anales de psicología
2006, vol. 22, nº 1 (junio), 143-154

ISSN edición impresa: 0212-9728. ISSN edición web (www.um.es/analesps): 1695-2294

Técnicas para detectar patrones de respuesta atípicos

Rosa Mª Núñez Núñez [1,*] y José Antonio López Pina[2]

[1]*Universidad Miguel Hernández de Elche (España),* [2]*Universidad de Murcia (España)*

Resumen: La identificación de patrones de respuesta atípicos es de gran utilidad para la construcción de tests y de bancos de ítems con propiedades psicométricas así como para el análisis de validez de los mismos. En este trabajo de revisión se han recogido los más relevantes y novedosos métodos de ajuste de personas que se han elaborado dentro de cada uno de los principales ámbitos de trabajo de la Psicometría: el escalograma de Guttman, la Teoría Clásica de Tests (TCT), la Teoría de la Generalizabilidad (TG), la Teoría de Respuesta al Ítem (TRI), los Modelos de Respuesta al Ítem No Paramétricos (MRINP), los Modelos de Clase Latente de Orden Restringido (MCL-OR) y el Análisis de Estructura de Covarianzas (AEC).
Palabras clave: Métodos de ajuste de persona; medición apropiada; patrón de respuesta atípico; escalograma de Guttman; teoría de la Generalizabilidad; Teoría de Respuesta al Ítem; modelos de respuesta no paramétricos; tests adaptativos informatizados.

Title: Aberrant patterns detection methods.
Abstract: Aberrant patterns detection has a great usefulness in order to make tests and item banks with psychometric characteristics and validity analysis of tests and items. The most relevant and newest person-fit methods have been reviewed. All of them have been made in each one of main areas of Psychometry: Guttman's scalogram, Classical Test Theory (CTT), Generalizability Theory (GT), Item Response Theory (IRT), Non-parametric Response Models (NPRM), Order-Restricted Latent Class Models (OR-LCM) and Covariance Structure Analysis (CSA).
Key words: Person-fit methods; appropriateness measurement; aberrant pattern; Guttman's scalogram; Generalizability Theory; Item Response Theory; non-parametric Response Models; computerized adaptive testing.

Introducción

Medición apropiada es la denominación que Levine y Rubin (1979) otorgaron a la parte de la Psicometría encargada de identificar y tratar patrones de respuesta atípicos (PRA) siguiendo los supuestos de la TRI. Un patrón de respuestas individual se define por una secuencia de 1s y 0s en función de si la respuesta es afirmativa o negativa, acierto o fallo, respectivamente; si se entiende que un PRA es el conjunto de respuestas del sujeto a un test que no representa a la variable psicológica objeto de medida, entonces la *medición apropiada* evalúa el ajuste de un modelo de respuesta al ítem (MRI) a las respuestas del sujeto. Sin embargo, Meijer y Sijtsma (1999, 2001) han preferido el término *métodos de ajuste de persona* ya que con él se estaría haciendo referencia no sólo a aquellas técnicas y estadísticos que evalúan el ajuste de un MRI a las respuestas de un sujeto, sino también a las que valoran el grado de acuerdo entre el patrón de respuestas de un sujeto y los patrones de respuestas de la muestra normativa a la que pertenece.

Empleando esta última terminología, los métodos de ajuste de persona se pueden clasificar en los siguientes tres grupos:

- Estadísticos para detectar PRA tomando como criterio un grupo normativo.
- Estadísticos de medición apropiada pertenecientes a la TRI.
- El índice de ajuste de persona según el AEC.
- Estadísticos de ajuste de persona para MCL-OR.

* **Dirección para correspondencia [Correspondence address]**: Rosa Mª Núñez Núñez. Universidad Miguel Hernández. Facultad de CC. Sociales y Jurídicas de Elche. Departamento de Psicología de la Salud. Área de Metodología de las CC. del Comportamiento. Edificio Altamira. Avenida de la Universidad s/n, 03203, Elche (Alicante, España). E-mail: rnunez@umh.es

Estadísticos para detectar PRA tomando como criterio un grupo normativo

En general, el procedimiento que siguen estas técnicas es escoger un patrón de respuestas de un sujeto en un test de ítems dicotómicos o dicotomizados, y compararlo con el patrón esperado según el grupo normativo al que pertenece.

El escalograma de Guttman y sus derivaciones estadísticas

Las primeras técnicas utilizadas para identificar PRA comparaban el patrón de respuesta observado de un sujeto con el patrón esperado según el modelo determinístico de Guttman (1944, 1950) para medir actitudes, posteriormente extrapolado al ámbito de las variables aleatorias (habilidades, rendimiento, conocimientos…). Este modelo fue elaborado con el propósito de construir escalas unidimensionales en las que los ítems estarían ordenados de modo que el acierto de un ítem implicaría el éxito de los ítems jerárquicamente inferiores.

Para un modelo de respuesta al ítem en el que el parámetro de habilidad del sujeto (θ_i) es conocido y δ_j es el parámetro de dificultad del ítem medido en la misma escala que la habilidad, entonces $\theta_i \geq \delta_j \Leftrightarrow P_j(\theta_i)=1$ y $\theta_i < \delta_j \Leftrightarrow P_j(\theta_i)=0$.

El planteamiento de Guttman (1944, 1950) ha sido el punto de partida de la elaboración de los siguientes estadísticos para identificar PRA:

El coeficiente de correlación biserial-personal. Este coeficiente valora la relación existente entre el grado de dificultad que tiene un sujeto para acertar un ítem y la que teóricamente debería tener por pertenecer a una determinada muestra. Asumiendo

que el rasgo subyacente al sujeto θ_i se distribuye según la ley normal, entonces las respuestas del sujeto a los ítems del test también se ajustarán a una distribución normal. Ordenados los ítems en dificultad creciente ($\delta_j < \delta_k$) –especificada a partir de las respuestas a los ítems de la muestra normativa–, se fija un punto de corte que divide el continuo de los ítems en aquellos que son más fáciles para el sujeto y los acierta, y los que son más difíciles y los falla. Bajo estos supuestos, el coeficiente de correlación biserial-personal (r_{bisper}) de Donlon y Fischer (1968) se obtiene comparando el patrón de respuestas de un sujeto y los índices de dificultad de los ítems en el grupo normativo.

