EVALUAR

Laboratorio de Evaluación Psicológica y Educativa

Evaluar, 8 (2008), 1 - 19

ARTICULO METODOLÓGICO

Nuevas tendencias en psicometría

María Cristina Richaud de Minzi[1]*

* Centro Interdisciplinario de Investigaciones en Psicología Matemática y Experimental (CIIPME), dependiente del Consejo Nacional de Investigaciones Científicas y Técnicas (CONICET)

Resumen En este artículo teórico-metodológico se describen sintéticamente los grandes desarrollos de la psicometría en el último siglo, que han surgido con el objetivo de alcanzar una medición cada vez más precisa de los constructos psicológicos. De este modo, se revisan los fundamentos de la teoría clásica de los tests y de la teoría de la generalizabilidad, los cambios acaecidos en el concepto de validez de constructo, los tests referidos a criterio y la teoría de respuesta al ítem. Se concluye que la ciencia psicométrica ha tenido una evolución importante en sus construcciones teóricas, procedimientos y métodos, permitiendo un avance significativo en la problemática de la medición psicológica.

Palabras claves: Desarrollos en psicometría, teoría de la generalizabilidad, teoría de respuesta al ítem.

Abstract. In this theoretical-methodological article is briefly described the major psychometrics developments in the last century, which have emerged with the aim to obtain an ever more precise measurement of psychological constructs. Thus, we review the fundaments of classical test theory and the generalizability theory; the changes in the concept of construct validity, tests criterion relating and the item response theory. Was concluded that psychometric science had an important development in their theoretical constructions, procedures and methods, allowing a significant progress in psychological measurement issues.

Keywords: Psychometrics developments, the generalizability theory, item response theory.

1. Introducción

Generalmente los modelos formales sobre los que se basan la construcción de los tests son de naturaleza matemático-estadística. El primer modelo de puntuación observada, a través del cual se intentó tratar el problema de la incertidumbre o error inherente a cualquiera

[1] Por favor dirigir la correspondencia relacionada con este artículo a:
María Cristina Richaud de Minzi
minzi@ciudad.com.ar

de las medidas realizadas mediante la aplicación de un test, fue el presentado por Spearman en 1904, donde planteó el clásico Modelo Lineal de Puntuaciones. La teoría basada en el modelo de Spearman, fue denominada también Teoría de las Puntuaciones Verdaderas o Teoría Clásica de los Tests.

Las hipótesis básicas de la teoría clásica de las puntuaciones verdaderas son las de un modelo aditivo lineal, donde la variable endógena o dependiente es la que corresponde a la puntuación observada en las pruebas, es la variable explicada que viene determinada por el fenómeno que el modelo traduce; y la variable exógena o independiente es la correspondiente a las puntuaciones verdaderas de los sujetos.

Hipótesis

1. *Hipótesis Fundamental:* Cualquier puntuación observada X es función de dos componentes, que son la puntuación verdadera V del sujeto y el error e.

$$X = V + e$$

En esta relación, X juega el papel de una variable aleatoria sobre una población Π de individuos que toma valores $V = v_g$ sobre g personas de una población de tamaño N. La variable V es otra variable aleatoria asociada que toma valores $V = v_g$ sobre Π.

2. *Hipótesis de nulidad de los errores:* Los errores, en promedio, se anulan. Por lo tanto, en el modelo hay que suponer que la media aritmética de los errores es cero. Esto indica que su esperanza matemática es cero:

$$E(e) = 0$$

Además de ésta, se pueden hacer considerar otras hipótesis acerca de los errores pues para hacer inferencias se supone además que los errores se distribuyen según una ley normal y que el modelo es homocedástico, es decir, que para cualquier i , j las variancias de los errores son iguales.

$$\text{Var}(e_i) = \sigma^2_{ei} = \sigma^2_{ej} = \text{Var}(e_j) \text{ para todo i, j}$$

3. No existe correlación entre las puntuaciones verdaderas y el error en una misma prueba.

$$r_{vg\ eg} = 0$$

4. No existe correlación entre los errores:

$$r_{ei\ ej} = 0 \text{ para todo i,j}$$

5. No existe correlación entre las puntuaciones verdaderas y los errores en formas distintas de un mismo test o en tests diferentes:

$$r_{vj\ ej} = 0 \text{ para todo j}$$

De las hipótesis del modelo lineal de Spearman se deducen las siguientes relaciones:

a) El valor esperado de la puntuación verdadera es igual al valor esperado de la puntuación observada:

$$E\,(\mathrm{V}) = E(\mathrm{X})$$

Es decir, bajo los supuestos del modelo, las medias de las puntuaciones observadas y las verdaderas coinciden.

b) La ecuación de regresión de la puntuación observada sobre la puntuación verdadera es la ecuación de una línea recta que pasa por el origen y que tiene pendiente unidad.

c) La variancia de las puntuaciones observadas es igual a la suma de la variancia de las puntuaciones verdaderas más la variancia de los errores

$$\sigma^2_X = \sigma^2_V + \sigma^2_e$$

Esta es una consecuencia inmediata de la hipótesis de no correlación entre la puntuación verdadera y el error.

