ACTA COLOMBIANA DE PSICOLOGÍA 5, 87-98-01

CONSIDERACIONES SOBRE LAS TEORÍAS DE MEDICIÓN: UN EJEMPLO DE APLICACIÓN BAJO LA TEORÍA DE RESPUESTA AL ITEM (TRI)

BERTHA LUCÍA AVENDAÑO PRIETO*
ELVERS WILLIAM MEDELLÍN LOZANO**

UNIVERSIDAD CATOLICA DE COLOMBIA

This article shows a review of the theories developed about the measurement of variables in psychology. It presents a historic revision and the main aspects of the Test Classic Theory (TCT), the Generalization Theory (GT) and the Item Response Theory (TRI). Given the importance that today has the TRI, a hipothetic example is used in order to explain and interpret some of the most important results of the information analyses supported by this theory.
Key words: Measurement, Test Clasic Theory, Generalization Theory, Item Response Theory, Item curve characteristic.

La Psicometría entendida como el conjunto de modelos formales que posibilitan la medición de variables psicológicas (Martínez, 1995), ha permitido a la Psicología el avance en la medición de todo tipo de entidades relacionadas con el estudio del comportamiento. De la misma forma como la física ha generado modelos para desarrollar y validar sus procesos de medición, la Psicología a través de la Psicometría ha elaborado modelos teóricos que le permiten llevar a cabo todo proceso de medición, sentando las bases para que se realicen de forma adecuada.

Se pueden resaltar dos modelos estadísticos que han tenido trascendencia en la construcción de pruebas: la teoría clásica de los test (TCT), y la teoría de respuesta al ítem (TRI)[1].

La TCT ha sido la más utilizada para validar los diferentes instrumentos empleados en Psicología, este modelo se basa en el concepto de puntuación observada de un test; que implica por una parte, la puntuación verdadera y por otra el error de medida, derivándose de esto el siguiente modelo lineal para las puntuaciones

$$x = b + e$$

Siendo x la puntuación observada, b la puntuación verdadera no observada de las

* Profesores Universidad Católica de Colombia. E-mail:psicologia@ucatolica.edu.co

[1] Adicionalmente se considera la teoría de la generalizabilidad (TG), pero como una extensión de la TCT.

diferentes personas y *e* una desviación no sistemática o aleatoria de la puntuación verdadera. A partir de éste modelo lineal la TCT desarrolló todo un conjunto de deducciones dirigidas a estimar la cantidad de error que afecta las puntuaciones de los test.

La TCT plantea además que no existe correlación entre las puntuaciones verdaderas de los participantes en un test y sus respectivos errores de medida, es decir que el tamaño de los errores no está asociado con el tamaño de las puntuaciones verdaderas.

Esta teoría supone también que los errores de medida de los participantes en un test no correlacionan con sus errores de medida en otro test, por lo tanto, si se aplican correctamente los tests, los errores serán aleatorios en cada ocasión, no existiendo razón apriori para que covarien sistemáticamente unos con otros (Muñiz, 1994).

Antes de utilizar una prueba como instrumento de medición apropiado para medir una variable psicológica bajo este modelo, debe obtenerse información sobre su confiabilidad y validez, dos conceptos fundamentales en los que se basa la TCT, que hacen referencia a la consistencia de los resultados y al grado en el cual la prueba mide aquello para lo que fue diseñada.

Las mediciones de variables psicológicas, como en cualquier otra ciencia, han de ser confiables, es decir deben tener pocos errores de medida. Los errores de medida de los que se ocupa la fiabilidad son aquellos no sometidos a control e inevitables en todo proceso de medir, sea químico, físico o psicológico. Con frecuencia las diferencias entre una medición y otra no dependen sólo de de estos errores, pudiendo explicarse además por los cambios operados en los sujetos, debidos a procesos madurativos, intervenciones o eventos de cualquier otro tipo. En estos casos la inestabilidad de las mediciones requiere de una explicación y carece de sentido atribuirla a los errores aleatorios. La confiabilidad no trabaja con este tipo de errores, los cuales deben venir explicados por los modelos manejados. En estas situaciones el evaluador debe identificar las fuentes de error que afecten a las mediciones y no considerar, como en muchas ocasiones, a la baja confiabilidad de los instrumentos de medida lo que puede ser sencillamente la variabilidad normal de la variable medida (Muñiz, 1998).