El índice de dificultad de los ítems se normaliza con la escala Δ del Educational Testing Service (ETS; Angoff, 1982). Los valores altos de r_{bisper} indican que el patrón del sujeto no es atípico, es decir, está de acuerdo o es similar a cualquier otro patrón de respuestas de un sujeto de esa muestra con el mismo o aproximado valor de rasgo subyacente; si $r_{bisper} < 0$ informaría de que el patrón de respuestas en cuestión está inversamente relacionado con la dificultad de los ítems en la muestra normativa y se calificaría de patrón atípico.

El índice de precaución. Para calcular el índice de precaución de Sato (1975; Tatsuoka y Linn, 1983) es necesario construir una matriz $N \times n$ de respuestas de N sujetos a n ítems dicotómicos, denominada *tabla S-P*. En las filas de esta matriz se colocan los sujetos ordenados de modo descendente en función de su puntuación total en el test y en las columnas aparecen los ítems dispuestos en dificultad creciente de acuerdo con el escalograma de Guttman. A partir de esta matriz se traza la curva del sujeto o curva-S (*student curve* o *S-curve*) que es una línea que comienza a la derecha de la n-ésima celdilla de la matriz del primer sujeto, siendo ésta la que coincide con el número de aciertos del sujeto (X_i). Al ir conectando las sucesivas celdillas de cada uno de los sujetos se obtiene una curva con la forma de la función de distribución de ogiva. Sato (1975) definió la *curva-S perfecta* como aquella que, manteniendo la curva-S observada invariante, se obtiene al permutar los 0s de la izquierda de la curva por 1s y los 1s de la derecha de la misma por 0s. En la matriz de la nueva tabla están las respuestas modificadas de los sujetos a los ítems del test y en la que las puntuaciones totales iniciales de los sujetos en el test y la puntuación total tras la transformación en la curva-S perfecta no varían; sin embargo, sí cambia el número de respuestas acertadas por ítem debido a dicha reestructuración de la matriz.

El *índice de precaución para el sujeto* i, C_i, se calcula a partir del análisis de las tablas que contemplan la curva-S observada y la curva-S perfecta. Si $C_i = 0$ el patrón está de acuerdo con el patrón Guttman; valores elevados de C_i indican que el patrón es atípico. Pero C_i no tiene límite superior en su rango de valores, lo cual dificulta su interpretación al no poder decidir apropiadamente acerca de si el patrón es atípico o no. Por ello, Harnisch (1983) sugirió que se debería considerar a un patrón como atípico si $C_i > 0'60$.

El índice de precaución modificado. Harnisch y Linn (1981) hicieron una modificación del índice de precaución C_i de Sato (1975), para solventar las dificultades de interpretación que éste presenta cuando su magnitud es elevada. El *índice de precaución modificado* C_i^* para el sujeto i toma valores entre 0 y 1, facilitando así su interpretación. Si $C_i^* = 0$ el sujeto i tiene un patrón de respuestas perfecto; el punto crítico para clasificar a un patrón como atípico es $C_i^* > 0'30$, de modo que cuanto más se aleje de este valor más atípico será, llegando al máximo de atipicidad si $C_i^* = 1$, lo cual indicaría que el sujeto presenta un patrón Guttman inverso.

Los índices basados en el número de errores Guttman. Guttman (1950, p. 70) definió el número de errores de un patrón según el modelo determinístico como el número de respuestas erróneamente pronosticadas para un sujeto a partir de su puntuación. Van der Flier (1977) elaboró un índice sencillo de calcular para detectar PRA a partir de una tabla S-P. Con este índice, denotado U_{1i}^*, se obtendría la desviación del patrón de respuestas del sujeto i respecto del patrón de la curva-S perfecta. Es un índice ponderado por el inverso del número máximo de errores Guttman correspondientes al patrón del sujeto i [$X_{i.}(n - X_{i.})$], lo cual estrecha el rango de U_{1i}^* en (0,1). Esto facilita que se puedan comparar patrones con diferente puntuación total y determinar si es un PRA o no; cuanto más se aleje U_{1i}^* de 0 más atípico será el patrón.

Meijer (1994) modificó el índice U_{1i}^* y sugirió que el uso del número de errores Guttman podría ser un buen índice para detectar PRA, siempre y cuando no se dejara influir ni por el número de ítems del test ni por el número de aciertos del sujeto en el test. Para ello, dividió el número de errores Guttman de un patrón por el número máximo de errores que le correspondería al número de respuestas correctas de ese patrón; a este estadístico lo denotó G_i^*. Si $G_i^* = 0$, el patrón es un vector Guttman y cuanto más se aleje de este valor más atípico será. Meijer (1995) aclaró que, aunque se emplee el número de errores Guttman, adoptó la definición de error dada por Loevinger (1947, 1948), la cual contempla

una versión probabilística del modelo determinístico de Guttman (1944, 1950). Por esta definición, el número de errores se contabiliza por pares de ítems.

El índice de conformidad con la norma. Un aspecto a tener en cuenta en los tests que evalúan los procesos cognitivos empleados en la resolución de problemas es la *consistencia* de las respuestas a lo largo del tiempo. La importancia de la identificación de patrones inconsistentes no sólo radica en detectar sujetos que fallan los ítems por emplear procesos cognitivos erróneos, sino también en localizar a aquellos sujetos que, aunque responden correctamente a los ítems, no han utilizado las operaciones mentales adecuadas. Tatsuoka y Tatsuoka (1982) desarrollaron un índice para dicho propósito al que denominaron *índice de conformidad con la norma* (*Norm Conformity Index*, NCI), el cual valora el grado de aproximación entre el patrón de respuestas de un sujeto y un patrón del grupo normativo ajustado a la escala Guttman con el mismo número de respuestas correctas que aquel. Este índice es un coeficiente de correlación entre la dificultad de los ítems ordenados según el grupo normativo y el patrón de respuestas del sujeto en estudio. Su intervalo de valores es (-1,+1); si $NCI = +1$, el patrón de respuestas es un patrón ajustado a la escala Guttman; si $NCI = -1$ indica que el patrón es un vector de respuestas inverso al que le correspondería en dicha escala. Cuanto más se aleje NCI de +1, más atípico será el patrón.

El coeficiente de escalabilidad. Sijtsma y Meijer (1992) propusieron una extensión del *coeficiente de escalabilidad* H de Loevinger (1948) para identificar PRA a partir de una matriz $N \times n$ ordenada según el escalograma de Guttman. Son dos los estadísticos que derivaron: H^T compara el patrón de un sujeto con el de otro de la misma muestra, y H_i^T compara el patrón de un sujeto con el resto de la muestra. Para calcular ambos es necesario conocer la proporción esperada de ítems acertados coincidentes por un sujeto i y por un sujeto g. Tanto H^T como H_i^T varían entre 0 y 1; si son iguales a 0 el patrón del sujeto es atípico; en el caso opuesto, cuando H^T o H_i^T valen 1, indicaría que es un patrón Guttman perfecto. Sin embargo, Sijtsma (1986) observó que en presencia de un patrón de respuestas perfecto estos coeficientes no siempre son iguales a 1.