d) El cuadrado del coeficiente de correlación lineal entre las puntuaciones observadas y sus correspondientes puntuaciones verdaderas, es igual a la razón de la variancia de las puntuaciones verdaderas con respecto a la variancia de las observadas

$$r^2_{XV} = \frac{\sigma^2_V}{\sigma^2_X}$$

e) De la expresión del coeficiente de determinación, como razón entre variancias, y de $\sigma^2_X = \sigma^2_V + \sigma^2_e$, se deduce otra expresión para r^2_{XV} :

$$r^2_{XV} = \frac{\sigma^2_X - \sigma^2_e}{\sigma^2_X} = 1 - \frac{\sigma^2_e}{\sigma^2_X}$$

f) El cuadrado de la correlación entre la puntuación observada y el error es igual a la razón de la variancia de los errores con respecto a la variancia de las puntuaciones observadas:

$$r^2_{Xe} = \frac{\sigma^2_e}{\sigma^2_X}$$

g) Una nueva relación es la complementariedad a uno del cuadrado de ambos coeficientes de correlación

$$r^2_{XV} + r^2_{Xe} = 1.$$

2. Un cambio fundamental en los conceptos de puntuación verdadera y de confiabilidad

Frente al modelo clásico de Spearman, la teoría de la Generalizabilidad *(Generalizability) (G)* es una teoría estadística acerca de la dependencia de las medidas del comportamiento, que abandona el sentido correlacional dado a la confiabilidad en el contexto clásico.

Esta teoría a la que Cronbach dio el nombre de "generalizability", como una contracción de las palabras *generalize* y *ability*, se desarrolla fundamentalmente en los años 50, aunque su exposición detallada aparece recién en 1972.

La teoría *G* no acepta el concepto tradicional de que un instrumento de medida es adecuado si su coeficiente de confiabilidad es alto, pues las decisiones las basa en los resultados del análisis de las fuentes y tipos de error (Martínez Arias, 1995).

El concepto clásico de la confiabilidad de las medidas está basado en el concepto de valor o puntuación verdadera y en la unicidad de esa puntuación, de manera que considera que cualquier test u observación posee un único valor verdadero en una familia de tests paralelos. El coeficiente de confiabilidad se interpreta como el valor estimado del cuadrado de la correlación, o como la razón de las variancias de los valores verdaderos con respecto a los valores observados.

La teoría *G* considera que esta es una concepción muy limitada de la confiabilidad en tanto implica una única definición de las variancias verdadera y del error, sin tomar en consideración todas las posibles fuentes de error y la naturaleza de su procedencia.

Cronbach, Gleser, Nanda y Rajaratnam (1972) establecieron la definición de *confiabilidad (dependability)* de la siguiente manera:

“El valor en un test u otra medida en la que se basa la decisión es sólo uno de los muchos valores que podrían servir al mismo propósito. Quien toma la decisión no está casi nunca interesado en la respuesta dada a objetos o preguntas estímulo particulares, a un

examinador particular, a un momento particular de prueba. Al menos algunas de estas condiciones de medición pueden ser alteradas sin que el valor sea menos aceptable para el que decide....El dato ideal en el cual basar la decisión debe ser algo como el valor medio de la persona a través de todas las observaciones aceptables" (p. 15).

La confiabilidad, entonces, se refiere al grado de exactitud al generalizar a partir de un valor observado de una persona en un test u otra medida (por ejemplo, observación de conducta, encuesta de opinión) al valor promedio que la persona podría haber recibido bajo todas las posibles condiciones que el examinador estaría deseoso de aceptar. La suposición de que el conocimiento, la actitud, la habilidad, u otro atributo medido es estable está implícita en esta noción de confiabilidad; es decir, suponemos que cualquier diferencia entre los valores obtenidos por un individuo en diferentes ocasiones se debe a una o más fuentes de error, y no a cambios sistemáticos en el individuo debidos a la maduración o el aprendizaje.

Un valor único, obtenido en una ocasión en una forma particular del test con un único examinador, no es completamente confiable; es decir, es improbable que se asemeje al valor promedio de esta persona a través de todas las ocasiones, formas de test y administradores. El valor obtenido por una persona usualmente sería diferente en otras ocasiones, formas del test o administradores. La teoría clásica de los tests puede estimar separadamente sólo una fuente de error por vez (por ejemplo, la variación en los valores a través de las ocasiones puede ser evaluada con el método test-retest de confiabilidad).

La fuerza de la teoría *G* consiste en que múltiples fuentes de error en una medida pueden estimarse separadamente en un solo análisis. Consecuentemente, de una manera similar a la forma en que la *fórmula de profecía* de Spearman-Brown es usada para predecir la confiabilidad como una función de la longitud del test, la teoría *G* posibilita al que decide determinar cuántas ocasiones, formas de test y administradores se necesitan para obtener valores confiables. En el proceso la teoría *G* provee un coeficiente resumen que refleja el nivel de confiabilidad, un coeficiente de generalizabilidad que es análogo al coeficiente de confiabilidad de la teoría clásica.

El concepto de confiabilidad en la teoría *G* es aplicable tanto a universos simples como complejos a los cuales quien toma las decisiones puede querer generalizar. El caso más simple es aquel en el cual el universo es definido por una fuente de variación principal llamada *faceta*.

Desde la perspectiva de la teoría *G*, una medida es una muestra extraída de un

universo de observaciones *admisibles* que quien decide desea tratar como intercambiables con el fin de tomar una decisión. (La decisión debe ser práctica, tal como la selección de los estudiantes con más altos valores para un programa acelerado, o puede ser la base de una conclusión científica tal como el impacto de un programa educativo sobre el logro en ciencia).

Un universo de una faceta está definido por una fuente de error de medición, es decir, por una sola faceta.