Es importante aclarar que la confiabilidad se refiere a la estabilidad de las mediciones cuando no se parte del supuesto, confirmado teórica y empíricamente, de que la variable medida no fue modificada de manera diferencial para los sujetos. Un ejemplo relacionado con este concepto, es el de la medición de la inteligencia; si hoy se aplica una prueba a un grupo de sujetos para establecer el nivel de desarrollo cognoscitivo y se vuelve a aplicar al día siguiente, las variaciones en los CI deberán ser pequeñas, indicando que la prueba es confiable y que los errores aleatorios son pequeños.

Existen diferentes estrategias para estimar la confiabilidad, entre las cuales se pueden resaltar el coeficiente de confiabilidad, la estimación empírica del coeficiente de confiabilidad, la estimación de las puntuaciones verdaderas, confiabilidad de las diferencias y los tipos de errores de medida.

Establecer la confiabilidad de un instrumento es necesario pero no suficiente para garantizar una medición apropiada, es esencial establecer la validez. Este concepto concepto se puede definir en varios niveles y de diversos modos. El sentido del concepto se puede comunicar mediante los diversos tipos de preguntas a los que intenta res-

ponder los análisis de validez: ¿Qué rasgos está midiendo la prueba? ¿Mide la prueba el rasgo para el que fue construida? ¿Qué se puede predecir a partir de las calificaciones de la prueba? ¿Qué porcentaje de la varianza en las calificaciones de la prueba se puede atribuir a la variable que mide? Por lo tanto, la validez de una prueba se define ya sea a través de 1) la extensión con que la prueba mide un rasgo subyacente hipotético o 2) la relación entre las calificaciones de la prueba y alguna medida de criterio externo (Brown, 1980, citado por Medellín, 2001).

De igual forma que en la confiabilidad, existen distintas formas para determinar la validez, sin embargo las que tradicionalmente se han utilizado en la TCT han sido, la validez de constructo, la validez de contenido y la validez relacionada con el criterio.

Adicionalmente a los principios de confiabilidad, validez y error en la medición, es importante resaltar que la TCT utiliza el modelo de la curva normal para representar la distribución de las mediciones obtenidas con las pruebas psicológicas. Es decir que esta teoría parte del supuesto que las características, atributos o rasgos se presentan en las poblaciones bajo una distribución normal. Sin embargo, En la década del 80 se comprende el hecho de que la distribución de los resultados que se obtenían en las pruebas correspondía a los niveles de dificultad de las preguntas utilizadas, esto significó que se puso en tela de juicio el hecho aceptado comúnmente de que los atributos medidos en poblaciones se distribuían normalmente (Pardo, 2000).

Una razón por la cual la TCT fue y sigue siendo empleada como marco para la construcción e interpretación de test psicológicos es que permite ser utilizada en diferentes campos de aplicación; pero tal vez el mayor aporte de la TCT fue encontrar un modelo estadístico que fundamentó las puntuaciones de los test y permitió la estimación de los errores de medida asociados a todo proceso de medición.

Por otro lado, la Teoría de la Generalizabilidad (TG) ha sido un modelo menos utilizado que la TCT para validar instrumentos, aplica procedimientos derivados de los modelos de análisis de varianza (Anova) y del diseño experimental a los datos de los test (Martínez, 1995). La TG fue desarrollada por L. J. Cronbach y sus colaboradores (Cronbach, Rajaratnam y Gleser, 1963; Cronbach, Gleser, Nanda y Rajaratnam, 1972.) y se basa esencialmente en la estimación de los diversos componentes de la varianza (Martínez, 1995).

El marco conceptual de la TG sustituye las nociones clásicas de puntuación verdadera y error que maneja la TCT, por otros conceptos más amplios que implican que una medida es una muestra de observaciones aceptable, es decir, los ítems representan solo una muestra de una población mucho más amplia de potenciales ítems, a partir de los cuales se intentan hacer generalizaciones más allá de la muestra particular de ítems.

Un concepto importante que incluye Cronbach en su teoría es el de Faceta, utilizado para designar cada una de las características de la situación de medida que pueden modificarse de una medición a otra y que, en consecuencia, pueden variar el resultado obtenido (Martínez, 1995). Si se traslada éste concepto al modelo estadístico en que se basa la teoría (Anova), las facetas son equivalentes a los factores, sus resultados representan efectos principales y sus combinaciones con otras facetas, o con el

objeto de medida, serían las interacciones. Un ejemplo de faceta pueden ser las diferentes situaciones en que se aplica el instrumento, esto significa que a través de la TG se pueden identificar las fuentes de error que afectan los resultados de las mediciones.