El estadístico g_2 de Frary, Tideman y Watts

A partir de los supuestos de la TCT, Frary, Tideman y Watts (1977) elaboraron el índice g_2 para identificar patrones con respuestas copiadas. Para ello, contrastaron el número observado y esperado de ítems con la misma respuesta de dos sujetos. Para cada par de sujetos, uno de ellos es el sujeto que copia (C) y el otro es el sujeto del que se copia la respuesta o sujeto *fuente* (F). Ya que las respuestas de los dos sujetos pueden ser iguales o no, el procedimiento de comparación de respuestas es un ensayo de Bernoulli. Por lo tanto, suponiendo que las covarianzas de los ítems son insignificantes, el índice g_2 es la diferencia entre el número observado y esperado de respuestas idénticas dividido por la desviación típica de la diferencia, un índice que sigue una distribución normal. No obstante, Frary *et al.* afirmaron que la distribución de g_2 podría alejarse de la normalidad si los algoritmos utilizados para calcular la probabilidad de que el sujeto C escoja la misma respuesta que el sujeto F no son los adecuados.

Estadísticos basados en Modelos de Respuesta al Ítem No Paramétricos

Van der Flier (1980, 1982) desarrolló el índice U_3 para probar el ajuste de personas dentro del campo de los modelos de medida no paramétricos, concretamente con el Modelo de Homogeneidad Monótona (MHM) de Mokken (1971). El índice U_3 toma como criterio a un grupo normativo para identificar PRA, el cual es el punto de referencia para comparar un patrón de respuestas observado con puntuación $X_{i.}$, con el patrón esperado en dicho grupo con la misma puntuación. El rango de valores de U_3 oscila entre 0 y 1, de modo que si $U_3 = 0$ indica que el patrón de respuestas es un patrón Guttman perfecto y si $U_3 = 1$ el patrón de respuestas es el patrón inverso al que le correspondería en el escalograma. Cuanto más se aleja U_3 de 0, más se aleja el patrón de ser un patrón Guttman perfecto y más atípico resulta. Pero U_3 depende de los valores de habilidad, por lo que el resultado podría llevar a errores en su interpretación. Para eliminar esta supeditación, van der Flier (1982) lo estandarizó y logró otro estadístico, denotado Z_{U_3}, distribuido según la ley normal.

Estadísticos desarrollados dentro de la Teoría de la Generalizabilidad

La Teoría de la Generalizabilidad (TG) fue elaborada por Cronbach, Gleser, Nanda y Rajaratnam (1972) para resolver las limitaciones de medida de la TCT cuando se utilizaban tests referidos al criterio y tests de dominio o maestría, ya que éstos estudian al sujeto y las conclusiones se elaboran respecto a él y no respecto a un grupo de referencia estandarizado. Kane y Brennan (1980) desarrollaron tres medidas de fiabilidad para tests referidos al dominio o tests de maestría. Aplicando los conceptos de la TG, sean dos tests de maes-

tría J y K compuestos por n y $n^{|}$ ítems, respectivamente, extraídos del mismo universo de ítems para medir el mismo dominio o nivel de maestría. Los tests J y K son aleatoriamente paralelos porque han sido construidos desde el mismo universo, son independientes, tienen igual número de ítems ($n = n^{|}$) e igual puntuación de corte (λ) para evaluar la maestría. Entonces, la *función de acuerdo referida al dominio* A_i del sujeto i evalúa la desviación de la puntuación media del sujeto en ambos tests, que será positiva y en el mismo sentido si en ambos tests se clasifica al sujeto maestro o no maestro, o negativa si en ambos tests la clasificación es diferente, i.e., existe desacuerdo en las puntuaciones medias del sujeto y, en consecuencia, las desviaciones son grandes y de sentido contrario. Si los recesos de la puntuación de corte son próximos a 0, A_i es próxima a 0 y se clasifica al sujeto de caso *borderline*. El segundo índice, la *función de acuerdo máximo esperado* $A_{max,i}$ mide el grado de acuerdo de la puntuación media de un sujeto en un test consigo misma.

Kane y Brennan (1980) consideraron que también era importante evaluar el grado de coherencia de un sujeto para contestar un test de maestría, ya que cuando éstos se emplean no interesa la posición relativa de un sujeto con respecto al resto de la población, sino la puntuación del sujeto en términos absolutos. Por ello, definieron la *función de desacuerdo esperado* L_i que se obtiene sustrayendo del acuerdo máximo esperado el acuerdo esperado.

Los estadísticos de medición apropiada de la Teoría de Respuesta al Ítem

Los estudiosos e investigadores de la TRI que, desde hace poco más de 20 años, se han interesado por identificar a aquellos sujetos cuyos patrones de respuesta no se ajustan al modelo seleccionado, han desarrollado una serie de procedimientos y estadísticos para tal fin. Su objetivo es detectar los patrones que rompen la relación que debe existir entre ellos y el nivel de habilidad o rasgo del sujeto al que corresponde dicho patrón; en definitiva, evaluar el ajuste entre el MRI y el sujeto. Al amparo de los supuestos de esta teoría pertenecen los índices que a continuación se describen.

Extensión de los índices de precaución. Tatsuoka (1984), y Tatsuoka y Linn (1983) establecieron un nexo entre los estadísticos para detectar PRA por comparación con un grupo normativo y los supuestos de la TRI, buscando una correspondencia entre la curva-S, el índice de precaución C_i de Sato (1975) y las curvas características de la TRI. Estos índices de precaución ECI (*Extended Caution Index*) se clasificaron en dos categorías. La primera incluye a $ECI2$ y $ECI3$ por ser medidas referidas al grupo que miden la relación entre el patrón de respuesta observado del sujeto i y la variable normalizada derivada del grupo al que pertenece; ambos son similares a NCI de Tatsuoka y Tatsuoka (1982) y, conceptualmente, a r_{bisper} de Donlon y Fischer (1968). En la segunda categoría se encuentran $ECI4$ y $ECI5$, orientados al sujeto por comparar el patrón de respuestas observado con el patrón teórico que le corresponde según la Curva Característica de Persona (CCP) en un nivel θ fijado; estos estadísticos están relacionados con el procedimiento de Trabin y Weiss (1979) que se describirá más adelante.