Si el investigador intenta generalizar a partir del conjunto de los ítem de un test a un conjunto mucho más grande de ítem, ITEM es una faceta de medición y el universo de ítem deberá definirse por todos los ítem posibles. Si el investigador está interesado en generalizar de una forma de test a un conjunto mucho más abarcativo de formas de test, la faceta es FORMA; el universo deberá definirse a través de todas las formas posibles de test (por ejemplo, todos los tests de personalidad desarrollados durante los últimos 15 años).

La faceta ítem puede entonces estar representada por una gran variedad de ítem que comprenden el *universo* de ítem del test.

Idealmente, los que utilizan los tests desean conocer el puntaje universal de cada persona. Dado que este valor ideal es desconocido, deseamos saber cuán exacta es la generalización realizada a partir del conjunto particular de ítem del test a todos los ítems admisibles, como un indicador del rendimiento del estudiante en un dominio específico.

Cuando se habla de un universo de dos facetas el mismo debe ser definido por dos facetas, por ejemplo *ítem* y *ocasiones*, tomadas juntas; es decir que, el universo de observaciones admisibles estaría definido por todos los ítem aceptables que pueden darse en muchos puntos de tiempo.

Tabla 1.

Fuente de variabilidad, Tipo de variación y su Notación correspondiente

Fuente de variabilidad	Tipo de variación	Notación
Personas (p)	Variancia del Valor universo (Objeto de la medición)	σ^2_p
Observadores (r)	Efecto constante para todas las personas debido a la rigurosidad del observador	σ^2_r
Ocasiones (o)	Efecto constante para todas las personas debido a sus inconsistencias en la conducta de una ocasión a otra	σ^2_o

Continúa en la pagina siguiente

Continuación de la Tabla 1

p x r	Inconsistencias de la evaluación de los observadores de la conducta de una persona en particular	σ^2_{pr}
p x o	Inconsistencias de una ocasión a otra en la conducta de una persona en particular.	σ^2_{po}
r x o	Efecto constante para todas las personas debido a diferencias en la rigurosidad de los observadores de una ocasión a otra	σ^2_{ro}
p x r x o, e	Residual que consiste en la única combinación de p, r, o; facetas no medidas que afectan la medición; y/o efectos aleatorios	$\sigma^2_{pro,e}$

Aunque la teoría *G* se focaliza en las fuentes de variación que contribuyen al error en las medidas, también proporciona un coeficiente de confiabilidad llamado *coeficiente de generalizabilidad (G).* El coeficiente *G* muestra cuán exacta es la generalización desde un *valor observado*, basado en una muestra de conducta de las personas, a su valor universo. Como el coeficiente de confiabilidad de la teoría clásica de los test, el coeficiente de generalizabilidad refleja la proporción de variabilidad en los valores de los individuos que es sistemática, es decir atribuible a la variabilidad del valor universo (valor verdadero). Debido a que la variancia de error es diferente según se trate de decisiones relativas o absolutas, la magnitud de los coeficientes *G* dependerá del tipo de decisiones.

3. La importancia del concepto de validez de constructo

En lo que respecta al concepto de validez éste comienza siendo de índole empírica (validez de criterio o predictiva, por ejemplo).

En términos científicos, la validez de las construcciones hipotéticas constituye uno de los adelantos más notables en la teoría y práctica de la medición. Representa un adelanto notable porque en ella se integran nociones psicométricas y teóricas.

Cuando el experto en medición indaga la validez de las construcciones hipotéticas de una prueba, desea saber qué propiedad o cuáles propiedades psicológicas y de otra índole pueden *explicar* la variancia de dicha prueba. Desea conocer el *significado* de la prueba (Cronbach & Meehl, 1955).

La validación de las construcciones hipotéticas y la investigación científica de carácter empírico están íntimamente ligadas. No se trata simplemente de validar una prueba. Es preciso intentar validar la teoría sobre la que ésta descansa. Según Cronbach dicha

validación consta de tres partes: sugerir cuáles son las construcciones que posiblemente fundamentan la eficacia de la prueba, deducir hipótesis a partir de la teoría con base en la cual se hizo la construcción, someter a prueba empírica las hipótesis (Cronbach, 1970).

El punto importante acerca de la validez de construcciones hipotéticas, que la distingue de los demás tipos de validez, es su interés por la teoría, las construcciones teóricas y las investigaciones científicas de carácter empírico para la comprobación de las posibles relaciones. La validación de construcciones hipotéticas en la medición contrasta agudamente con los enfoques empíricos que definen la validez de una medida únicamente por su eficacia para predecir un criterio.

La comprobación de hipótesis alternativas reviste particular importancia en la validación de construcciones hipotéticas, pues son necesarias la convergencia y la discriminación. Convergencia significa que todos los datos recabados de distintas fuentes y con métodos diferentes revelan que las construcciones tienen un significado igual o similar. Los diversos métodos de medición han de converger en la construcción. Los datos que se obtienen al aplicar el instrumento de medición a grupos distintos en lugares también diversos, deben producir resultados parecidos, o, en caso contrario, explicar las diferencias.

Discriminación significa que empíricamente puede distinguir la construcción hipotética de otras similares y que se puede identificar lo que no guarda relación con ella. Dicho de otro modo, indicamos cuáles son las otras variables que se correlacionan con esa construcción y la manera en que lo hacen pero también indicamos cuáles no deberían estar correlacionadas con ella. Por ejemplo, se determina que una escala para medir el conservadorismo debe correlacionarse, y en efecto se correlaciona, con las medidas de autoritarismo y rigidez, no así con las medidas de aceptabilidad social.