Los presupuestos de la TG solucionan algunos de los principales problemas que presenta la TCT. En primer lugar resuelve el problema de la concepción unitaria e indiferenciada del error de medida, ya que la TCT concibe el error de manera global y unitaria como consecuencia de una combinación de diferentes fuentes no diferenciadas.

Cambia el concepto limitado de confiabilidad, por uno más amplio de generalización, es decir, permite realizar inferencia estadística a una población a partir de una puntuación observada.

Reemplaza el postulado de medidas paralelas por el de medidas aleatoriamente paralelas, al considerar que las diferentes condiciones de un procedimiento de evaluación son una muestra aleatoria de un universo más amplio.

Sin embargo, la TG no logra solucionar dos problemas importantes que presenta la TCT, ya que en ambos planteamientos las características del instrumento dependen de la muestra, y las medidas de los sujetos dependen del tipo de prueba utilizada.

Los modelos de medición psicológica y educativa denominados genéricamente Teoría de Respuesta a los Items (TRI), poseen ventajas sobre los Test clásicos. La TRI presenta un planteamiento nuevo, ya que la TG puede considerarse más como una extensión de la TCT.

La TRI partiendo de supuestos o hipótesis fuertes, intenta dar una fundamentación probabilística al problema de la medición de rasgos y constructos no observables, liberando a la TCT de los problemas antes mencionados (Martínez, 1995). Como señala Lord (1980), citado por Muñiz (1990), la TRI no contradice ni los supuestos ni las conclusiones fundamentales de la TCT, si no que hace asunciones adicionales que permiten responder los interrogantes que esta última no soluciona. El nombre Teoría de Respuesta a los Items proviene de que el ítem es la unidad básica de análisis en lugar de las puntuaciones totales o globales de la prueba como sucede en la TCT. La TRI también ha sido denominada teoría del Rasgo Latente o Teoría de la Curva característica del ïtem (Anastasi, 1997).

Como se señaló anteriormente, la TCT y la TG se basan en la distribución normal, la TRI utiliza diferentes funciones matemáticas basadas en diversos grupos de suposiciones, algunos modelos usan la función de ojiva normal (distribución normal acumulada) y otros usan la función logística (distribución basada en propiedades logarítmicas) en general, los resultados obtenidos con los diferentes modelos son sustancialmente similares (Anastasi, 1997).

Los modelos más usados en la TRI se fundamentan en la ojiva logística o función de distribución logística, que representan una familia de curvas cuya forma es similar en apariencia a la ojiva normal. Estas funciones fueron utilizadas como modelo para el estudio de desarrollo de plantas y animales desde el nacimiento hasta la madurez. Los especialistas en biomatemáticas usaron la distribución logística para estudiar las tasas de crecimiento y mortalidad en los años 20. Estas aplicaciones biométricas fueron retomadas por Bradley y Terry en los años 50, y a partir de ahí Birnbaum (1968)

y Baker (1961) formularon un modelo logístico pero en análisis de ítems. Baker desarrolló programas para computador que permitían la aplicación del análisis logit y probit de los ítems y estudió su comportamiento con datos empíricos y simulados (Baker, 1959, 1963, citado por Wright y Stone, 1998).

En el campo de la medición estas distribuciones se expresan en funciones que al ser graficadas presentan forma de S. Los modelos de TRI asumen que existe una relación funcional entre los valores de la variable que miden los ítems y la probabilidad de acertar estos, denominando dicha función Curva Característica de los ítems (CCI). Expresado en otras palabras, esto significa que la probabilidad de acertar un ítem sólo depende de los valores de la variable medida por el ítem; por tanto personas que tengan diferentes puntuaciones en dicha variable tendrán probabilidades distintas de superar determinado ítem (Muñiz, 1990). La característica de cada ítem puede describirse por su CCI, siendo esta curva la unidad conceptual básica de la TRI.

