Teoría de Respuesta al Ítem
Conceptos básicos y aplicaciones para la medición de constructos psicológicos

Horacio Félix Attorresi,
Gabriela Susana Lozzia,
Facundo Juan Pablo Abal,
María Silvia Galibert
*y María Ester Aguerri**

Resumen

El objetivo de este trabajo es introducir al lector en las nociones básicas de la Teoría de Respuesta al Ítem (TRI). La Primera Parte describe las limitaciones de la perspectiva clásica y define los conceptos fundamentales de la TRI: Curva Característica, Parámetros, Función de Información, Estimación y Ajuste de los modelos. La Segunda Parte revisa desarrollos aplicados efectuados a partir de los nuevos avances de esta teoría. Se trata el análisis del funcionamiento diferencial de los ítems y la construcción de bancos de ítems y tests adaptativos informatizados. Se reconoce la necesidad de una formación más sólida sobre la TRI en los especialistas en Evaluación Psicológica de Argentina.

Palabras clave: *teoría de Respuesta al ítem, psicometría, modelos psicométricos.*
Key words: *item response theory, psychometrics, psychometric models.*

Los tests son procedimientos de recolección de información sobre un individuo o grupo utilizados habitualmente en Evaluación Psicológica. La construcción de los mismos se basa en modelos psicométricos que permiten evaluar la calidad de la medida y dar garantías de la misma. En el campo de la Psicometría, la Teoría de los Tests constituye el marco de referencia teórico y metodológico que reúne los modelos que subyacen a la elaboración y uso de tests (Muñiz, 1997).

Los modelos que componen la Teoría de los Tests formalizan las interrelaciones de tres componentes que intervienen en la medición mediante tests: a) la puntuación observada tras la administración del test (el puntaje total en un test o la respuesta de un individuo a un ítem), b) un valor inobservable del dominio o rasgo psicológico que se pretende medir y; c) el error de medida que conlleva todo proceso de medición.

Desde su primera formulación (Spearman, 1904), la Teoría Clásica de los Tests (TCT) ha servido como modelo para dar una interpretación a los puntajes de las personas en los tests. A pesar de su expansión y vigencia, la literatura sobre Teoría de los Tests de los últimos cuarenta años ha registrado un desplazamiento gradual hacia teorías y técnicas de medición psicológica superadoras de la perspectiva clásica. Aunque se han propuesto numerosos modelos a lo largo de la historia de la Psicometría, la Teoría de Respuesta al Ítem (TRI) es el desarrollo más reconocido (Hambleton & Swaminathan, 1985; Lord, 1980; Martínez-Arias, 1995; Muñiz, 1997).

Quizás es conveniente resaltar la idea de que tanto la TCT como la TRI persiguen el mismo objetivo: estimar el error que se comete al intentar medir un fenómeno psicológico específico. Son construcciones teóricas (con menor o mayor grado de complejidad y profundidad) respecto de un mismo hecho. Ambas

* *Horacio Félix Attorresi, Gabriela Susana Lozzia, Facundo Juan Pablo Abal, María Silvia Galibert y María Ester Aguerri.*
Instituto de Investigaciones de la Facultad de Psicología, UBA.
Rivera Indarte 132, 1°"A", (1406), Ciudad Autónoma de Buenos Aires.
E-Mail: hattorre@psi.uba.ar
REVISTA ARGENTINA DE CLÍNICA PSICOLÓGICA XVIII p.p. 179-188

teorías plantean un modelo y un conjunto de supuestos, que si se cumplen, garantizan la precisión de la medida. También cabe destacar que la coexistencia de ambas teorías no implica su incompatibilidad. Lejos de competir, estos modelos se complementan en la práctica psicométrica para realizar un análisis más profundo y exhaustivo de la calidad y/o del funcionamiento del test.

El objetivo de este trabajo es definir brevemente los conceptos básicos de la TRI para introducir en las características más importantes de este enfoque psicométrico. Por este motivo, se han pensado dos partes para organizar este texto. La Primera Parte recorre a nivel teórico las nociones fundamentales de la TRI. La Segunda Parte revisa un conjunto de desarrollos efectuados a partir de los nuevos avances de esta teoría y, al mismo tiempo, muestra ejemplos para que el lector interesado conozca el alcance potencial de estas aplicaciones.

Desarrollo

Primera parte. Conceptos elementales para comprender la TRI.

Teoría Clásica de Tests. Generalidades y limitaciones.

La TCT surgió del modelo lineal de puntuaciones formulado por Spearman (1904) y alcanzó su formalización más precisa en la obra de Novick (1966). Su formulación matemática es bastante simple y supone que el puntaje observado de un sujeto en un test es el resultado de la suma del valor real (puntaje verdadero) y el error de medición.

La propuesta de Spearman se asemeja a la forma de medir en las ciencias duras. Si se realizan múltiples mediciones de una longitud, la mejor estimación de la "longitud verdadera" resultará del promedio de todas las observaciones efectuadas. Si cada una de las mediciones que se ejecutan son independientes y el error de medida en cada una de ellas es aleatorio, el promedio de éste tenderá a cero (dado que las observaciones que subestimaron a la puntuación verdadera se cancelarían con aquéllas que la sobreestimaron). De esta forma, tras infinitas mediciones, el promedio de las puntuaciones empíricas obtenidas podría considerarse igual a la verdadera.

Pero las críticas al modelo y sus supuestos surgieron tempranamente (Thurstone, 1928):

1) El resultado obtenido al medir una variable es inseparable del test usado, lo cual sería como pensar que el peso de un objeto depende de la balanza que se utilice. Si se mide una variable psicológica con dos tests diferentes (que miden el mismo constructo) las puntuaciones obtenidas no son estrictamente equiparables, dado que no se encuentran en la misma escala.