La validación de constructo es un proceso continuo mediante el que se realizan múltiples estudios para poner a prueba distintas hipótesis acerca de la estructura interna de constructo y de sus relaciones con otras variables.

En comparación con muchas otras técnicas multivariadas, que tienen una aproximación exploratoria en el análisis de los datos, el modelo de las ecuaciones estructurales tiene una aproximación confirmatoria (Byrne, 1994), es decir que el patrón de relaciones entre las variables se especifica a priori, con base en expectativas teóricas. Las características distintivas del modelo de las ecuaciones estructurales lo hacen especialmente útil cuando se desea poner a prueba modelos teóricos, a través de los datos empíricos

(Crowley & Fan, 1997).

Dentro de los modelos de las ecuaciones estructurales está el Análisis Factorial Confirmatorio (AFC) que examina las relaciones causales entre las variables observadas y los constructos latentes (factores).

El AFC comienza con un modelo teórico plausible que se supone describe, explica o da cuenta de los datos empíricos. La construcción del modelo se basa ya sea en una información anterior acerca de la naturaleza de la estructura de los datos o en teorías sustantivas en el campo. En tal forma, las variables se limitan a pesar solamente en uno o unos pocos factores.

Aunque el AFE es útil para generar hipótesis, es necesario poner a prueba esa hipótesis con una técnica más rigurosa como en AFC. Finalmente, el AFE difiere del confirmatorio en términos de los efectos potenciales de las fluctuaciones de muestreo en los resultados de investigación. Dado que se trabaja con los datos naturales, el AFE tiende a estar más influido por la idiosincrasia de una muestra particular. En un grado considerable, el AFC evita este problema ajustando un modelo teórico preespecificado a los datos muestrales.

Debido a que es una técnica guiada por la teoría, la construcción del modelo no está afectada por los datos de una muestra particular, y se reduce en gran proporción la probabilidad de la influencia de las fluctuaciones e idiosincrasias de la muestra.

4. Últimas concepciones acerca de la validez

Angoff (1988) y Cronbach y Quirk (1976) afirman que la validez de constructo no puede expresarse en un simple coeficiente. No existe un índice matemático de la validez de constructo ya que su naturaleza es en realidad cualitativa. Cuando un atributo es expresado en términos de los múltiples ítems de un instrumento, el análisis factorial es utilizado para establecer la validez de constructo. Para Messick (1995), la forma convencional de entender la validez (contenido, criterio, constructo) es fragmentada e incompleta, especialmente debido a que no tiene en cuenta las implicaciones del significado del puntaje como base para la acción y las consecuencias sociales del uso del puntaje. La validez no es una propiedad del test o evaluación, sino del significado de las puntuaciones del test, es decir que lo que se valida son las inferencias derivadas de las puntuaciones del test o de otros indicadores, sobre el significado de las puntuaciones o la interpretación para propósitos aplicados y sobre las implicaciones para la acción, es decir, las consecuencias sociales y éticas (Messick, 1989).

Aunque Messick (1980, 1989) aboga por un concepto unitario de validez, concepción que ha sido adoptada por los últimos estándares publicados (AERA, APA, NCME, 1999), también señaló que diferentes tipos de inferencias a partir de los tests requieren distintos tipos de evidencia. En general, Messick (1989, 1995) señala como aspectos a considerar en la validez:

- Contenido: relevancia y representatividad del test
- Sustantivo: razones teóricas de la consistencia observada de las respuestas
- Estructural: configuración interna del test y dimensionalidad
- Generalización: grado en que las inferencias hechas a partir del test se pueden generalizar a otras poblaciones, situaciones o tareas. Este aspecto tiene especial importancia en la adaptación y/o traducción de escalas y tests de una cultura a otra.
- Externo: relaciones del test con otros tests y constructos. Análisis de la utilidad de la medida.
- Consecuencial: consecuencias éticas y sociales del test. Evaluación del sesgo del test.

La posición de Messick en cuanto a la validez contempla entonces un aspecto más referido a los valores sociales y las consecuencias éticas. El debate actual se centra en la cuestión de si la investigación de las posibles consecuencias de la administración y uso de los tests debe incluirse como una parte más del plan de validación de un instrumento, es decir, si la validación es un proceso científico o sociopolítico (Crocker, 1997). Las posturas están encontradas y, aunque todos asumen o destacan la importancia de las consecuencias sociales del uso de los tests, disienten en si deben ser valoradas como parte de la validez del test y uso del mismo o, por el contrario, deben ser valoradas por aquellos que desarrollan los tests pero no incluidas en la validez del mismo.

Entre los defensores de la primera postura cabría citar a Linn (1997), Shepard (1997) y al propio Messick, y de la segunda, a Popham (1997) y Mehrens (1997). Para este último autor, el uso de un instrumento para llevar a cabo una medición y la precisión de la medida obtenida es un asunto muy distinto de las consecuencias que se obtengan de esa medida.

Las posturas actuales acerca de la validez podrían resumirse como sigue: 1) lo que se valida no es el test sino las puntuaciones del test, y por lo tanto la pregunta que tratamos de responder es ¿es válido el uso o la interpretación de las puntuaciones de este test?, 2) la

validez no se puede resumir en un solo indicador o índice numérico de información, si no se asegura mediante la acumulación de evidencia teórica, estadística, empírica y conceptual del uso de las puntuaciones, 3) una puntuación puede ser válida para un uso y no para otro, 4) la validación es un proceso continuo y dinámico y 5) la teoría juega un papel muy importante como guía tanto del desarrollo de un test como de su proceso de validación (Gómez Benito, Hidalgo Montesinos, 2002).

