09	-0.07
25	0	0.03
26	0.1	0.12
27	0.19	0.22
28	0.28	0.31
29	0.38	0.41
30	0.47	0.51
31	0.57	0.61
32	0.66	0.71
33	0.76	0.81
34	0.87	0.92
35	0.97	1.02
36	1.08	1.14
37	1.2	1.25
38	1.31	1.37
39	1.44	1.5
40	1.57	1.64
41	1.72	1.78
42	1.87	1.94
43	2.04	2.11
44	2.23	2.3
45	2.45	2.52
46	2.71	2.78
47	3.03	3.11
48	3.47	3.55
49	4.2	4.28
50	4.91	4.99

En la Tabla 2, correspondiente a la estimación de habilidad para 10 estudiantes, ya sea cuando se encuentran con el Grupo Alto o con el Grupo Total, se observa la diferencia en la estimación que refleja la dependencia de las calibraciones en el grupo poblacional particular donde se realice. En este caso, y como se mencionó antes, se calcula la habilidad de 10 personas cuando ellas se encuentran en el Grupo Total y en elGrupo Alto. Se esperaría que los resultados sean iguales.

Esto es, si calculamos la habilidad de las personas con el programa BILOG-MG, a un grupo de personas, los resultados son diferentes si estos se encuentran en un grupo o en otro, aunque su propias respuestas a cada pregunta no cambian. Lo que sucede es que la estimación de los parámetros es diferente en cada grupo para cada pregunta. Esto hace que la calibración de habilidades sea dependiente de la muestra de personas.

Segundo Problema

Equivalencia de resultados en dos aplicaciones distintas. La solución al segundo problema es el proceso de ***equating***, que se usa para ajustar las puntuaciones que se obtienen en formas diferentes de una misma prueba de tal manera que sean intercambiables a pesar de que los niveles de dificultad de las pruebas sean distintos. En la TCP se utiliza un procedimiento semejante y que consiste en comparar estadísticas de grupos poblacionales.

Es un problema reconocido que cuando se aplican dos pruebas diferentes a distintas poblaciones, las escalas de resultados pueden no ser comparables, especialmente cuando una de las dos pruebas es más difícil que la otra. Debido a esto se requiere realizar algún procedimiento particular que garantice que las calificaciones sean comparables. En este caso, y como ya se mencionó, el procedimiento se conoce con el nombre de equating. Se exige que las estimaciones de los parámetros de los ítems sean independientes del grupo de personas que los abordan.

En el caso específico del Examen de Estado, se utiliza el procedimiento de grupos no equivalentes con items comunes. Como se sabe, en dos exámenes de estado diferentes (digamos aplicados en dos años consecu-

Tabla 2. Estimación de la habilidad de 10 estudiantes a partir de datos con el Grupo Total y con el Grupo Alto.

	Grupo Bajo	Grupo Alto
1	-0.96	-1.19
2	1.58	1.51
3	0.57	0
4	0.56	-0.05
5	0.72	0.2
6	1.01	0.28
7	1.49	1.23
8	0.98	0.38
9	1.44	1.04
10	0.86	0.23

tivos), se procesan resultados de diferentes estudiantes; para realizar el equating es necesario que ambos grupos de personas respondan a algunas preguntas comunes. En este sentido se requiere que la estimación de los parámetros de los ítems sea independiente de la muestra de examinados, porque cualquier cambio en sus parámetros puede considerarse como un cambio real en la calidad de la educación de la población que responda; de esta manera se pueden reconocer tendencias en los resultados de la población.

El método utilizado consistió en calibrar los parámetros de los items con ambos modelos, en dos poblaciones semejantes de tal manera que se esperaría que los resultados fueran muy semejantes, casi idénticos. Para ello se utilizó una base de datos con las respuestas de 36394 estudiantes a la prueba de biología aplicada en 1998 y una submuestra aleatoria de esta misma base de datos con las respuestas de 18196 de esos estudiantes.