2) Las propiedades de los ítems y del test están determinadas por las características de los examinados. Esto implica que aquello que se está midiendo afecta al instrumento utilizado para medir. Sería como pensar que un kilo de acero puede pesar distinto a un kilo de plumas.

Estas limitaciones de la TCT fomentaron la aparición de nuevas teorías de medición psicológica. En la década del 60, surgieron los primeros desarrollos de la TRI, un enfoque emergente del campo educativo que se propuso profundizar el estudio de las propiedades psicométricas de los ítems y de los tests.

Teoría de Respuesta al Ítem

La denominación TRI agrupa líneas de investigación psicométricas independientes iniciadas por Rasch (1960) y Birnbaum (1968). El factor común de estos desarrollos es que establecen una relación entre el comportamiento de un sujeto frente a un ítem y el rasgo responsable de esta conducta (*rasgo latente*). Para ello, recurren a funciones matemáticas que describen la probabilidad de dar una determinada respuesta al ítem para cada nivel del rasgo medido por este.

El objetivo sustancial de la TRI es la construcción de instrumentos de medición con propiedades invariantes entre poblaciones. Si dos individuos presentan idéntico nivel de rasgo medido ambos tendrán igual probabilidad de dar la misma respuesta, independientemente de la población de pertenencia. Esto conlleva un gran beneficio respecto de la TCT en tanto que es posible evitar el uso de un grupo normativo.

Mientras que en la TCT se modeliza sobre el puntaje verdadero en una prueba particular, en la TRI se toma al ítem como unidad de análisis y se modeliza directamente sobre el rasgo latente. El nivel de rasgo latente que presenta un individuo es fruto de una estimación a partir del patrón de respuestas manifestado en un conjunto de ítems. Si se varía el conjunto de ítems utilizado se mantiene la puntuación estimada aunque eventualmente hayan cambiado las propiedades psicométricas de los reactivos. Por lo tanto, la TRI permite mediciones invariantes más allá de los ítems que componen el instrumento.

Hambleton y Swaminathan (1985) también rescatan la importancia de las medidas locales de precisión que proporciona la TRI. Desde la TCT se indica la fiabilidad como un valor global y constante para todos los niveles del rasgo. Sin embargo, se sabe que los tests suelen ser más precisos para discriminar en un determinado rango de la variable y menos en otros.

La TRI provee información respecto del grado de exactitud con que se mide la variable en función de sus diferentes niveles. Estas medidas de precisión locales se hacen operativas mediante las Funciones de Información de los Ítems y del Test desarrolladas por Birnbaum (1968).

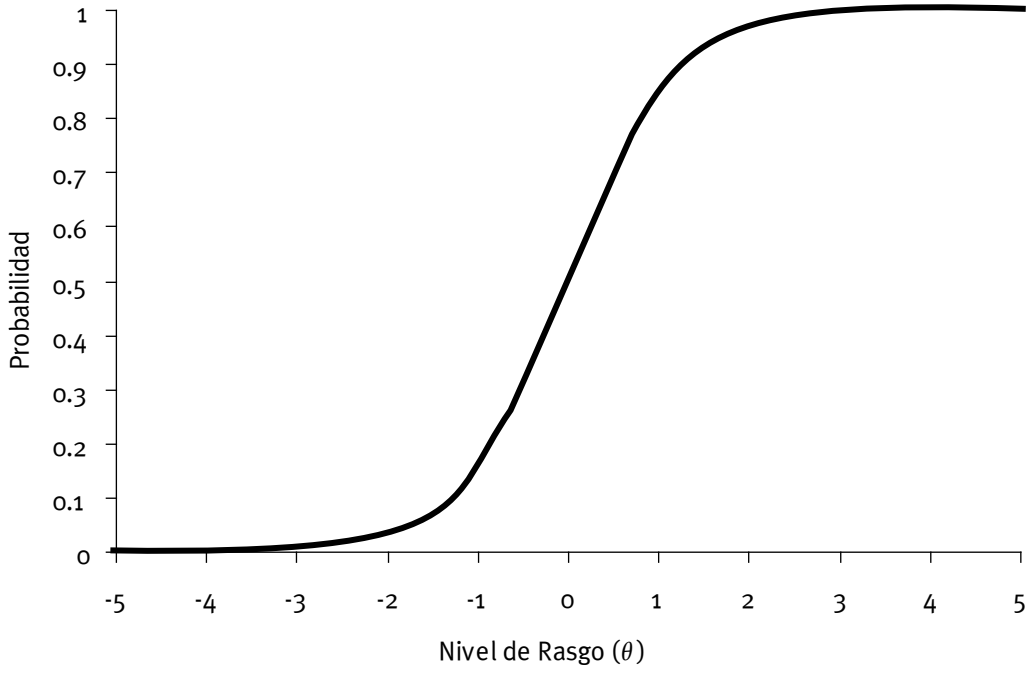
Todas estas características básicas de la TRI son las que ayudan a encontrar respuestas a los principales inconvenientes observados en la TCT. Pero obtener estas garantías de precisión en la medición de un constructo no es una tarea simple para el investigador. La fuerza de esta teoría se sostiene en un conjunto de supuestos exigentes a los que la mayoría de los datos empíricos difícilmente se acomodan, y que por ende, condicionan su aplicabilidad. Los siguientes apartados describen los conceptos teóricos elementales necesarios para comprender la TRI.

Curva Característica del Ítem y sus parámetros. En el marco de la TRI se postula la existencia de una relación directa entre el comportamiento de un individuo frente a un ítem y el rasgo que genera esta conducta. La formalización de esta relación adopta la forma de una función matemática que vincula la probabilidad de dar una determinada respuesta (opción-clave o respuesta clave) a un ítem con cada nivel del rasgo latente. En los ítems de habilidades la clave es la respuesta correcta y en los de personalidad es aquélla opción que indica la presencia de un nivel mayor de rasgo en el individuo. El gráfico de dicha función matemática se denomina *Curva Característica del Ítem (CCI)*.