5. La segunda gran teoría de la nueva psicometría: La Teoría de la Respuesta al ítem

Un precursor de esta teoría fue Guttman (1944) cuya preocupación era la propiedad de la unidimensionalidad en una escala. Según Guttman, con una escala unidimensional, el conocimiento del valor escalar total del sujeto evaluado debería permitir reproducir el patrón de valores escalares del mismo. En una escala unidimensional, los ítem pueden ordenarse (por refrendación, descripción o cualquiera sea la dimensión subyacente) de manera tal que la respuesta positiva a un ítem (por ej., de acuerdo, en una escala de actitudes) debería implicar una respuesta positiva a todos los otros ítem más bajos de la escala, e inversamente, la respuesta negativa a un ítem, debería implicar una respuesta negativa a todos los ítem más altos de la escala. Para verificar la unidimensionalidad, Guttman desarrolló la técnica del escalograma.

Se necesitó que pasaran muchos años y que se desarrollaran los modelos matemáticos de las regresiones logísticas y los actuales sistemas de computación para que fuera posible abordar una solución para las limitaciones de los modelos clásicos de medición.

El principal inconveniente que presenta la teoría clásica de la medición es la imposibilidad de separar las características del examinado de las características del test: uno puede ser interpretado sólo en el contexto del otro. La característica del examinado en la que estamos interesados es la *habilidad* medida por el test. ¿Qué significa la habilidad? En la teoría clásica de los tests la *noción de habilidad* es expresada a través del *valor verdadero* que es definido como el valor esperado del rendimiento en el test de interés. La habilidad del examinado es definida sólo en términos de un test particular. Cuando el test es *difícil* parecerá que el examinado tiene poca habilidad; cuando el test es *fácil* parecerá que el examinado tiene mayor habilidad. ¿Qué significa que un test es fácil o difícil? La *dificultad del ítem* se define como la proporción de examinados de un grupo de interés, que lo contesta correctamente. El grado de dificultad del ítem depende de la habilidad de los examinados y la

habilidad de los examinados depende de la dificultad de los ítems del test.

El poder discriminativo de los ítems, la validez y la confiabilidad son también definidos en términos de un grupo particular de examinados. El test y las características de los ítems cambian en función de la muestra de examinados y las características de los examinados cambian cuando se modifica el contexto del test. Por lo tanto, es muy difícil comparar examinados a los que se han administrado diferentes tests y comparar ítems cuyas características se han determinado con diferentes grupos de examinados.

Otros dos problemas relativos a la teoría clásica se refieren a la definición de confiabilidad y a su inversa conceptual: el error estándar de medición. En la teoría clásica de los tests, la *confiabilidad* se define como la correlación entre los resultados del test en formas paralelas del mismo. En la práctica es muy difícil sino imposible satisfacer esta definición. Los diferentes coeficientes de confiabilidad con los que se cuenta proveen estimaciones más bajas de confiabilidad o estimaciones con sesgos desconocidos (Hambleton & van der Linden, 1982). El problema con el error estándar de medición, que es función de la confiabilidad de la prueba y de la variancia, es que se supone igual para todos los examinados. Sin embargo, los valores en cualquier test no son medidas igualmente precisas para los examinados con diferente grado de habilidad. Por lo tanto, la suposición de errores de medición iguales para todos los examinados no es plausible.

Una última limitación de la teoría clásica es que está orientada al test y no al ítem. El modelo clásico del resultado verdadero no considera cómo responde el sujeto a un ítem dado.

Una teoría alternativa de los tests debería incluir: a) características de ítems no dependientes del grupo, b) resultados que describan capacidades de los examinados no dependientes del test, c) un modelo expresado al nivel del ítem y no al nivel del test, d) un modelo que no requiera tests estrictamente paralelos para la evaluación de la confiabilidad, y e) un modelo que provea una medida de precisión para cada habilidad.

El modelo de la teoría de la respuesta al ítem propone que una dimensión subyacente simple genera un conjunto de respuestas observables a ítems dicotómicos. Esta teoría es un modelo logístico multivariado con un predictor no observable que tiene un objetivo principal en las propiedades de los ítems y un objetivo secundario en las estimaciones o puntuaciones de la persona a lo largo de la dimensión del atributo subyacente (θ) (Panter, Swygert, Dahlstrom & Tanaka, 1997).

En relación con el tema central en la construcción de una escala (cuales items

funcionan mejor como indicadores del rasgo), los métodos de la teoría clásica (correlaciones ítem-test, dificultad del ítem y análisis factorial) funcionan bien. Los parámetros de discriminación y dificultad de la TRI no han llevado a decisiones diferentes acerca de la calidad de los ítems. Esto se debe a que todos los índices (correlación item-test, pesaje factorial o la pendiente de la función de respuesta al ítem) están altamente relacionados (Reise & Henson, 2003).

Sin embargo, los parámetros de los ítems de la TRI tienen una propiedad de invariancia lineal que no tienen los índices de la TC y que facilitan importantes aplicaciones como el Funcionamiento diferencial del ítem (DIF). El modelo TRI supone un estructura subyacente unidimensionaI donde todos lo items sirven como indicadores de un factor subyacente. Esta aproximación permite a los investigadores probar hasta qué punto los rasgos del modelo subyacente son invariantes a través de las distintas subpoblaciones como género, edad o estatus clínico.