Tatsuoka (1984, 1996) desarrolló la estandarización de algunos ECI anteriores debido a la dependencia que éstos tienen de los valores de habilidad. Además, elaboró el índice $ECI6$ con objeto de obtener información acerca del sujeto, comparando el vector de respuesta observado con las probabilidades de respuesta $P_{i.}$ obtenidas con el MRI.

El análisis de residuales. Trabajando con el modelo de Rasch, Wright y Stone (1979) investigaron acerca de cómo este modelo podía explicar la factibilidad de un patrón de respuestas. Siguiendo las mismas suposiciones de Sato (1975) pero aplicando los principios de la TRI, un patrón *esperado* sería el correspondiente o bien a sujetos más hábiles que tienen mayor probabilidad de acertar ítems fáciles que sujetos menos hábiles, o bien a sujetos que ante dos ítems de distinta dificultad tienen mayor probabilidad de acertar el ítem fácil que el ítem con mayor dificultad. Las desviaciones de lo que sería un patrón esperado, patrón *inesperado*, se pueden detectar mediante el cálculo de un índice de residuales estandarizado (Z_{ij}) en términos de la habilidad del sujeto y dificultad del ítem. Pero para justificar el patrón de respuestas, más adecuado que Z_{ij} es su media cuadrática, el índice U_i que sigue una distribución χ^2_ν con grados de libertad igual al número de ítems del test menos 1 ($\nu = n-1$), con el que se sabría cómo de inesperado es el patrón de respuestas. Al probar este índice para detectar PRA, los autores apreciaron que se solía rechazar la hipótesis nula cuando la habilidad del sujeto y la dificultad del ítem eran dispares entre sí. Para solventar este inconveniente, Wright y Masters (1982) ponderaron el índice U_i; el nuevo índice W_i es la media cuadrática ponderada de los residuales, en donde el factor de ponderación son las varianzas de cada uno de los cuadrados de los residuales. Cuando los datos se ajustan al modelo, el índice W_i sigue aproximadamente una distribución de media 1 y varianza $q_i^2 = \left\{\sum_{j=1}^{n}\left[C_{ij} - Var_j\left(\hat{\theta}\right)^2\right]\right\}\left\{\left[\sum_{j=1}^{n} Var_j\left(\hat{\theta}\right)\right]^2\right\}^{-1}$, donde C_{ij} es el índice de curtosis de la respuesta del sujeto

al ítem. Por conveniencia y para poder comparar distintos patrones de respuesta, Wright y Masters (1982), y Wright y Stone (1979) expresaron U_i y W_i como estadísticos normalizados con media 0 y desviación típica 1 (ZU_i y ZW_i).

Rudner (1983) generalizó los estadísticos del análisis de residuales al modelo de 3-p. Los dos índices que propuso fueron el residual estandarizado de la media cuadrática F_1 y, W_3 proporcional a W_i, basado en la media cuadrática ponderada de ajuste para el modelo de 3-p. La media de W_3 es 1, por lo que valores superiores a éste indican que el patrón de respuesta no se ajusta al modelo y, por lo tanto, es un PRA; si $W_3 < 1$ se considera al patrón observado en acuerdo con el esperado según el modelo.

Smith (1985) propuso dos estadísticos relacionados con los dos índices anteriores trabajando con subtests. Si un test de n ítems es dividido en S subtests ($s = 1,2,...,S$) que contienen k ítems, el estadístico de residuales intrasubtests no ponderado de ajuste para un patrón es el estadístico UB_i que sigue, según los análisis posteriores de Kogut (1988), una distribución χ^2 con S grados de libertad si se emplea θ ó $S-1$ grados de libertad si se trabaja con $\hat{\theta}$ estimada por el método de máxima verosimilitud (MV). El estadístico de residuales intersubtests ponderado por el número de ítems del subtest en estudio es UW_i.

La Curva de Respuesta de Persona (CRP). Para estudiar el ajuste de las respuestas de un sujeto al MRI, Trabin y Weiss (1979, 1983) apostaron por un procedimiento gráfico al que nombraron Curva de Respuesta de Persona (CRP). Existen dos CRP por sujeto, una que se obtiene a partir de los datos empíricos –la CRP observada– y otra resultado del MRI al que se supone que se ajustan los datos –la CRP esperada–. Para graficar la CRP observada o empírica, primero hay que agrupar los ítems del test en S subtests representativos de cada nivel de dificultad y disponerlos en orden de dificultad creciente, con la particularidad de que todos los subtests contengan el mismo número de ítems (k). Dentro de cada subtest, los ítems se ordenan por su grado de discriminación desde el ítem más discriminativo al de menor discriminación. A continuación, se calcula la proporción de respuestas correctas en cada uno de los subtests. La CRP aparece al unir estas proporciones (eje de ordenadas) como función de los niveles de dificultad de los ítems (eje de abscisas).

Para poder tomar decisiones acerca de lo típico o atípico del patrón de respuestas, la CRP observada debe ser comparada con una CRP esperada en los mismos ejes de coordenadas. La CRP esperada se obtiene una vez calculadas las probabilidades esperadas de acertar cada uno de los ítems del test [$P_j(\hat{\theta})$] según el modelo de 1-p, 2-p ó 3-p, previamente estimados los parámetros de habilidad y conocidos a priori los parámetros de los ítems. La comparación de ambas curvas podría ser un indicio del grado de ajuste del patrón de respuestas. Sin embargo, esto sería una prueba intuitiva carente de fundamento para catalogar a un patrón de típico o atípico. Por esto, los autores propusieron realizar una prueba χ^2 de bondad de ajuste sobre el índice $D(\hat{\theta})$ de las diferencias entre las curvas observada y esperada; $D(\hat{\theta})$ se contrasta con un valor χ^2_ν con grados de libertad $\nu = S-1$. Si la prueba es estadísticamente significativa, el patrón de respuesta es atípico y, evaluando las diferencias entre las proporciones medias de aciertos por subtests, el investigador podría describir el patrón de respuestas como consecuencia de que el sujeto haya contestado al azar, descuidado sus respuestas o haya hecho trampa.