Es importante aclarar que la CCI no es la regresión ítem-test, ya que en este tipo de curvas la variable que miden los ítems no es la puntuación que las personas obtienen en las pruebas, puesto que los valores de q están comprendidos entre menos infinito y más infinito ($-\mu$, $+\mu$) mientras que los de un test suelen estar entre 0 (cero) y la puntuación máxima posible en ese test.

La formula general de la función logística es:

$$P = \frac{e^x}{1 + e^x}$$

donde *e* es la base de los logaritmos neperianos y *x* cualquier valor o función.

Para definir adecuadamente la CCI es necesario tener en cuenta tres parámetros denominados *a, b* y *c*.

El parámetro *a* corresponde al índice de discriminación, al parámetro *b* se le denomina índice de dificultad y el parámetro *c* representa la probabilidad de acertar el ítem al azar, cuando no se conoce la respuesta. Según el tipo de función matemática adoptada y el valor de los parámetros, se tendrán diferentes modelos de CCI.

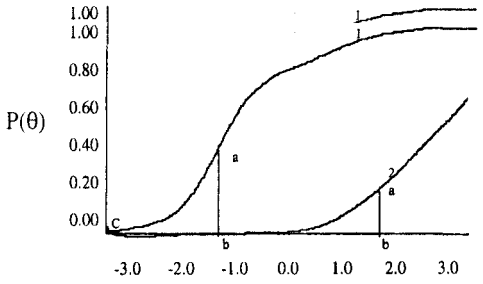

Figura 1. Calibración de los participantes y de los ítems a lo largo de la variable.

La figura 1 muestra dos CCI. Los cambios de la variable se representan en la abscisa o eje x en valores logit y en la ordenada o eje y se representa P (q) que indica la probabilidad de éxito del ítem.

El parámetro a o índice de discriminación representa la derivada de la función o valor de la pendiente de la recta tangente a la curva en el punto de máxima pendiente. Cuanto mayor sea la pendiente mayor será el índice de discriminación. El valor de a cuando q se distribuye según la curva normal con media 0 y desviación típica 1 y no hay aciertos al azar (c=0), viene dado aproximadamente por la expresión

$$a \equiv \frac{r_b}{\sqrt{[\,1-(r_b)^2\,]}}$$

donde r_b es la correlación biserial ítem-test, o sea el índice de discriminación en el modelo clásico.

El parámetro *b* o índice de dificultad es el valor de q correspondiente al punto de máxima pendiente de la CCI. Como en el caso de *a*, tampoco el significado de *b* es exactamente el mismo que en la teoría clásica aunque, se refiere a la dificultad del ítem. Aquí la dificultad del ítem se mide en la misma escala que q. Si se mantienen las condiciones de normalidad de q, el parámetro *b* se relaciona con el índice de dificultad de la teoría clásica aproximadamente según la siguiente expresión.

$$b \equiv \frac{-Z_p}{r_b}$$

donde Z_p es la puntuación típica que corresponde en la curva normal a la proporción de aciertos en el ítem (índice de dificultad en la teoría clásica) y r_b es la correlación biserial ítem-test.

El parámetro *c* representa la probabilidad de acertar el ítem al azar cuando no se sabe nada, es decir, es el valor de $P(\theta)$ cuando $\theta = -\infty$.

A medida que las CCI se ubican más a la derecha en el eje de abscisas el nivel del ítem incrementa pues *b* aumenta. En la figura 1 es más fácil el ítem 1 que el ítem 2. El poder discriminativo *a* viene indicado por la pendiente de las CCI; el ítem 1 tiene más poder discriminativo que el ítem 2. El parámetro *c*, aciertos al azar, es 0 para los dos ítems.

La curva característica de los ítems presenta un resumen gráfico del análisis de la dificultad, poder de discriminación y probabilidad de adivinar el item (Murphy y Davidshofer, 1994).