En el caso del funcionamiento diferencial del ítem, los individuos provenientes de dos grupos tienen diferentes probabilidades de acertar el ítem, aún cuando los dos grupos están en el mismo punto en la dimension del atributo subyacente. Por ejemplo, hombres y mujeres con un nivel similar de depresión estimada deberían tener la misma probabilidad de acordar con un ítem particular. Si la probabilidad varía según el género, el item presenta sesgo (Panter, *et al.,* 1997).

Hay otras ventajas conectadas con la interpretación de los parámetros TRI. Por ejemplo, el parámetro b (dificultad del ítem) es más fácil de interpretar que la proporción de acuerdos de la TC. En la TC el significado cambia a través de la escala, mientras que en la TRI la dificultad del ítem está en la misma escala del nivel del rasgo del examinado (la dificultad del ítem es la cantidad de constructo necesaria para alcanzar una probabilidad de acuerdo de .50).

La diferencia más significativa entre la TRI y la TC se basa en la conceptualización del error de medición. La TC provee un solo índice de confiabilidad y error estándar que es constante para todos los examinados. La TRI por el contrario, permite calcular la Función de información del item y la Función de información de la escala, permitiendo que el error de medición cambie a través del continuo de la variable latente dependiendo de las propiedades de la medida.

Específicamente en la TRI, un ítem es juzgado de acuerdo a su *Función de*

información. La información indica cuán bien un item discrimina entre respondentes que están a diferentes niveles de la variable latente. La Función de información es aditiva a través de los ítems, por lo que las funciones de información pueden sumarse para determinar la función de información de la escala que indica cuán bien funciona un conjunto de ítems como un todo. Además, la información se relaciona inversamente con el error estándar de medición. Las medidas pueden proveer diferentes cantidades de información. Los respondentes tendrán diferentes errores de medición dependiendo de donde están ubicados en la variable latente (Reise & Henson, 2003).

Gray-Little, Williams, y Hancock (1997) hicieron un análisis TRI de la escala de Autoestima de Rosenberg y encontraron que, aunque el test tenía alta consistencia interna, la información TRI mostraba que la medida era bastante pobre para diferenciar entre los examinados con alta puntuación en el rasgo.

Steinberg y Thissen (1995) describieron los análisis separados del modelo TRI de dos parámetros para las dimensiones Acción y Pensamiento de la escala de Acción de Kuhl (1985). Los parámetros de discriminación (a) de los ítems de Pensamiento mostraron que algunos ítems de la escala se relacionaban más que otros con el constructo subyacente y que los parámetros umbral (b) para los ítems de Acción también discriminaban entre los ítems. El examen de las curvas de cada ítem permitió a los investigadores ver qué ítems eran útiles para medir cada dimension y cuáles podían eliminarse sin disminuir la precisión de la medición (Steinberg & Tissen, 1995).

La aplicabilidad del tercer parámetro de la TRI a los datos de personalidad no se ha desarrollado totalmente hasta la fecha. Esto puede deberse a que contrariamente a lo que ocurre en educación, los respondentes en el extremo inferior de la dimensión del rasgo pueden no estar motivados a dar una respuesta “correcta” o, en términos de personalidad, a acordar con un ítem. En algunos ítems, los respondentes con bajo nivel del rasgo pueden interpretarlos simplemente en forma diferente (Panter, *et al.*, 1997).

Junto con el análisis factorial exploratorio que indica la dimensionalidad subyacente a un conjunto de ítems, la TRI es útil para la construcción y validación de inventarios de personalidad.

Otra importante función de la TRI es el diseño de índices de “propiedad”. En la medición de la personalidad se incluyen escalas de validez para identificar protocolos que pueden ser caracterizados como respuestas anómalos a un conjunto de ítems. Las respuestas

son anómalas o inapropiadas si los respondentes con bajos niveles en el rasgo acuerdan con ítems que indican altos niveles de rasgo, o inversamente (Reise & Waller, 1993; Swygert, Panter, Dahlstrom & Reise, 1996). El índice Iz (Drasgow, Levine & Williams, 1985) puede emplearse para identificar puntuaciones inapropiadas en un inventario de personalidad con ítems dicotómicos que satisface los supuestos de la TRI, (Birenbaum, 1985).

En un estudio sobre las escalas de validez del MMPI-2 (Swygert *et al.,* 1996), se demostró que fue difícil discriminar entre un respondente con altos niveles de psicopatología y uno que respondía a los ítems de un modo tradicionalmente conocido como inconsistente. La inconsistencia en una escala fue predictiva de la inconsistencia en otra, y la puntuación en la escala Esquizoide fue el indicador más altamente predictivo del valor Iz de un responderte. Aquí los índices de propiedad sirvieron para examinar la validez de las respuestas a los ítems y para resaltar que las respuestas no válidas en una escala de un inventario de personalidad son acompañadas probablemente por respuestas no válidas en otras escalas. (Panter *et al.,* 1997).

6. Recientes desarrollos acerca de la aplicación conjunta de TRI y TG

El proceso de medición de una variable latente es una mezcla de desarrollo y diseño. La TRI es una herramienta útil para el análisis y escalamiento de las facetas del test. Los modelos TRI pueden ayudar en la caracterización de los ítems, y en las puntuaciones y formas a incluir en un instrumento de medición. La TRI es menos útil en el diseño de preguntas que reflejen la variabilidad relativa de las facetas del diseño del test. Para tales decisiones en el contexto de los instrumentos de medición multifacético, la TG puede tener un importante papel.