Sin embargo, la distribución condicional de $D(\hat{\theta})$ es desconocida y no se podría asegurar que fuera un índice asintóticamente estandarizado independiente de θ. Klauer y Rettig (1990) aportando tres pruebas estadísticas asintóticas estandarizadas; la primera de ellas, el estadístico χ^2_{SC} de bondad de ajuste por subtest es similar al propuesto por Trabin y Weiss (1979, 1983), el cual se contrasta con una χ^2_ν con $\nu = S-1$ grados de libertad; si el objetivo era comparar las habilidades estimadas en los subtests, la segunda prueba es el estadístico χ^2_W de Wald, distribuido según χ^2_ν con grados de libertad $\nu = S-1$; el tercer índice es el criterio de razón de verosimilitud de Neyman-Pearson χ^2_{LR} y, de nuevo, la prueba estadística sobre el patrón de respuestas se contrasta con una χ^2_ν cuyos grados de libertad son $\nu = S-1$.

Del estudio de simulación Montecarlo que llevaron a cabo Klauer y Rettig (1990) se dedujo que $\chi^2_{SC} \leq \chi^2_{LR} \leq \chi^2_W$, los resultados de χ^2_{SC} eran más estables y χ^2_W fue la menos robusta frente a una mala aproximación de la estimación de la habilidad a la distribución normal.

Sijtsma y Meijer (2001) han adaptado el procedimiento de identificación de PRA mediante la CRP al ámbito de los modelos de respuesta no paramétricos.

Aunando el Modelo de Doble Monotonicidad de Mokken (1971) y el concepto de CRP, Emons, Sijtsma y Meijer (2004) han probado el modelo de regresión y *kernel smoothing* para detectar PRA. Ambos son procedimientos gráficos; por el primero se contrastan los modelos de regresión logística saturado y reducido una vez estimados los parámetros de los

dos modelos; por el segundo, se comparan la CRP observada y esperada estimadas mediante *kernel smoothing*.

El método de comparación de las curvas características de los ítems. Rosenbaum (1987) sugirió que a partir de las curvas características de los ítems (CCI) se podrían identificar PRA siempre y cuando los ítems y las respuestas de los sujetos estén definidos en una escala latente, es decir, que estén ordenados en dificultad creciente. Esta ordenación es la misma que la del escalograma de Guttman (1944, 1950) con la diferencia de que éste es un modelo determinístico y Rosenbaum emplea modelos de variable latente –el modelo de Rasch, los modelos de doble monotonicidad de Mokken, el modelo logístico de 2-p, algunos modelos multidimensionales…– para la detección de los patrones atípicos.

Sea un sujeto i con un patrón de respuestas en n ítems dicotómicos $U_i = u_1, u_2, ..., u_n$ y $P(U = u)$ es la distribución de ese vector de respuestas en una población; entonces, un modelo de variable latente establece que para el vector U hay asociada una distribución condicional para cada uno de los sujetos que depende de θ, variable latente e inobservable característica de cada sujeto. De acuerdo con este supuesto, el ítem j es de modo uniforme más difícil que el ítem k si $P(u_j = 1 \mid \theta) \leq P(u_k = 1 \mid \theta)\ \forall \theta$, i.e., la CCI del ítem j está por debajo de la CCI del ítem k. Gráficamente, si las CCIs no se cruzan para un determinado θ, los patrones de respuesta son normales, pero si se cruzan serán un indicador de un PRA.

Los estadísticos de curvatura de la función de verosimilitud. Drasgow, Levine y McLaughlin (1987) propusieron dos estadísticos para identificar PRA basados en la función de verosimilitud. Una respuesta atípica influye en la forma de la función de verosimilitud alejándola del punto máximo, achatando la curva debido a que no hay ningún valor de habilidad que permita al MRI seleccionado ajustarse al patrón de respuestas. El primer estadístico de curvatura es la estimación de la varianza *jackknife* normalizada, la cual se obtiene tras estimar dos parámetros de habilidad por MV con el modelo de 3-p: $\hat{\theta}$ para el test de longitud n y $\hat{\theta}_{(j)}$ para el test con $n-1$ ítems por eliminación del ítem j Los *pseudo-valores* son $\hat{\theta}_j^* = n\hat{\theta} - (n-1)\hat{\theta}_{(j)}^*$ para $j = 1, 2, ..., n$.

La varianza jackknife del parámetro de habilidad $\sigma^2(\hat{\theta}^*)$ no es un índice de medición apropiada ya que depende de θ. Para saldar este problema, los autores recurrieron al empleo de la función de información (FI), ya que ésta es el recíproco de la varianza asintótica de $\hat{\theta}$ y así la estimación jackknife puede ser normalizada. Entonces, el índice de ajuste del patrón de respuestas $JK = \sigma^2(\hat{\theta}^*)I(\hat{\theta})$. Cuando el patrón de respuestas no se ajusta al modelo, la función de verosimilitud es relativamente plana y la estimación de la varianza es mayor que para un patrón de respuesta normal.

El segundo índice (O/E) compara las curvaturas de las funciones de verosimilitud esperada y observada. Si la función de verosimilitud es más chata para el PRA que para el patrón de respuesta normal, entonces la FI observada será menor que la FI esperada. En consecuencia, el índice de medición apropiada es una razón de funciones de información.

En el estudio en el que Drasgow, Levine y McLaughlin (1987) pusieron a prueba los índices JK y O/E concluyeron que, a pesar de estar estandarizados, padecían de una pobre identificación de PRA.

Los estadísticos de ajuste de persona óptimos. Levine y Drasgow (1984), y Drasgow, Levine y Zickar (1996) consideraron que un índice de medición apropiada era *óptimo* cuando ningún otro índice obtenía tasas de identificaciones correctas de PRA mejores que él, i.e., cuando tiene efectividad absoluta para detectar un determinado tipo de respuesta atípica. El estadístico de ajuste óptimo es una razón de probabilidad basada en el lema de Neyman-Pearson, por el cual se contrasta la probabilidad de ocurrencia de un patrón de respuesta normal según un MRI [$P_n(U)$] frente a la de patrón según un modelo de respuesta atípico [$P_a(U)$]. Para este último, existen $S_k = \binom{n}{m}$ formas de escoger m de los n ítems del test para generar una respuesta atípica. La probabilidad $P_a(U)$ se calcularía mediante un algoritmo numérico de cuadratura. El índice de ajuste óptimo es el resultado de la razón de probabilidades $\lambda(U) = P_a(U)/P_n(U)$. Si $\lambda(U)$ se aleja de 1 el patrón es atípico. Además, el lema de Neyman-Pearson aporta una prueba estadística en la que se fija la tasa de error Tipo I para rechazar la hipótesis nula de que el patrón es normal en un valor $\alpha = cteP_n(U)$, rechazando la hipótesis nula si $P_a(U) \geq cteP_n(U)$.