Si se supone que el rasgo medido es una habilidad, de un modo intuitivo se podría conjeturar que cuanto más hábil es la persona mayor probabilidad tendrá de contestar correctamente el ítem que mide esa habilidad. De igual manera podría entenderse la respuesta a un ítem dicotómico de un test de personalidad. Por ejemplo, cuanto más alto sea el nivel de Extraversión de un individuo, mayor probabilidad tendrá de dar una respuesta afirmativa al ítem *Soy una persona conversadora*. Esta descripción racional de la respuesta a un ítem ayuda a entender que la curva de la CCI deberá ser siempre creciente.

Dentro de la TRI existen distintas familias de modelos matemáticos probabilísticos que cumplen con este requisito de monotonía creciente, pero sólo dos fueron empleados con mayor frecuencia como base para la CCI: las funciones de distribución de probabilidad normal acumulada y la logística. Ambas curvas adoptan una forma de *S* suavizada con un punto de inflexión en el que tienen su máxima pendiente. Presentan dos asíntotas horizontales (superior a derecha e inferior a izquierda). La superior corta al eje vertical en el valor de máxima probabilidad (i.e. igual a 1) mientras que la inferior lo corta en un valor comprendido entre 0 y 1 (ver Figura 1).

Figura 1. **Curva Característica de un Ítem con parámetros a = 1, b = 0 y c =0**



Además del tipo de función adoptada (normal o logística), la CCI de cada ítem queda determinada cuando se especifican los parámetros que describen al reactivo. Estos parámetros se refieren a la localización del ítem en la escala del rasgo medido (*b*), a su potencia discriminatoria (*a*) y la probabilidad de dar la respuesta clave para un nivel muy bajo de rasgo (*c*). Por lo tanto, cada ítem tendrá una CCI propia según los valores que adopten estas propiedades psicométricas.

Supuestos. La aplicación de los modelos se asienta sobre conjunto de hipótesis que deben suponerse de partida. Asumir que la conducta de una persona ante un ítem sigue un modelo probabilístico con una forma determinada (CCI) es en sí mismo un supuesto de la TRI.

La cantidad de rasgos latentes que intervienen al momento de contestar a un ítem establece una importante distinción clasificatoria de los modelos. Si se supone que la respuesta al ítem está en función de un único rasgo los modelos son denominados unidimensionales. En la práctica es difícil que este supuesto de *unidimensionalidad* del rasgo latente se satisfaga totalmente ya que múltiples factores pueden afectar a la respuesta a un ítem. Por este motivo, para los modelos unidimensionales sólo se exige un rasgo fundamental (factor dominante) que explique las respuestas de los sujetos. Así también, para situaciones en las que se contempla la incidencia de dos o más rasgos afectando en la respuesta al ítem se han generado modelos multidimensionales. No obstante, estos desarrollos son relativamente incipientes y su puesta en práctica está resultando verdaderamente compleja para los investigadores.

Otro supuesto de los modelos de la TRI es la independencia local de los ítems. Este requisito supone que conocido el nivel θ de un sujeto, las respuestas a cualquier subconjunto de ítems no agregan ninguna información para el cálculo de probabilidad de respuesta a un ítem en particular. Es decir, que las respuestas a distintos ítems son estadísticamente independientes. La unidimensionalidad del rasgo latente y la independencia local de los ítems son las dos caras de un mismo requerimiento. Esto es, si se confirma que dos ítems no son independientes significa que otro factor ajeno al que se pretende medir incide en la estimación del nivel del rasgo violando el supuesto de unidimensionalidad (Lord & Novick, 1968).

Modelos. La respuesta de una persona a los ítems obedece, por un lado, a la cantidad del rasgo que tiene dicha persona y, por otra parte, a las características propias del ítem que está contestando. Por ende, todos los modelos de la TRI vinculan el nivel del rasgo del evaluado con las propiedades psicométricas que describen al ítem y la probabilidad de optar entre las opciones del reactivo. Los tres parámetros de la CCI mencionados arriba se corresponden con los definidos en la primera generación de modelos de la TRI, los cuales son unidimensionales y suponen que la respuesta al ítem sólo admite dos opciones. Estos modelos dicotómicos se utilizan tanto para ítems que evalúan rendimiento o habilidad (correcto – incorrecto) o rasgos de personalidad (Acuerdo – Desacuerdo / Si – No).

De la combinación del tipo de función matemática adoptada para la CCI (logística o normal) y el número de parámetros considerados (uno, dos o tres) es posible definir seis modelos diferentes. Si un ítem es modelizado con el Modelo de Rasch, la CCI sólo se describe a partir del parámetro *b* y considera que el parámetro *a* es constante y el *c* es nulo para todos los ítems. El Modelo de Dos Parámetros (ML2p) contempla, además del *b*, el parámetro *a*, mientras que el *c* sigue considerándose nulo. Por último, el Modelo de Tres Parámetros (ML3p) utiliza los parámetros *b*, *a* e incorpora el parámetro *c*. A modo de ejemplo, a continuación se desarrollarán las particularidades del ML3p. Su formulación es la siguiente:

El símbolo θ corresponde al nivel del rasgo latente que se desea medir con el ítem i y Pi(θ) es la probabilidad de dar la respuesta clave al ítem i para un nivel dado de θ. Para el lector acostumbrado a trabajar con la TCT es importante aclarar que no se debe confundir el nivel θ con el puntaje total (observado) de una persona en un test. θ es el equivalente de la puntuación verdadera de la TCT. Esto significa que la puntuación total de una persona en un test (puntaje bruto) es una estimación de θ, de la misma manera que lo es de la puntuación verdadera en el marco de la TCT. La escala adoptada para medir θ tiene un rango teórico de -∞ a +∞ y su origen está determinado por consenso según la escala estandarizada, con media o y desvío típico 1.