En conclusión, las estimaciones del parámetro de las preguntas con el modelo de Rasch no dependen del grupo poblacional particular donde se realice ya que el promedio de sus diferencias es inferior al error promedio de medición. Para el modelo de 3 parámetros, las diferencias en las estimaciones del parámetro de dificultad (que se utiliza en el proceso de equating) son más altas y no existe una estimación del error de medición para determinar la bondad de la estimación.

Tercer Problema

La solución al tercer problema, el de calificación de preguntas de crédito parcial, es evidente. Sencillamente se requiere que el modelo pueda procesar información diferente a la dicótoma, es decir diferente a los 1 (unos) y 0 (ceros) conque tradicionalmente ha calificado la psicometría. La respuesta es sencilla: el modelo de Rasch si puede procesar información con pesos diferentes para las distintas opciones, mientras que el modelo de 3 parámetros no lo puede hacer.

En términos reales, hoy en día se están utilizando diversos formatos de preguntas que evalúan competencias en distintos sistemas de evaluación de la educación[8], los cuales requieren, en algunas ocasiones, reconocer la ponderación de algunas opciones de respuesta que son válidas aunque no sean la respuesta más correcta.

En el Examen de Estado para ingreso a la Educación Superior, se utilizan items con formato de respuesta múltiple válida en la prueba de matemáticas del núcleo común, en donde dos opciones son consideradas correctas, y en la prueba de profundización en lenguaje y la interdisciplinaria de medio ambiente, todas las opciones son válidas pero cada una tiene diferente grado de validez. En estos casos es indispensable recurrir a un modelo que pueda analizar este tipo de datos, como lo es el modelo de Rasch.

[8] En Colombia se viene haciendo una evaluación por competencias desde 1991 en el programa de Evaluación de la Calidad de la Educación, SABER, cuyas pruebas ha diseñado el ICFES. En época más reciente, el Examen de Estado para Ingreso a la Educación Superior, ha adoptado la evaluación por competencias desde el año 2000.

A MANERA DE CONCLUSIÓN. UNA ÚLTIMA PALABRA

De acuerdo con las necesidades mencionadas en este documento es claro llegar a la conclusión que el modelo de Rasch las satisface mientras que el de 3 Parámetros no lo hace de la misma manera o no lo hace definitivamente. Creo que podríamos continuar con la especificación de problemas de medición que deben ser resueltos y llegaríamos a las mismas conclusiones: el modelo de Rasch los resuelve mientras que el de tres parámetros no en todos los casos. Sin importar la prueba específica y su uso particular. Este sería el caso siempre y cuando quisiéramos hacer medición.

Uno de los aspectos que explica los resultados de las pruebas realizadas a los dos modelos es la diferencia en las ecuaciones de estimación de los parámetros en cada modelo.

En el modelo de 3 Parámetros, según WRIGHT (1992):

$$\underset{i}{a_i X_{\theta i}} = \underset{i}{a_i P_{\theta i}} \blacklozenge \theta$$

$$\underset{\theta}{\theta X_{\theta i}} = \underset{\theta}{\theta P_{\theta i}} \blacklozenge a_i$$

Como se puede observar hay una ponderación cruzada entre el parámetro de habilidad y el de discriminación en este modelo lo que garantiza la divergencia. Es decir, en los procesos iterativos para estimar los parámetros, no se llega a ningún valor y las iteraciones pueden ser infinitas, inclusive con datos ideales que se ajusten al modelo. De alguna manera hay que hacer un poco de trampa y detener el proceso de iteración en algún momento[9].

En el modelo de Rasch, según WRIGHT (1992):

$$\underset{i}{X_{ni}} = \underset{i}{P_{ni}} \blacklozenge B_n$$

$$\underset{n}{X_{ni}} = \underset{n}{P_{ni}} \blacklozenge D_i$$

En este caso se llega inevitablemente a la convergencia en los procesos de iteración. El software utilizado para el procesamiento, presenta la información del número de iteraciones necesarias para encontrar convergencia en las estimaciones.