Drasgow, Levine y Zickar (1996) elaboraron cinco modelos de respuesta atípica consecuencia de hacer trampa porque se conocen las respuestas, por falsificar respuestas en escalas de personalidad y en inventarios biográficos, por la inexperiencia con los ordenadores y con los tests informatizados, por copiar respuestas en un test desconocido y para espurias bajas según la definición de Levine y Rubin (1979).

Otro estadístico óptimo de ajuste que contempla un algoritmo para crear modelos para respuestas atípicas aplicado al modelo de Rasch fue elaborado por Klauer (1995). Este autor definió modelos alternativos a las generalizaciones biparamétricas del modelo de Rasch que contienen un paráme-

tro específico de persona (η_i) añadido al parámetro de habilidad θ_i. El parámetro η_i describe la magnitud y dirección de las desviaciones del modelo original.

Los estadísticos basados en la función de verosimilitud. Levine y Rubin (1979) barajaron la posibilidad de que la existencia de más de un rasgo latente por sujeto en la ejecución del test fuera la causa de PRA. Por la supuesta presencia de un parámetro de habilidad variable, los MRI clásicos –o como los autores denominaron *modelos estándar o constantes*– no serían válidos para evaluar el grado de acuerdo entre un patrón de respuestas y el nivel de habilidad del sujeto, ya que para ellos la habilidad θ de un sujeto es constante en todos los ítems del test. Si se asume un modelo de respuesta que contemple valores de habilidad θ_i independientes en cada uno de los ítems del test, valores que en conjunto siguieran una distribución normal de media θ_0 y varianza σ^2, este modelo se ajustaría mejor al patrón de respuesta del sujeto que un modelo constante. A este modelo alternativo que permite valores de habilidad variables lo llamaron *modelo gaussiano*. La relación entre ambos es que el modelo constante es el caso límite del gaussiano cuando $\sigma^2 = 0$.

A partir del modelo gaussiano, Levine y Rubin (1979) propusieron el índice de medición apropiada l_0 para detectar aquellos patrones de respuesta que no estaban en concordancia con los niveles de habilidad del sujeto, obteniendo una medida de la bondad de ajuste de un modelo psicométrico al patrón de respuestas individual ítem-por-ítem mediante el cálculo del logaritmo natural de la función de verosimilitud. Este índice depende de la capacidad del MRI para pronosticar, en función de θ_i, la respuesta dada por el sujeto. Cuanto más se aleje el logaritmo de la función de 1 tanto más atípico será el patrón.

Drasgow (1982) amplió el trabajo de Levine y Rubin (1979) al cuestionarse cuál sería el MRI que se debería ajustar a los datos. Si bien el modelo de 3-p fue el primero que se empleó para describir la probabilidad de respuesta en tests de elección múltiple, dicho modelo ha sido criticado por las dificultades que conlleva la estimación de los parámetros de los ítems con MV, concretamente la del parámetro de pseudo-azar (Lord, 1968, p. 1014). Por esta razón, Drasgow plantea la posibilidad de poder emplear el modelo de Rasch en tests de elección múltiple con el fin de obtener una óptima estimación de los parámetros y, en consecuencia, una mejoría de la tasa de identificaciones de PRA. Junto con esta hipótesis sobre el modelo de respuesta, este autor modificó el estadístico l_0 para paliar los efectos de la longitud del test y del número de ítems omitidos por los que se dejaba afectar. El nuevo índice es la media geométrica de l_0, $l_g = \exp^{l_0/m}$, donde m es el número de ítems contestados en el test.

El índice l_0 obtenía un alto porcentaje de identificación de PRA pero presentaba dos inconvenientes: a) no estaba estandarizado, lo cual implica que clasificar un patrón de respuesta como normal o atípico dependa de θ, y b) la distribución de l_0 era desconocida siendo ésta necesaria para poder probar la hipótesis nula de que el patrón de respuestas es normal. Para solucionar estos dos problemas Drasgow, Levine y Williams (1985) propusieron la versión estandariza del estadístico l_0. El objetivo era poder comparar los valores del índice de medición apropiada de sujetos en distintos niveles de habilidad. Suponiendo que l_0 seguía una distribución normal, derivaron analíticamente su esperanza matemática y desviación típica, dotando a este nuevo índice (l_z) de una prueba estadística que compara los valores observados con los teóricos, ya que l_z sigue una distribución normal de media 0 y desviación típica 1 si los datos se ajustan al MRI.

Molenaar y Hoijtink (1990, 1996) cuestionaron la capacidad de l_z para detectar PRA cuando se trabaja con el modelo de Rasch, con parámetros de habilidad estimados por MV y con tests cortos (50 ítems o menos) ya que, bajo estas condiciones, la distribución de l_z se aleja de la normalidad. Para cuando el test tiene pocos ítems ($n \leq 20$) o cuando en tests largos los ítems tienen parámetros de dificultad con un amplio rango de variabilidad, el estadístico propuesto para evaluar el ajuste del patrón fue $M(U)$ y su prueba estadística es una χ^2.

Snijders (2001), partiendo de los estudios y de la misma idea que Molenaar y Hoijtink (1990, 1996), ha propuesto el índice l_z^* para emplearlo cuando el parámetro de habilidad es desconocido y debe ser estimado, y han ampliado su utilidad a los modelos de 2-p, 3-p y de respuesta politómica. El estadístico l_z^* sigue una distribución próxima a la normal tipificada, ya que su varianza se aleja de 1 cuando aumenta el número de ítems del test y, en general, es sesgada negativa y leptocúrtica.

Drasgow, Levine y McLaughlin (1991) extendieron el uso del estadístico l_z al ámbito de los tests *multiunidimensionales* (l_{zm}). Un test multiunidimensional es un test multidimensional cuyos ítems cumplen el supuesto de independencia local y está dividido en subtests unidimensionales.

El estadístico ω. Wollack (1997) retoma la idea sobre la que Frary *et al.* (1977) elaboraron el índice g_2. Cuando los datos son explicados por el modelo de respuesta nominal de

Bock (1972), el estadístico ω de Wollack pretende detectar PRA, en concreto, patrones con respuestas copiadas. A todos los sujetos que contestan el test se les considera posibles copiadores (C) y sus respuestas son contrastadas con los patrones de sujetos fuente (F) según la disposición de asientos en el momento de la realización del test. El estadístico ω compara el número de ítems con la misma respuesta para cada par de sujetos y el número esperado a consecuencia del azar; ω sigue una distribución normal estandarizada ya que, asumiendo que las respuestas a los ítems son localmente independientes, el número de ítems contestados con la misma opción (h_{CF}) es la suma de n ensayos de Bernoulli independientes y, por el teorema central del límite, la distribución de h_{CF} se aproxima a la normal cuando el número de ítems tiende a infinito.