La TRI, a pesar de no ser un modelo reciente, ha venido tomando importancia en el campo de la psicometría, especialmente en la última década del siglo XX. En el campo nacional, el Instituto Colombiano para el Fomento de la Educación Superior ICFES, se basó en este modelo, a partir del año 2000, para analizar el nuevo examen de estado. La interpretación y el uso de los puntajes cambiaron sustancialmente en comparación con los resultados obtenidos en el anterior tipo de examen. El puntaje obtenido en el nuevo examen es la expresión del valor logit de cada persona que responda las pruebas, en una escala que oscila entre cero (0) y cien (100) puntos, aproximadamente. Una puntuación de cero (0) equivale a no responder ninguna pregunta correctamente. Se considera que la parte superior de la escala es abierta; es decir cien (100) puntos es apenas una aproximación y una referencia de la máxima competencia que puede obtener un estudiante (ICFES, 2000). Sin embargo, un estudiante con alto nivel de competencia puede obtener puntajes muy superiores a cien (100). Debido a que la escala logit de donde se obtiene el puntaje, es independiente de los resultados de los diferentes grupos de personas, es decir no se construye a partir de promedios o desviaciones estándar de alguna población particular, permite comparaciones directas en el tiempo, contrastando las ejecuciones de los estudiantes contra lo que mide una prueba en particular, tomando una primera aplicación como referencia (marzo del año 2000). En términos generales el puntaje refleja la competencia global del estudiante en el área medida (ICFES, 2000).

EJEMPLO DE APLICACIÓN BAJO LA TEORÍA DE RESPUESTA AL ITEM

A continuación se presenta un ejemplo práctico del análisis de una prueba de conocimiento bajo la TRI. La base de datos es hipotética y se destacaron los resultados más relevantes para establecer las características psicométricas de los ítems bajo este modelo.

La tabla 1 presenta la distribución de las respuestas de 10 personas en una prueba de conocimientos de 15 ítems. El máximo puntaje posible es 15, debido a que las respuestas correctas son calificadas con 1 y las incorrectas con 0. Las calificaciones oscilaron entre 0 y 15; Nina obtuvo 0 puntos, porque todas sus respuestas fueron incorrectas y José puntuó 15 ya que respondió correctamente toda la prueba. La pregunta número 01 fue acertada por 9 de los 10 participantes, la pregunta número 15 fue contestada correctamente solo por 2 personas. Las preguntas 11, 12, 13 y 14 fueron respondidas acertadamente por 3 personas. Se resaltan en negrilla las respuestas correctas que corresponden a las dadas por José.

La figura 2 representa uno de los resultados más importantes que arroja el programa Winsteps, utilizado para realizar el análisis de ítems bajo la TRI. En ésta se pueden identificar los siguientes aspectos: la primera columna presenta los puntajes logit, que para el caso indica que los 15 ítems de la prueba se distribuyeron entre –2 y +3 desviaciones logit. Al lado de la escala logit, se encuentran los nombres de los participantes; José aparece en la parte superior del gráfico, ya que fue la persona con más alta puntuación, seguido de Diana, Ana y Lola. Sara y Omar se encuentran en el mismo nivel, seguidos de Emma, Pepe y Luis. En la parte inferior del gráfico se encuentra Nina, quien fue la persona que obtuvo el puntaje más bajo.

La línea punteada tiene tanto a la izquierda como a la derecha las letras M, S y T. La M indica la media, la S una desviación y la T dos desviaciones, todas en escala logit. Las letras del lado izquierdo corresponden a la distribución de las personas y las del lado derecho a la distribución de los ítems.

Tabla 1. Base de datos prueba de conocimientos

Sujetos	01	02	03	04	05	06	07	08	09	10	11	12	13	14	15	Total
PEPE	B	B	C	D	W	W	A	B	C	D	A	B	C	D	A	5
LOLA	B	A	B	D	C	C	A	B	D	A	C	C	A	W	A	10
OMAR	B	D	C	C	D	B	C	A	D	A	B	D	C	B	D	7
ANA	B	A	B	D	C	C	A	B	D	A	D	C	W	W	C	10
EMA	B	A	B	C	A	C	A	B	C	B	D	C	A	W	B	6
SARA	B	A	B	C	A	C	A	B	C	B	D	C	A	A	A	7
LUIS	B	B	B	D	C	A	C	D	W	W	W	W	W	B	B	4
JOSE	**B**	**A**	**B**	**D**	**C**	**C**	**A**	**B**	**D**	**A**	**B**	**D**	**C**	**A**	**D**	**15**
NINA	A	B	C	C	A	A	B	D	A	B	C	C	A	D	W	0
DIANA	B	A	B	D	C	C	A	B	D	A	B	D	A	A	C	13
Total	9	6	7	6	5	6	7	7	5	5	3	3	3	3	2	