Como modelos teóricos parecen incompatibles, al menos en la superficie. Brennan (2001), por ejemplo, afirma que la TG es fundamentalmente un modelo muestral, mientras que la TRI es fundamentalmente un modelo de escalamiento. No obstante, dado que cada enfoque puede proveer información esencial para el desarrollo y diseño de instrumentos de medición, se podría esperar ver a la TRI y la TG utilizadas en conjunto (por ejemplo en Bock, Brennan & Muraki, 2002). De hecho, el uso de la TRI parece más prevalente en la bibliografía de medición que el uso secuencial de la TRI y la TG.

Cuando no se toma en cuenta la variabilidad de las diferentes facetas en los tests multifaceta, el uso de la TRI sola puede llevar a que no se examinen ciertas cuestiones, lo que

es particularmente importante cuando deben tomarse decisiones (Briggs, 2005).

Existe un nuevo enfoque llamado Generalizabilidad en TRI (GIR, por sus siglas en inglés). La fundamentación para este enfoque fue realizada en un trabajo inédito de Michael Kolen y Deborah Harris. El enfoque GIRM incorpora esencialmente la TG en la TRI haciendo supuestos de distribución acerca de las facetas de medición relevantes. El desarrollo de Kolen y Harris hace posible estimar los componentes de la variancia de la TG junto con los parámetros tradicionales de la TRI. La TG y la TRI pueden integrarse en el contexto de un diseño de una faceta con ítems binarios y de un diseño multifacético con ítems politómicos.

7. Tests referidos al criterio

Uno de los avances importantes de la Teoría de los Tests en los últimos 25 años fue el desarrollo y creciente interés por los Tests Referidos al Criterio (TRC). Estos tests representan procedimientos para evaluar el rendimiento y/o conducta de los sujetos con relación a dominios de contenidos bien definidos, en vez de por referencia a la conducta de otros sujetos (como en los Tests Referidos a las Normas, TRN).

La definición de los TRC que hoy es más aceptada es la de Popham (1978): “un test referido al criterio se utiliza para evaluar el status absoluto del sujeto con respecto a algún dominio de conductas bien definido” (p.93).

Hambleton y Rogers (1991) hacen una serie de precisiones a esta definición. En primer lugar, además de dominio de conductas, puede hablarse intercambiablemente de objetivos, destrezas y competencias. En segundo lugar, el dominio debe estar bien definido, siendo variables la amplitud y los contenidos de este dominio, ya que éstos dependen de la finalidad del test. En tercer lugar, cuando un TRC incluye más de un objetivo, los ítems que cubre cada uno de los objetivos suelen organizarse en subtests y el rendimiento de los sujetos es evaluado en cada uno de los objetivos. En cuarto lugar, aunque es una práctica frecuente establecer estándares de rendimiento o puntos de corte, la definición de TRC no incluye explícitamente este requisito, ya que pueden darse interpretaciones meramente descriptivas del rendimiento de los sujetos.

Los requisitos básicos para que un test pueda ser considerado TRC son los siguientes:

1) La existencia de un conjunto de objetivos claramente definidos.

2) Una proposición explícita de la finalidad del test.

8. Conclusiones

A lo largo de este artículo he tratado de presentar lo que a mi juicio son los grandes avances de la psicometría, que han surgido en el afán de acercarse cada vez más a una medición psicológica precisa del atributo en estudio.

Sabemos que el objetivo es ambicioso, ya que el problema está en la base misma de nuestra ciencia, cuyos objetos en su gran mayoría responden a mediciones derivadas o convencionales. En toda medición, aún en la fundamental, se cometen errores, pero en la psicometría la imperfección es mayor debido a su mediatez.

A pesar de todo, si la ciencia avanza desde la descripción precisa de los hechos a la explicación de los mismos, parecería que en psicometría se ha progresado mucho. De una concepción de la confiabilidad, que partió de la caracterización del error como un todo indiferenciado, con base en un estadístico descriptivo como la r de Pearson, se evolucionó a una teoría muestral que revoluciona el concepto de puntuación verdadera y explica el error estudiando sus fuentes o facetas, a través de una técnica inferencial como el análisis de la variancia; de un análisis de la validez, que comenzó con procedimientos empíricos para determinarla con referencia a un criterio o a una predicción, se llega al desarrollo del concepto de validez de constructo, que se basa en la lógica experimental de la puesta a prueba de hipótesis y análisis probatorios como el AFC; del estudio de los atributos subyacentes, a través de pruebas que describen su funcionamiento basadas en la comparación con una norma relativa a los sujetos de una población particular, se pasa a desarrollar instrumentos que incluyen ítems que explican los resultados obtenidos por una persona por la cantidad de aptitud o rasgo que posee, en términos absolutos.

Podríamos concluir entonces que la ciencia psicométrica ha tenido una evolución importante en procedimientos y métodos, pero fundamentalmente en conceptualizaciones teóricas, que nos han permitido avanzar decididamente en el apasionante problema de la medición en psicología.

Referencias

Angoff, W. H. (1988). Validity: An evolving concept. En H. Wainer y H. Braun (Eds.) *Test validity*. Hillsdale, NJ: Lawrence Erlbaum Associates.

Birenbaum, M.(1985). Comparing the effectiveness of several IRT based appropriateness measures in detecting unusual response patterns. *Educational and psychological Measurement, 45*, 523-533.