Desde otro punto de vista, el modelo de 3 Parámetros y todos los demás que se encuentran en la IRT, están diseñados para imitar datos, para aceptar cualquier clase de datos. Dicen cómo son los datos, los describen, pero están lejos de establecer si los corresponden a una verdadera medición

Los análisis de datos con el modelo de Rasch requieren la investigación y cuantificación de exactitud, precisión, confiabilidad, validez de constructo, control de calidad del ajuste de estadísticas, información estadística, linealidad, dependencia local y unidimensionalidad. El modelo de Rasch implementa la ordenación estocástica de Guttman, aditividad conjunta, la concatenación de Campbell, suficiencia y divisibilidad infinita.

---

[9] El software empleado para el modelo de 3 Parámetros el BILOG-MG o el Multilog, traen 10 iteraciones por defecto, pero es una variable que se puede controlar en el archivo de comandos.

Todos los aspectos mencionados en el presente documento nos colocan a las puertas de avances gigantescos en la medición realizados por la psicometría. Puertas que nos abren un futuro mucho más prometedor en la investigación y en el conocimiento.

REFERENCIAS

Andersen, E. 1980. *Discrete Statistical Models With Social Science Applications*. North-Holland publishing company, Amsterdam.

Choppin, B. 1985. «Bruce Choppin on measurement an education». *Evaluation in Education: An international Review series, 9, # 1*.

Embretson, S. Y Hershberger, s. 1999. *The New Rules of Measurement: what every psychologist and educator should know*. Lawrence Earlbaum Associates. New Jersey.

Fischer, G. y Molenaar, I. 1995. *Rasch Models: foundations, recent developments and applications*. Springer-Verlag. New York.

Haladyna, T. y Roid, G. 1983. A comparison of two approaches to criterion-referenced test construction. *Journal of Educational Measurement*, 20, pp. 271-282.

Hambleton, R., Cook, L. Latent trait models and their use in the analysis of educational test data. *Journal of Educational Measurement*, 14, pp. 75-96 1977.

Hambleton, R., Gruijter, D. 1983. Application of item response models to criterion-referenced test item selection. *Journal of Educational Measurement*, 20, pp. 355-367.

Hambleton, R. y Swaminathan, H. 1996. *Item Response Theory: principles and applications*. Kluwer-Nijhoff Publishing. Boston.

Holmes, S. 1982. Unidimensionality and vertical equating with the Rasch model. *Journal of Educational Measurement*, 19, pp. 139-147.

Linacre, J. Y Wright, B. 2000. A *User's Guide to WINSTEPS. MESA press*. Chicago.

Lord, F. Y Novick, M. 1968. *Statistical Theories of Mental Test Scores*. Addison-Wesley.

Marco, G. Item characteristic curve solutions to three intractable testing problems, *Journal of Educational Measurement*, 14, pp. 139-160.

Mehrens, W. Y Lehmann, I. 1982. *Medición y Evaluación en la Educación y en la Psicología*. CECSA. México.

Novick, M. Y Jackson, P. 1974. *Statistical Methods for Educational and Psychological Research*. McGraw Hill.

Pardo. C. 1999. Transformaciones en las Pruebas para Obtener Resultados Diferentes. *Serie: Nuevo Examen de estado Cambios para el Siglo XXI*. ICFES. Bogotá.

Thorndike, R. 1989. *Psicometría aplicada. Limusa*, México.

Van Der Linden, W. 1982. Criterion-referenced measurement: its main applications, problems and findings. *Evaluation in education*, t, pp. 97-118.

Wright, B. 1992. IRT in the 1990s: Which Models Work Best?. En: Rasch Measurement Transactions. Vol. 6 N° 1 pp 196 - 200.

Wright, B. Solving measurement problems with the Rasch Model. *Journal of Educational Measurement*, 14, pp. 9.

Zimowski, M., Muraki, E., Mislevy, R. Y Bock D. 1996. BILOG-MG Multiple-Group IRT analysis and test maintenance for binary items. *Scientific Software International*. Chicago.