bi es el índice de dificultad o parámetro de localización del ítem i. Coincide con el valor θ necesario para tener probabilidad 0,5 + ci / 2 de contestar la respuesta clave al ítem i. Un ítem tendrá mayor b que otro si se requiere de un mayor nivel de rasgo para tener la misma probabilidad de seleccionar la opción-clave. Al igual que el parámetro θ, puede oscilar entre -∞ y +∞ aunque, en la práctica, sus valores generalmente están dentro del intervalo (-4, 4). En el contexto de la medición de la personalidad se describe como el punto de transición (en la escala del rasgo) entre la probabilidad de tomar al enunciado del ítem como no descriptivo del evaluado y la de considerarlo como descriptivo (Richaud, 2005).

ai es el índice de discriminación del ítem i. Indica en qué medida el ítem diferencia a los examinados con un nivel en el rasgo por encima o por debajo del parámetro de localización. Se vincula con la pen-

diente de la CCI, cuanto más empinada sea la curva, mayor será el valor del parámetro a e indicará una mejor discriminación del ítem. El valor del parámetro a es siempre positivo. La capacidad discriminatoria se da para los valores de θ que están en torno al parámetro b; lo cual tendrá importantes consecuencias en la construcción de tests, pues según la zona de θ que sea de interés discriminar, se elegirán unos ítems u otros.

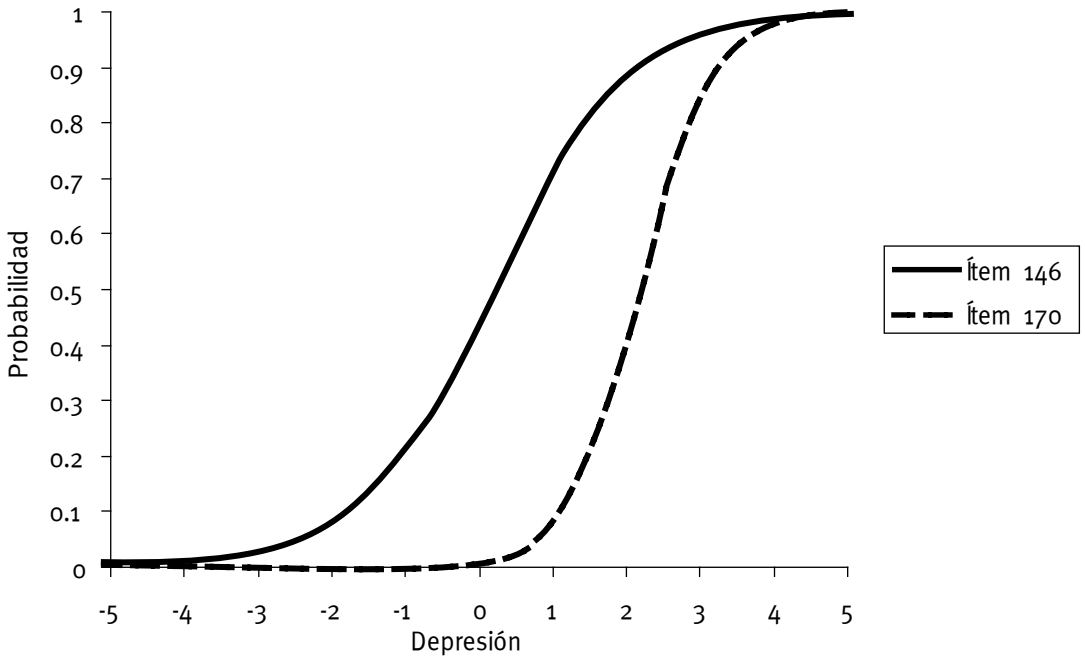
ci es el valor de la asíntota a izquierda, es decir cuando θ tiende a -∞. En ítems que miden habilidades refleja la probabilidad que tienen los individuos con muy bajo nivel de rasgo de responder correctamente. En tests de personalidad, algunos autores lo han interpretado como un indicador de la incidencia de la deseabilidad social (e.g. Rouse, Finger & Butcher, 1999) pero los resultados son bastante acotados como para generalizar dicha interpretación.

D es una constante de escalamiento. Cuando se adopta el valor 1.7 la función logística tiene una buena aproximación a la normal acumulada.

En la Figura 2 se observan las curvas características de dos ítems de la escala Depresión del MMPI-2 graficadas a partir de los resultados publicados por Childs, Dahlstrom, Kemp y Panter (2000). Estos autores aplicaron el ML2p, que utiliza los parámetros *b* y *a* para describir la CCI y supone nulo el parámetro *c*. Los parámetros del ítem 146 (*Lloro con facilidad*) estimados para las mujeres son: a_{146} = 0.66 y b_{146} = 0.19; mientras que los parámetros del ítem 170 (*Tengo miedo de estar perdiendo la razón*) son: a_{170} = 1.21 y b_{170} = 2.2. En relación con el parámetro *b* se observa que es necesario un nivel de Depresión (θ) más bajo para responder afirmativamente al ítem 146 que al ítem 170 ($b_{146} < b_{170}$), lo cual coincide con lo esperable clínicamente según los indicadores. Respecto del índice *a*, el ítem 170 discrimina mejor que el ítem 146. Esto puede observarse en que la pendiente del primero es más "empinada" que la del segundo ($a_{170} > a_{146}$). Esto significa que "llorar con facilidad" no es tan buen indicador de Depresión como "tener miedo de enloquecer". Pero hay que recordar que la capacidad discriminatoria se da para los valores de θ que están en torno al parámetro *b*; por tanto la adecuada interpretación sería que el ítem 146 tiene su mayor poder discriminativo y este es de magnitud moderada en torno a los niveles de Depresión intermedios (b_{146} = 0,19); mientras que el ítem 170 tiene su mayor poder discriminativo y este es de magnitud alta en torno a los valores de Depresión más elevados (b_{170} = 2.2). Cada ítem aporta información en diferentes niveles de la escala de Depresión.