Los estadísticos para Tests Adaptativos Informatizados (TAI)

Debido al creciente uso de los TAI o CAT (*Computerized Adaptive Testing*), Bradlow, Weiss y Cho (1998) propusieron un estadístico *outlier* para identificar diferentes clases de PRA en este tipo de tests. En los TAI, los ítems son seleccionados para maximizar la información acerca de la habilidad del sujeto en cada ítem administrado, por lo que, en un diseño TAI perfecto siempre $P_j(\theta_i) = 0'5$, es decir, $b_j \approx \theta_i$ y al concluir el test el promedio de dificultad de los ítems es aproximadamente igual al nivel de habilidad del sujeto. Bradlow *et al.* estaban interesados en detectar sujetos con patrones outlier multivariantes a partir del estadístico para patrones de respuesta univariante basado en la suma acumulada $w_h = \sum_{j=1}^{h} u_j$. A este estadístico lo denotaron $K(\theta)$, el cual sigue una distribución normal.

McLeod y Lewis (1999) idearon un estadístico Z_c de medición apropiada adecuado para identificar PRA en los TAI resultantes de la memorización de los ítems más difíciles del test que tienen mayor frecuencia de aparición. El estadístico Z_c es una extensión del índice de precaución $ECI4_z$ de Tatsuoka y Linn (1983).

Van Krimpen-Stoop y Meijer (1999, 2000) simularon una distribución T para detectar PRA basándose en l_z de Drasgow, Levine y Williams (1985), y l_z^* de Snijders (2001). Van Krimpen-Stoop y Meijer (2001, 2002) han vuelto a retomar el estadístico l_z^* de Snijders para utilizar el procedimiento de suma acumulada CUSUM (*Cumulative Sum*) de Page (1954) para detectar PRA en TAIs con ítems dicotómicos y politómicos. El inconveniente de este procedimiento es, que los valores críticos para la toma de decisiones acerca de la atipicidad de un patrón, se determinan mediante simulación.

Bradlow y Weiss (2001) han listado una serie de estadísticos para identificar patrones outliers y han propuesto métodos de normalización de los mismos para estandarizarlos.

Recientemente, McLeod, Lewis y Thissen (2003) han ideado un índice basado en la razón de probabilidad para detectar a aquellos sujetos que se someten a un TAI y, previamente, conocen el contenido de algunos de los ítems del banco que se han mostrado en ocasiones anteriores a otros sujetos. El *FLOR* (*Final Log Odds Ratio*), elaborado a partir del teorema de Bayes, compara la probabilidad de acertar un ítem como consecuencia de su memorización, por conocimiento a priori del mismo, con la probabilidad de responderlo por propia habilidad. Su expresión es la que sigue:

$$FLOR = \log_{10}\left[\frac{p(s \mid u_1,\ldots,u_m)/[1-p(s \mid u_1,\ldots,u_m)]}{p(s)/[1-p(s)]}\right]$$

donde, $p(s \mid u_1,\ldots,u_m)$ es la probabilidad de que el sujeto esté empleando el conocimiento acerca de los ítems de banco (memorización) después de la administración del ítem *m*; $1-p(s \mid u_1,\ldots,u_m)$ es la probabilidad de que el sujeto no esté empleando el conocimiento acerca de los ítems de banco (memorización) después de la administración del ítem *m*; $p(s)$ es la probabilidad esperada o probabilidad a priori de que un sujeto haya tenido la oportunidad de memorizar los ítems, valor derivado de la modificación del modelo de 3-p. El valor $p(s)$ es un valor específico (p.e., 0'0001), fijado bien a raíz de la evidencia empírica para detectar PRA debidos a la copia en tests tradicionales bien a partir de la literatura acerca de la teoría de la decisión. Si *FLOR* =1, hay 10 veces más sospechas de que el sujeto esté haciendo trampa antes de que el último ítems sea administrado; si *FLOR* =0, indica que no hay porqué tener sospechas de que el sujeto esté memorizando los ítems; y si *FLOR* =-1, hay 10 veces menos sospechas de que el sujeto esté haciendo trampa antes de que el último ítem sea administrado o, lo que es lo mismo, una razón negativa evidencia que el sujeto no se está copiando.

La ventaja que presenta el índice *FLOR* es que, en su cálculo, incorpora tanto las restricciones del TAI como el algoritmo de selección de los ítems del banco.

El índice de ajuste de persona según el Análisis de Estructuras de Covarianza

Este índice propuesto por Reise y Widaman (1999) valora el ajuste al modelo de AEC estimado a nivel de sujeto para obtener la proporción de éstos que no se ajustan a dicho modelo. Un modelo de AEC es un modelo lineal que pretende, o bien enlazar las variables latentes y las observadas, o bien especificar la relación entre varias variables latentes o entre varias variables observadas. Para evaluar el ajuste del sujeto al

modelo, estos autores propusieron un índice basado en el logaritmo de la función de máxima verosimilitud, IND_{χ^2}, cuya argumentación es similar a la de l_0 de Levine y Rubin (1979), y se define a partir de la diferencia para cada sujeto entre los logaritmos de la función de verosimilitud calculados con un modelo sustantivo (modelo de un factor) y con un modelo saturado (modelo que reproduce exactamente la matriz de covarianzas). El índice IND_{χ^2} es una razón de probabilidad que se contrasta con una prueba de ajuste χ^2 con grados de libertad igual a la diferencia en el número de parámetros estimados en el modelo saturado y en el modelo sustantivo. Valores positivos y elevados de IND_{χ^2} aparecerían en sujetos con PRA o en sujetos que no contribuyen al ajuste del modelo.

Estadísticos de ajuste de persona empleando Modelos de Clase Latente de Orden Restringido

Los Modelos de Clase Latente de Orden Restringido (MCL-OR), aplicados al ámbito de la medición apropiada, se podrían delimitar como un grupo intermedio de modelos entre los de la TRI paramétricos y no paramétricos. De los primeros, comparten la posibilidad de precisar las distribuciones muestrales de los estadísticos de medición apropiada y, de los segundos, el moldeamiento a patrones de respuesta que no se adaptan a las restricciones marcadas por los modelos paramétricos.