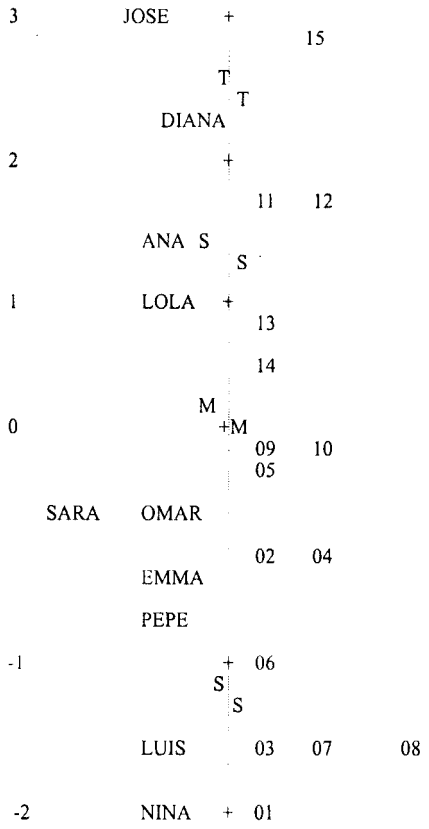

Figura 2. Calibración de los participantes y de los ítems a lo largo de la variable

Por último, aparece la numeración de los ítems con la calibración de la dificultad a lo largo de la variable. El ítem número 15 es el más difícil, seguido por los ítems 11 y 12 que se encuentran al mismo nivel, les sigue en dificultad los ítems 13 y 14. Aparecen después, los ítems 09 y 10 (con igual nivel de dificultad), luego el 05, seguido por las preguntas 02 y 04 (igual dificultad), a continuación aparece el 06, continuando con los ítems 03, 07 y 08 (que tienen la misma dificultad), y por último se encuentra el ítem 01 que es el más fácil de todos.

Si las medias y las desviaciones de personas e ítems se corresponden en la figura, esto significa que el modelo se ajusta perfectamente. En el ejemplo se puede considerar que hay un buen ajuste.

El propósito de la medición es lograr que el instrumento evalué todo el continuo de la variable, los espacios entre ítems en la figura indican una pobre definición o evaluación en varias regiones de la variable. Por lo tanto, es necesario elaborar más ítems que permitan medir las porciones de la variable que no están siendo evaluadas. Así por ejemplo, deben diseñarse preguntas que

evalúen el espacio entre el ítem 15 y los ítems 11 y 12.

Otro aspecto importante a tener en cuenta en el análisis de la figura 1, está relacionado con la discriminación de los ítems respecto a la ejecución de los sujetos; las preguntas 03, 07 y 08 son las que mejor discriminan la ejecución de Luis.

Por otro lado se puede afirmar que las preguntas 14, 09, 10 y 05 tienen una dificultad promedio y que en términos generales el 50% de los evaluados se encuentran entre –1 y +1 desviación logit. La figura 1 muestra que entre –1 y +1 desviación logit están Lola, Sara, Omar , Emma y Pepe. José se encuentra a +3 desviaciones logit y Nina a -2 desviaciones logit. Respecto a las preguntas se puede afirmar, que el ítem número 15 se encuentra entre +2 y +3 desviaciones logit y la pregunta número 01 se encuentra a –2 desviaciones logit. Entre –1 y +1 desviación logit están localizadas 8 preguntas.

En los resultados que aparecen en la tabla 2, se presenta el resumen de los estadísticos que arrojó el modelo de las personas con respecto al instrumento, la confiabilidad es de 0.58, lo que indica un nivel aceptable de precisión en la información dada por las personas. El nivel de ajuste para los puntajes por medio del método de mínimos cuadrados estandarizados (MNSQ) alrededor de la media es de 0.93 y en los extremos es de 1.0, esto indica que el modelo predice que los puntajes de los participantes se encuentran dentro del rango de ajuste que está entre 0.7 y 1.3 en puntaje logit, es decir la muestra se encuentra bien distribuida según los parámetros que plantea el modelo. Los puntajes máximo y mínimo en desviaciones logit confirman el análisis realizado en la figura 1, mostrando con precisión los valores logit superior e inferior, que para el ejemplo indican que Diana se encuentra exactamente a 2.25 desviaciones logit por encima de la media y Luis a 1.33 desviaciones logit por debajo de la media. Es importante anotar que los datos extremos (José y Nina) no son tenidos en cuenta para el análisis, ya que no contribuyen al ajuste en el modelo (El programa Winsteps no los incluye).