La limitación que presentan los modelos de la primera generación es la dificultad de dar un tratamiento satisfactorio a ítems puntuados politómicamente. Al respecto, una nueva generación de modelos de la TRI ha hecho importantes contribuciones (Ostini & Nering, 2005). Los modelos politómicos de la TRI son útiles para el análisis de ítems con respuesta nominal (examen multiple choice) u ordinal (como las escalas tipo Likert). Si bien el desarrollo de los modelos politómicos excede los objetivos de esta introducción es importante destacar que no implican un nivel de complejidad mucho mayor al de los modelos dicotó-

Figura 2. **Comparación de la Curvas Características de dos Ítems**



micos. Los modelos politómicos no son más que una extensión de algunos modelos de la primera generación. Por ejemplo, el Modelo de Respuesta Graduada de Samejima (1969) constituye una generalización del ML2p a ítems de respuesta politómica ordenada. La clave de esta extensión reside en dividir conceptualmente la respuesta múltiple (nominal u ordenada) en una serie de variables dicotómicas.

Estimación y Ajuste del modelo. A partir de los patrones de respuestas observados en la muestra, en la TRI se estiman los parámetros de cada ítem y el valor del rasgo θ para cada sujeto. La estimación puede efectuarse siguiendo diversas estrategias y aplicando distintos métodos. El método de estimación más frecuentemente utilizado es el de Máxima Verosimilitud que consiste en asignar a θ y a los parámetros el valor que hace máxima la probabilidad de los datos observados. La noción que subyace a este procedimiento es análoga a la de un clínico que, tras indagar los síntomas durante la anamnesis, propone el diagnóstico que considera más probable en función de la sintomatología presente (Martínez-Arias, 1995).

Una vez realizada la estimación se requiere de una prueba de *bondad de ajuste* que permita evaluar hasta qué punto el modelo alcanzado representa bien a los datos observados. La evaluación del ajuste se basa en análisis de la discrepancia entre las CCIs teóricas (dadas por el modelo) y las CCIs empíricas. Estas últimas se construyen distribuyendo a los individuos en submuestras en función de su nivel de habilidad estimado y calculando las frecuencias de acierto observadas en los distintos intervalos del rasgo. Si el modelo se ajusta a los datos empíricos, se puede suponer que la curva característica representa de forma apropiada la relación entre el rasgo latente y la probabilidad de dar una determinada respuesta al ítem. En caso contrario, o el modelo no es el adecuado para relacionar el rasgo con la respuesta al ítem (se podría probar qué ocurre con otro modelo de la TRI) o el ítem no tiene un comportamiento que permita predecir el rasgo (se debería eliminar el ítem).

El ajuste está vinculado con los objetivos de invarianza de las mediciones y de las poblaciones (recuérdese las limitaciones de la TCT). La invarianza consiste en que las CCI quedan determinadas por sus propios parámetros independientemente de las de los demás ítems y de la distribución del rasgo latente en la población de individuos que sirvieron para estimarlas.

Función de Información. Además de la CCI, cada ítem aporta una Función de Información (FI) sobre el rasgo, la que indica para qué niveles del mismo el ítem proporciona mediciones más precisas. Como se podrá intuir, esta información también está aludiendo al error cometido. Por la fórmula con que se define la FI, a mayor valor de la misma, menor valor del error que se comete. Si un reactivo no es preciso para medir un nivel del rasgo es porque esa medida incluye un error demasiado elevado. La FI tiene una formulación bastante compleja que difícilmente podría explicarse en pocas palabras. Por esto se ha preferido resaltar su utilidad en el plano conceptual. Para un desarrollo más extenso pueden consultarse Muñiz (1997) y Martínez-Arias (1995).

La FI constituye un valioso instrumento para el análisis de los ítems ya que indica para qué valor de θ aporta más información el reactivo y cuál es la magnitud de la información aportada para dicho valor de θ. Si el rasgo que se está midiendo es una habilidad, la FI de un ítem difícil será mayor para niveles de habilidad altos y menor para niveles bajos; o sea que tal ítem será más útil para medir a los individuos más hábiles que para los menos hábiles. La FI es para la TRI lo que el concepto de confiabilidad es para la TCT; sólo que en la TRI un ítem no será más o menos confiable en términos absolutos sino para determinados niveles de la escala.

Lo expresado en relación a un ítem también puede ser aplicado al test. La FI de un test se define como la suma de las funciones de información de los ítems que lo componen. Por ejemplo, Childs et al. (2000) encontraron que los ítems de la escala Depresión del MMPI-2 arrojaban mayor información en los niveles relativamente altos de rasgo, este resultado es consistente con el propósito clínico que tiene el test.

Puntuación del individuo evaluado. El fin último de la administración de un test es asignar al evaluado una puntuación que refleje el nivel de rasgo que posee. En el marco de la TCT, el evaluador suma las respuestas codificadas numéricamente ya sea con una puntuación dicotómica o politómica y luego transforma el puntaje total (la puntuación observada) a una puntuación típica o percentilar. Desde la perspectiva de la TRI, la asignación de una puntuación a una persona es un proceso mucho más complejo y que requiere de un soporte informático, ya que resulta imposible realizar a mano los cálculos. Esto es importante porque implica una modificación de las prácticas usuales de puntuación de los tests.