Bock, R. D., Brennan, R. L., & Muraki, E. (2002). The information in multiple ratings. A*pplied psychological*

measurement, 26 (4), 364-375.

Brennan**,** R. L. (2001). *Generalizability theory*. New York: Springer-Verlag.

Briggs, D. (2005). *Generalizability in Item Response Modeling.* University of Colorado at Boulder. Extraído el 16 de octubre de 2006 de la Web: http://bearcenter.berkeley.edu/seminars/seminarsSpring2005.php#feb1

Byrne, B. (1994). *Structural equation modeling with EQS and EQS/Windows: Basic concepts, applications, and programming*. Newbury Park, CA: Sage.

Crocker, L. (1997). Assessing content representativeness of performance assessment exercises. *Applied Measurement in Education, 10* (1), 83-95.

Cronbach, L. J. (1970). *Essentials of Psychological Testing.* Nueva York: Harper and Row.

Cronbach, L. J. & Quirk, T. J. (1976). Test validity. *In International Encyclopedia of Education*. New York: McGraw-Hill.

Cronbach, L. J. & Meehl, P. E. (1955). Construct validity in psychological tests. *Psychological Bulletin, 52,* 281-302.

Crowley, S. L. & Fan, X. (1997). Structural equation modeling: Basic Concepts and applications in personality assessment research. *Journal of Personality Assessment, 68*, 3, 508-531

Drasgow, F. Levine, M., & Williams, E. A. (1985). Appropriateness measurement with polychotomous item response models and standardized indices. Bri*tish Journal of Mathematical and Statistical Psychology, 38*, 67-86.

Gómez Benito, J., & Hidalgo Montesinos, M.D. (2002). *La validez en los tests, escalas y cuestionarios*. Informe del proyecto número BSO2001-3751-C02-02 financiado por el Ministerio de Ciencia y Tecnología y la FEDER. Consultado el 6/8/2005. Extraído de la Web http://huitoto.udea.edu.co/~ceo/Validez02.htm.

Gray-Little, B., Williams, V. S. L., & Hancock, T. D. (1997). An item response theory analysis of the Rosenberg Self-Esteem Scale. *Personality and Social Psychology Bulletin, 23*, 443-451.

Guttman, L. (1944), A Technique for Scale Analysis, *Educational and Psychological Measurement*, 4, 179-190.

Hambleton, R.K. & Rogers, H.J. (1991). Advances in criterion-referenced testing and measurement : a review. En R.K. Hambleton & J.N.Zaal (Eds.). *Advances in educational and psychological testing : Theory and applications*. Boston-Kluwer.

Hambleton, R.K. & van der Linden, W.J. (1982). Advances in item response theory and applications: An introduction. *Applied Psychological Measurement, 6*, 373-378.

Kuhl, J. (1985). Volitional mediators of cognition-behavioral consistency: Self-regulatory processes and action versus state orientation. In J. Kuhl & J. Beckman (Eds.), *Action control: From cognition to behavior* (pp.101-128). Berlin: Springer-Verlag.

Linn, R.L. (1997). Evaluating the validity of assessments: The consequences of use. *Educational Measurement: Issues and Practice*, 16, 14-16.

Martínez Arias, R. (1995). Psicometría: Teoría de los tests psicológicos y educativos. Madrid: Síntesis Psicológica.

Mehrens, W.A. (1997). The consequences of consequential validity. *Educational Measurement: Issues and*

Practice, *16*, 16-18.

Messick, S. (1980). Test validity and the ethics of assessment. *American Psychologist, 35*, 1012-1027.

Messick, S. (1989). Validity. En R.L. Linn (Ed.), *Educational Measurement* (3th. Ed.). New York: American 200 Council on Education and Macmillan publishing company.

Messick, S. (1995). Standards of validity and the validity of standards in performance assessment. *Educational measurement: Issues and Practice*, 14, 5-8.

Panter, A. T., Swygert, K. A., Dahlstrom, W. G., & Tanaka, J. S. (1997). Factor analytic approaches to personality item-level data. *Journal of Personality Assessment, 68(3),* 561-589.

Popham, W. J. (1978). *Criterion-referenced measurement*. Englewood Cliffs, NJ: Prentice-Hall.

Popham, W.J. (1997). Consequential validity: Right concern-wrong concept. *Educational Measurement: Issues and Practice*, 16, 9-13.

Reise, S. P., & Henson, J. M. (2003). A discussion of modern versus traditional psychometrics as applied to personality assessment scales. *Journal of Personality Assessment, 81(2),* 93-103.

Reise, S. P., & Waller, N. G (1993). Traitedness and the assessment of response pattern scalability. *Journal of Personality and Social Psychology*, *65*, 143-151.

Shepard, L.A. (1997). *The centrality of test use and consequences for test validity*. Educational Measurement: Issues and Practice, 16, 5-8, 13, 24.

Steinberg, L., & Thissen, D. (1995). Item response theory in personality research. In P. E. Shrout and S. Fiske (Eds.), *Personality research, methods, and theory: A festschrift honoring Donald W. Fiske* (pp. 161-181). Hillsdale, NJ: Lawrence Erlbaum Associates, Inc.

Swygert, K. S., Panter, A. T., Dahlstroom, W. G., & Reise, S. (1996). *The use of appropriateness indices in the MMPI-2*. Chapel Hill: University of North Carolina, L.L. Thurstone Psychometric Laboratory.