En los resultados que aparecen en la tabla 3, se presenta el resumen de los estadísticos que arrojó el modelo de los ítems con respecto al instrumento; la confiabilidad de la totalidad de la prueba es de 0.23,

Tabla 2. Resumen general de resultados de los puntajes obtenidos por las personas

	Media puntaje Logit	Error del modelo	Ajuste a la media MNSQ	Ajuste a los extremos MNSQ
Media	0.10	0.70	0.93	1.0
Desviación estándar	1.17	0.08	0.60	1.08
Puntaje Max.	2.25	0.85	2.39	3.75
Puntaje Min.	-1.33	.60	0.39	0.28
Indice de confiabilidad	Indice real	0.58		

Tabla 3. Resumen general de resultados de los puntajes obtenidos en los ítems

	Media puntaje Logit	Error del modelo	Ajuste a la media MNSQ	Ajuste a los extremos MNSQ
Media	0.00	0.92	1.03	1.03
Desviación estándar	1.16	0.10	0.43	0.68
Puntaje Max.	2.56	1.17	2.33	2.66
Puntaje Min.	-1.32	0.81	0.63	0.54
Indice de confiabilidad	Indice real	0.23		

reflejando un bajo nivel de precisión en la medición del atributo. El nivel de ajuste para los puntajes por medio del método de mínimos cuadrados estandarizados (MNSQ) alrededor de la media es de 1.03 y este mismo valor se obtuvo para los extremos, esto indica que el modelo predice que los puntajes de los ítems se encuentran dentro del rango de ajuste que está entre 0.7 y 1.3 en puntaje logit, es decir las preguntas se encuentran bien distribuidas según los parámetros que plantea el modelo, aunque como se anotó anteriormente, el nivel de precisión es bajo. Los puntajes máximo y mínimo en desviaciones logit se relacionan con el análisis realizado en la figura 1, mostrando con precisión los valores logit superior e inferior, que para el caso indica que el ítem 15 se encuentra a 2.56 desviaciones logit por encima de la media y los ítems 03, 07 y 08 a -1.32 desviaciones logit por debajo de la media. Es importante anotar que el programa no tuvo en cuenta el ítem 01 para el análisis, ya que no contribuye al ajuste en el modelo.

Antes de estimar la medida de una persona a partir de su puntaje, se debe examinar su patrón de respuestas. Se debe observar si su patrón es uniforme con la manera en la que se espera que los ítems susciten respuestas. Cuando los ítems con los cuales una persona es examinada han sido calibrados a lo largo de una variable de fáciles a difíciles, como en el ejemplo planteado, entonces se presume que el patrón de respuestas de los evaluados sea más o menos consistente con el orden de dificultad de esos ítems a lo largo de la variable. Se espera que los evaluados tengan éxito en los ítems fáciles y fallen en los difíciles (Wright y Stone, 1998). El análisis del patrón de respuestas bajo la TRI se obtiene a través de un escalograma de Guttman (arrojado por el programa Winsteps), el cual se presenta en la tabla 4, que corresponde al patrón de respuestas de los 15 evaluados presentados en el ejemplo.

En la parte superior del escalograma se encuentran las preguntas, ordenadas de la más fácil a la más difícil, en la columna de la izquierda los evaluados ordenados de mejor ejecución a más baja ejecución y en el cuerpo de la figura aparecen los símbolos 1, 0, @, y A. El número 1 indica una respuesta correcta que se esperaba fuera correcta. El número 0 indica una respuesta incorrecta que debió ser incorrecta. El símbolo @ corresponde a una respuesta

Tabla 4. Escalograma de Guttman que representa el patrón de respuestas dado por las personas

Sujetos	Número de la pregunta														
	01	03	07	08	06	02	04	05	09	10	14	13	11	12	15
José	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
Diana	1	1	1	1	1	1	1	1	1	1	1	@	1	1	0
Ana	1	1	1	1	1	1	1	1	1	1			0	0	0
Lola	1	1	1	1	1	1	1	1	1	1		@	0	0	0
Omar	1	@	@	@	@	@	@	0	A	A	0	A	A	A	A
Sara	1	1	1	1	1	1	@	0	0	0	A	0	0	0	0
Emma	1	1	1	1	1	A	0	0	0	0		0	0	0	0
Pepe	1	@	1	1		0	A		0	0	0	A	0	0	0
Luis	1	A	0	0	0	0	A	A			0				0
Nina	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0

incorrecta que se esperaba fuera correcta y la letra A corresponde a una respuesta correcta que debió ser incorrecta. Los espacios en blanco indican las preguntas que no tuvieron respuesta por parte de los evaluados.