Una posibilidad es que el evaluador haya realizado una administración masiva de un conjunto de ítems sobre los que desconoce sus parámetros (como los exámenes para el ingreso a la Residencia en Psicología de nuestro país). En ese caso, se debe hacer una estimación tanto de los parámetros de cada ítem como del valor del rasgo θ manifestado por los examinados. Otra posibilidad es que, por estudios previos, ya se conozcan las propiedades de los ítems y que sólo se necesite la prueba para medir el nivel de un rasgo determinado de un individuo. Este es el escenario típico de un profesional que administra

en su consultorio un test validado. En este caso, ingresando el patrón de respuesta a los ítems del examinado el software estima solamente el valor θ. El proceso de estimación utilizado no es otro que el de máxima verosimilitud, explicado más arriba.

La TRI también ha permitido diseñar índices para la detección de respuestas aberrantes o deshonestas. Dado un valor de θ de un individuo, existe un patrón de respuestas esperable. El grado de discrepancia entre el patrón observado y el esperado es un índice para identificar respuestas inapropiadas (Zickar & Drasgow, 1996). Se evalúa el ajuste del sujeto al modelo. Ahora bien, un patrón inusual no implica necesariamente una tendencia deliberada a manipular las respuestas sino que puede ser la consecuencia de una comprensión errónea de los ítems, develar diferencias culturales o simplemente mostrar deseabilidad social (Li & Olejnik, 1997).

Segunda Parte. Principales aplicaciones de la TRI

Gracias a estos desarrollos de la TRI y a la generación del software necesario para poder aplicar sus modelos psicométricos se han podido evaluar con mayor profundidad las propiedades de numerosos tests elaborados a partir de la TCT y construir nuevos instrumentos. Asimismo, se renovó el interés en áreas de la medición psicológica que se hallaban estancadas como son el estudio del funcionamiento diferencial de los ítems y la construcción de Banco de Ítems y Tests Adaptativos Informatizados.

Análisis del Funcionamiento Diferencial de los Ítems (DIF)

Si la probabilidad de seleccionar la opción-clave a un ítem para un nivel dado de rasgo depende de alguna otra característica que el rasgo en cuestión, dicha probabilidad podrá variar entre las poblaciones que difieran en tal característica, con lo que el ítem resultaría *sesgado* al tener un funcionamiento diferencial. El funcionamiento diferencial se presenta cuando no se satisface el supuesto de unidimensionalidad. Así, uno de los problemas centrales de la TRI es el estudio del *Funcionamiento Diferencial del Ítem (Differential Item functioning, DIF*). Se considera que un ítem presenta funcionamiento diferencial cuando sujetos de distintos grupos y de un mismo nivel de rasgo tienen diferente probabilidad de dar la respuesta clave. Es decir, cuando el ítem presenta una CCI diferente para cada uno de los grupos. Existen diferentes tipos de DIF (uniforme y no uniforme) y múltiples métodos para su detección basados tanto en la TCT (e.g. Mantel & Haenszel, 1959) como en la TRI (e.g. Camilli & Shepard, 1994, Hambleton & Swaminathan, 1985). La aplicación de estos análisis garantiza que los ítems introducidos en un banco o en un test no funcionen diferencialmente para distintos grupos de personas, perjudicando a uno de los grupos cuando en realidad ambos tienen el mismo nivel de rasgo.

Camilli y Shepard (1994) distinguieron el concepto de DIF del concepto de *sesgo del ítem*. Mientras el primero es puramente estadístico, el segundo considera las causas de tal funcionamiento diferencial. Se hará referencia al sesgo de los ítems sólo cuando se hayan dado explicaciones debidamente fundadas para el funcionamiento diferencial. Así, el análisis del DIF puede ser útil no sólo para la creación de instrumentos de medición invariantes entre poblaciones (por lo que es una herramienta utilizada habitualmente en la adaptación de instrumentos de una cultura a otra) sino también para detectar diferencias entre grupos cuyas interpretaciones podrían generar hipótesis de interés psicológico. Pero antes de dar una interpretación convendría estar seguros de que no se está en presencia de un falso DIF. La vertiente metodológica del estudio del DIF valora la eficacia de los distintos métodos en diferentes condiciones generadas intencionalmente por medio de simulación computacional.

Como ejemplo de una de sus aplicaciones prácticas, el estudio del DIF es utilizado por las compañías más importantes dedicadas a la construcción de pruebas como el último control de calidad al que son sometidos los ítems. A fin de evitar pleitos legales, consideran de particular interés la detección del DIF para grupos raciales, étnicos y de género. Al margen de los problemas éticos y legales, la presencia de DIF es una amenaza de validez para los ítems y el test por lo que su estudio puede ser útil para la comparación de otros grupos que el investigador considere pertinentes. Por ejemplo, se podría estudiar si los indicadores de pruebas gráficas como el Test Guestáltico Visomotor presentan un funcionamiento diferencial según la lateralidad del evaluado (ser diestro o zurdo).

En la misma línea, otra de las aplicaciones más frecuentes de las técnicas de detección del DIF pretende evaluar si las diferencias encontradas entre dos grupos se deben a diferencias genuinas en el rasgo (esto es denominado *impacto*) o son generadas artificialmente por un instrumento que contiene ítems con funcionamiento diferencial. Abad, Colom, Rebollo y Escorial (2004) estudiaron el DIF según el género en los ítems de la Prueba de Matrices Progresivas Avanzada de Raven. Los autores se preguntaban si la naturaleza viso-espacial de los ítems no favorecía a los varones, grupo que suele puntuar más elevado en tests espaciales. Sus resultados mostraron que varios ítems de la prueba presentaban un DIF que perjudicaba a las mujeres. Al descartar los reactivos con DIF los varones continuaban teniendo un mejor desem-

peño, pero la diferencia con el puntaje promedio de las mujeres había disminuido. Este estudio permitió eliminar las diferencias artificiales generadas por el instrumento e identificar la verdadera magnitud del impacto según el género de los individuos.

Banco de ítem y Tests Adaptativos Informatizados (TAIs)

Un banco de ítems es un conjunto de reactivos que miden un mismo rasgo y cuyos parámetros están calibrados; esto es, estimados en una misma escala (Barbero, 1996). Los ítems junto con sus características tanto de contenido como psicométricas son almacenados en una base de datos. De esta manera pueden formar parte de un sistema informatizado de evaluación.

La invarianza de los parámetros de los ítems respecto de las poblaciones y de las mediciones respecto de los instrumentos cobra sentido cuando se dispone de un banco. Como las puntuaciones obtenidas por los individuos a partir de cualquier subconjunto de ítems del Banco dan una medida del rasgo en la misma escala, para comparar los resultados no es necesario que todas las personas realicen el mismo test, sino que se puede elegir el conjunto de ítems que sea más adecuado a su nivel de habilidad o a los objetivos de la medición, garantizando la validez de los resultados obtenidos. De esta forma, se pueden elegir distintos conjuntos de ítems para construir Tests Paralelos tan útiles en el ámbito educativo cuando se requieren frecuentes evaluaciones o múltiples formas de un test. También permite confeccionar test con características psicométricas prefijadas, como por ejemplo, seleccionar los ítems con cierto grado de discriminación, o nivel de dificultad o con mayor función de información.

Asimismo, los bancos de ítems son muy utilizados en el desarrollo de Tests Referidos al Criterio. Su objetivo es determinar si los evaluados dominan ciertos contenidos de conocimiento, para lo cual se suele fijar un punto de corte que permita diferenciar entre expertos y no expertos en la materia en cuestión. Por tanto se pueden elegir del banco los ítems que presenten una discriminación máxima en el nivel del rasgo asociado al punto de corte (Martínez Arias, 1995). En el área de la salud el punto de corte puede estar referido por ejemplo, a la sintomatología que presenta una persona o el nivel en que posee una determinada característica de personalidad.

Una de las aplicaciones de la TRI que ha tenido mayor repercusión es la construcción y administración de Tests Adaptativos Informatizados (Wainer, 2000). Esta aplicación también requiere de un Banco de ítems, pero en este caso un software selecciona progresivamente los ítems más apropiados para la medición de una persona en función del nivel de rasgo que va manifestando en cada respuesta; por lo que resulta una medida más eficiente (Olea & Ponsoda, 2003). En el caso de test de habilidades, si el evaluado responde correctamente, el programa presentará un ítem más difícil. Si la respuesta es incorrecta, presentará un ítem más fácil. La administración de los ítems continúa hasta que se alcanza un número de ítems previamente especificado o un valor determinado de precisión o error típico. Como la dificultad de cada ítem seleccionado se halla en torno a la del anterior, un individuo al que se le administra un TAI nunca tendrá que responder ítems demasiado difíciles o demasiado fáciles para su nivel. Esto lo diferencia de un test convencional de longitud fija en que se presentan en la misma secuencia todos los ítems que lo integran a todos los individuos. De esta forma se evita la tendencia de las personas a contestar al azar y desmotivarse cuando los ítems superan sus conocimientos, así como, el aburrimiento si los ítems son muy fáciles. En el caso de test de personalidad, se presentarán ítems que impliquen un mayor o menor nivel de rasgo en función de que la persona seleccione o no la opción-clave. Aquí también se evita responder a ítems irrelevantes para determinar el nivel de rasgo del evaluado.

Un TAI aporta mayor precisión de la medida en todos los niveles del rasgo, a diferencia de un test convencional que posee su máxima precisión en los niveles medios del rasgo. Esto es posible porque los ítems que conforman el TAI serán aquéllos que maximicen la FI del test para el nivel de rasgo correspondiente al evaluado. A esto se suma un ahorro de tiempo debido a que para proporcionar la misma información sobre el nivel de rasgo se requiere sólo entre un 10 y un 50% de los ítems que se necesitarían si se usara un test no adaptado.

Otra ventaja que presenta está relacionada a la seguridad de la prueba. Como los individuos reciben distintos ítems, no sabrán de antemano cuáles les tocarán. Esto es un asunto de suma importancia cuando es necesario aplicar los tests de forma continua a muestras numerosas de personas (Olea & Ponsoda, 2003).

Actualmente son muchos los test convencionales para los cuales existen versiones adaptativas, por ejemplo, el Graduate Record Exam (GRE), varios tests de aptitudes intelectuales (como el Differential Aptitude Test), y múltiples tests desarrollados tanto en Estados Unidos como en Europa para selección de personal (es el caso del CAT-ASVAB), admisión a centros educativos (e.g. Law School Admission Test), evaluación y certificación educativa (e.g. COMPASS placement tests). En el contexto de los tests de personalidad, Forbey y Ben-Porath (2007) diseñaron una versión adaptativa del MMPI-2. Estos

autores confirmaron el ahorro de ítems y de tiempo de administración con respecto a la versión informatizada pero convencional del MMPI-2, obteniendo resultados comparables en términos de puntuaciones y validez.