En la primera fila aparece la descripción del patrón de respuestas de José, está fila solo tiene unos y estos resultados son consistentes con el orden de dificultad de los ítems. En segundo lugar aparece la descripción de las respuestas dadas por Diana, quien fue la que siguió en puntaje a José (13 respuestas correctas), en este patrón se observa una inconsistencia señalada con el símbolo @, ya que la respuesta dada en la pregunta número 13 fue incorrecta pero debió ser correcta. En el caso de Omar se observan varias inconsistencias, ya que él contestó correctamente las preguntas 09, 10, 11, 12, 13 y 15, que eran más difíciles que las 02, 03, 04, 05, 06, 07 y 08. Por esto, aparecen los símbolos @ y A que indican que las primeras preguntas debieron ser contestadas correctamente y en las últimas debió fallar. Los espacios en la fila que representa la ejecución de Luis muestran que él dejó de responder las preguntas 09, 10, 11, 12 y 13. El patrón de respuesta de Nina indica que sus resultados son consistentes con su nivel de ejecución.

Solo se presentaron y analizaron 3 de las 22 tablas que arroja el programa Winsteps, utilizado para el procesamiento de la información bajo la TRI, y basado específicamente en el modelo de un parámetro (índice de dificultad) propuesto por el matemático Danés Georg Rasch.

La IRT se proyecta para el siglo XXI como el modelo para sustentar psicométricamente la información arrojada por las pruebas, no solo de conocimientos, si no también de cualquier instrumento que mida variables psicológicas.

La TRI es una sofisticada y poderosa alternativa sobre los métodos tradicionales para el análisis de items y más que establecer la ejecución de las personas en la prueba o de comparar los items con el test, la teoría define la relación entre la habilidad para medir el atributo y la construcción individual de los items del test, esto abre nuevos caminos en la investigación psicométrica.

REFERENCIAS

Aiken, L. (1996). *Tests Psicológicos y Evaluación*. México: Prentice Hall.

Anastasi, A. y Urbina, S. (1997). *Psychological Testing*. New Jersey: Prentice Hall.

Brown, F. (1980). *Principios de Medición en Psicología y Educación*. México: Manual Moderno.

ICFES (2000). *Admisión a la Educación Superior: Algunos Temas de Discusión*. Bogotá.

Kerlinger, F. (1988). *Investigación del Comportamiento*. México: McGraw-Hill.

Linacre, J. y Wright, B. (1999). A User's Guide to Winsteps. Chicago: MESA Press.

Martínez, R. (1995). *Psicometría: Teoría de los Tests Psicológicos Educativos*. México: Editorial Síntesis Psicológica.

Medellín, E. (2001). *Fundamentos de la Medición en Psicología. En Construcción de Pruebas Objetivas para la Evaluación de Conocimientos en el Aula*. Revista Aula Psicológica, Volumen 2 páginas 129 a 171. Bogotá: Facultad de Psicología, Universidad El Bosque.

Muñiz, J. (1990). *Teoría de Respuesta a los Ítems*. Madrid: Ediciones Pirámide S.A.

Muñiz, J. (1994). *Teoría Clásica de los Tests*. Madrid: Ediciones Pirámide S.A.

Muñiz, J. (1998). *Teoría Clásica de los Tests*. Madrid: Ediciones Pirámide S.A.

Murphy, K; Davidshofer, C. (1994) *Psycological Testing*. New Jersey: Prentice Hall.

Nunnally, J. y Bernstein, I. (1995). *Teoría Psicométrica*. México: Editorial McGraw-Hill

métrica. México: Editorial McGraw-Hill.

Pardo, C. (2000). *Transformaciones en las Pruebas para Obtener Resultados Diferentes*. Bogotá: ICFES.

Rocha, M. y Pardo, C. (2000). *Admisión a la Educación Superior*. Bogotá: ICFES.

Wrigth, B. y Stone, M. (1998). *Diseño de Mejores Pruebas*. México: CENEVAL.