CONSIDERACIONES FINALES

La Psicometría mundial se encuentra atravesando un período de transición. Mientras que la TRI se encuentra en auge en Europa y EE.UU., Latinoamérica ha ignorado por mucho tiempo estos nuevos desarrollos. En Argentina, el estudio y aplicación de los modelos de la TRI ha comenzado a dar sus primeros pasos despacio pero de forma auspiciosa. En 1998, Cortada aplicó un modelo logístico de la TRI para la construcción del Test Verbal Buenos Aires. Asimismo, recientemente la TRI ha sido utilizada para el análisis de los datos obtenidos en los estudios de evaluación educativa en nuestro país, tanto en el Operativo Nacional de Evaluación (ONE) dependiente del Ministerio de Educación, Ciencia y Tecnología como en el internacional PISA (Programme for Indicators of Student Achievement) de la Organización para la Cooperación y el Desarrollo Económico (OCDE).

La facilidad conceptual de la TCT hacía compatible su presentación con la enseñanza de sus aplicaciones en las técnicas psicométricas. La formación de los profesionales podía estar centrada en los criterios de calidad que debían reunir los instrumentos para incorporarlos en un proceso de Evaluación Psicológica y en las garantías del proceso mismo. Sin embargo, el crecimiento de la TRI sumó un bagaje teórico sofisticado sobre la fundamentación de la medición psicológica que obligó a reorganizar los contenidos que se dictan en esta área. Se debe vencer una gran resistencia por parte de los alumnos de grado y de posgrado, para quienes la teoría de la medición todavía parece resultar una temática ajena a su propia disciplina. Una formación al margen de la TRI pone a los especialistas en Evaluación Psicológica argentinos en una situación de clara desventaja. Quizás resulte un momento favorable para acortar la brecha existente entre la formación de nuestros profesionales y los de otros países.

BIBLIOGRAFÍA

Abad, F. J., Colom, R., Rebollo, I. Escorial, S. (2004). Sex differential item functioning in the Raven's Advanced Progressive Matrices: evidence for bias. Personality and Individual Differences, 36, 1459–1470.

Barbero, M. I. (1996). Banco de ítems. En J. Muñiz (Ed.). *Psicometría* (pp. 139-170). Madrid: Universitas.

Birnbaum, A. (1968). Some latent trait models and their use in inferring an examinee's ability. En F. Lord & M. Novick (Eds.). *Statistical Theories of Mental Test Scores*. Reading, MA: Addison Wesley.

Camilli, G. & Shepard, L. (1994). *Methods for identifying biased test items*. Thousand Oaks: Sage.

Childs, R., Dahlstrom, W., Kemp, S. & Panter, A. (2000). Item response theory in personality assessment: A demonstration using the MMPI-2 Depression Scale. *Assessment, 7,* 37-54.

Cortada, N. (1998). La Teoría de Respuesta al Ítem y su aplicación al "Test Verbal Buenos Aires". *Interdisciplinaria, 15*, 101-129.

Forbey, J. & Ben-Porath, Y. (2007). Computerized adaptive personality testing: A review and illustration with the MMPI-2 Computerized Adaptive Version. *Psychological Assessment*, 19, 14-24.

Hambleton, R. K. & Swaminathan, H. (1985). *Item response theory: Principles and applications*. Boston: Kluwer.

Li, M. F. & Olejnik, S. (1997). The power of Rasch person-fit statistics in detecting unusual response patterns. *Applied Psychological Measurement, 21*, 215-231.

Lord, F. M. & Novick, M. R. (1968). *Statistical theories of mental tests scores*. Reading, MA: Addison-Wesley.

Lord, F. M. (1980). *Applications of item response theory to practical testing problems.* Hillsdale, N. J.: Lawrence Erlbaum.

Mantel, N. & Haenszel, W. (1959). Statistical aspects of the analysis of data from retrospective studies of disease. *Journal of the National Cancer Institute, 22,* 719-748.

Martínez Arias, M. R. (1995). *Psicometría: Teoría de los Tests Psicológicos y Educativos.* Madrid: Síntesis.

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

Novick, M. R. (1966). The axioms and principal results of classical test theory. *Journal of Mathematical Psychology, 3*, 1-18.

Olea, J. & Ponsoda, V. (2003). *Tests adaptativos informatizados*. Madrid: UNED.

Ostini, R. & Nering, M. (2005). *Polytomous item response theory models.* Newbury Park, CA: Sage.

Rasch, G. (1960). *Probabilistic Models for Some Intelligence and Attainment Tests*. Copenhague: The Danish Institute for Educational Research.

Richaud, M. C. (2005). Desarrollos del análisis factorial para el estudio de ítems dicotómicos y ordinales. *Interdisciplinaria, 22,* 237 – 251.

Rouse, S. V., Finger, M. S. & Butcher, J. N. (1999). Advances in clinical personality measurement: An item response theory analysis of the MMPI-2 PSY-5 scales. *Journal of Personality Assessment, 72*, 282-307.

Samejima, F. (1969). Estimation of latent ability using a response pattern of graded scores. *Psychometrika Monograph, 17*.

Spearman, C. E. (1904). General Intelligence objetively determined and meassured. *American Journal of Psychology, 15*, 201 -293.

Thurstone, L. L. (1928). Attitudes can be measured? *American Journal of Sociology, 33*, 529-554.

Wainer, H. (2000). Computer Adaptive Tests: Whither and whence. *Psicologica, 21,* 121-133.

Zickar, M. J. & Drasgow, F. (1996). Detecting Faking on a Personality Instrument using appropriateness measurement. *Applied Psychological Measurement, 20*, 71-87.

Abstract: This study sets out to introduce the reader to the basic notions of the Item Response Theory (IRT). The First Part outlines the limitations of the classical perspective and defines IRT fundamental concepts: Item Characteristic Curve, Parameters, Information Function, Estimation and Models Fit. The Second Part examines the developments applied on the basis of the latest advances in this theory. Topics such as the Differential Item functioning as well as the construction of Item Banks and Computerized Adaptive Tests are also enlarged upon. It has been concluded and acknowledged that there is a need for a more solid background in the IRT among the specialists in Psychological Assessment in Argentina.