Emons, Glas, Meijer y Sijtsma (2003) han redefinido el índice basado en el número de errores Guttman G_i^* de Meijer (1994), el estadístico U_3 de van der Flier (1980), el estadístico l_0 de Levine y Rubin (1979) y el índice $ECI6$ de Tatsuoka (1984), para adaptarlos a los MCL-OR. A estos índices los han denotado como $G^*(q)$, $U_3(q)$, $LogL(q)$ y $\varsigma(q)$, respectivamente, donde q ($q = 1,2,\ldots,Q$) es cada una de las clases latentes que se corresponde con un punto del continuo de θ. Para la detección de PRA, los autores implementan la prueba bayesiana predictiva *a posteriori* para obtener las distribuciones de los citados estadísticos de medición apropiada.

Discusión

Los métodos de ajuste de persona pretenden identificar a aquellos patrones de respuesta que o bien son incoherentes con los patrones de la muestra normativa a la que pertenecen o bien un MRI no se ajusta a ellos. La identificación y el tratamiento de PRA es de gran utilidad para la construcción de tests y el uso de bancos de ítems con propiedades psicométricas, así como para el análisis de la validez de los mismos. Todos ellos han sido empleados e incluso comparados en efectividad en tests simulados (Emons *et al.*, 2003; Emons *et al.*, 2004; Hendrawan, Glas y Meijer, 2005; Karabatsos, 2003; Li y Olejnik, 1997; McLeod, Lewis y Thissen, 2003; Nering, 1995, 1997; Nering y Meijer, 1998; Noonan, Boss y Gessaroli, 1992; Meijer, 2003; Núñez, 2002; Reise, 1990; Rogers y Hattie, 1987; van Krimpen-Stoop y Meijer, 1999, 2002) y con patrones de respuesta reales o simulados de tests de actitudes (Meijer, 2002; Parsons, 1983), de personalidad (Ferrando, 2004; Meijer y Nering, 1997; Reise, 1995; Reise y Flannery, 1996; Reise y Waller, 1993; Reise y Widaman, 1999; Schmitt, Chan, Sacco, McFarland y Jennings, 1999; Zickar y Drasgow, 1996), y de tests cognitivos y de aptitudes (Birenbaum, 1986; Drasgow, 1982; Drasgow y Levine, 1986; Drasgow, Levine y McLaughlin, 1987, 1991; Drasgow, Levine y Williams, 1985; Harnisch y Linn, 1981; Harnisch y Tatsuoka, 1983; Klauer y Rettig, 1990; Levine y Drasgow, 1982, 1983a, 1983b; Levine y Rubin, 1979; McLeod y Lewis, 1999; Meijer, 1994; Meijer, Muijtjens y van der Vleuten, 1996; Meijer y Nering, 1997; Meijer, Sijtsma y Smid, 1990; Miller, 1986; Molenaar y Hoijtink, 1996; Rudner, Bracey y Skaggs, 1996; Schmitt *et al.*, 1999; Tatsuoka, 1996; Tatsuoka y Tatsuoka, 1982, 1983; Trabin y Weiss, 1983; van der Flier, 1982; Wollack, 1997; Wollack, Cohen y Serlin, 2001; Wright y Masters, 1982; Wright y Stone, 1979).

Tras haber descrito todos ellos, el investigador interesado en la detección de PRA se enfrenta ante la duda de cuál de ellos escoger para el análisis. La elección de un método u otro depende de si se toma como criterio comparativo a un grupo normativo o si se evalúa el ajuste de un MRI a los datos. Para el primer caso, hay que puntualizar que, con excepción de Z_{U_3}, la distribución teórica de estos estadísticos es desconocida y que la distribución de los datos empíricos depende de la puntuación total, por lo que las decisiones acerca de la atipicidad de un patrón se quedan en un plano meramente descriptivo. Dentro de la TRI, la opción del uso de un índice para la identificación de PRA estaría en función de varios factores: a) el hipotético tipo de patrón atípico que podría mostrarse –azar, copia, errores de transcripción, impericia,…– según el constructo o variable que se examina y del tipo de test –en la literatura, los PRA por excelencia han sido los de azar y copia–; b) el modelo de respuesta –modelo de Rasch, modelo logístico de 2-p, modelo de 3-p, modelos no paramétricos,…–; c) el valor verdadero de habilidad del sujeto –extremos *vs.* moderados–; d) el método de estimación de los parámetros de habilidad; e) la longitud del test, directamente relacionada con el punto anterior.

Para calcular estos índices, se han creado algunos programas informáticos:

- FORTRAN TCC3 de Baillie y Tatsuoka (1982) para $ECI1_z$, $ECI2_z$, $ECI4_z$ de Tatsuoka (1984, 1996) y l_z, utilizado por Birenbaum (1986).

- Un programa elaborado por Drasgow (1985) que calcula l_z, $ECI4_z$ y W_3 de Rudner (1983), e implementado por Noonan *et al.* (1992).
- RASCH/ECIZ de Nelson y Chatman (1985) para hallar los índices U_i de Wright y Stone (1979), *ECI2*, *ECI4* de Tatsuoka y Linn (1983), $ECI2_z$ y $ECI4_z$.
- IPARM de Smith (1991) para el análisis de residuales con *UB* y *UW* de Smith (1985).
- RSP de Glas y Ellis (1994) calcula los índices de precaución de Tatsuoka (1984) adaptados al modelo de Rasch.
- WPERFIT de Ferrando y Lorenzo (2000) para el cómputo de l_z, $ECI4_z$ y la prueba χ^2_{SC} de bondad de ajuste para la CRP de Trabin y Weiss (1979, 1983; Klauer y Rettig, 1990).
- X-PAT de Doval, Núñez, Renom y Solanas (2001) analiza patrones de respuesta atípicos mediante el número de errores de Loevinger (1947, 1948), n_{bisper} de Donlon y Fischer (1968), C_i de Sato (1975), U_1^* de van der Flier (1977), C_i^*de Harnisch y Linn (1981) y *NCI* de Tatsuoka y Tatsuoka (1982).

Pese a la variedad de métodos de ajuste de persona, este ámbito de estudio de la Psicometría requiere todavía de más investigación que afiance aspectos tales como las características distribucionales que poseen cuando se emplean las estimaciones de los parámetros así como su extensión a modelos de respuesta politómica y no paramétricos, también examinar la posible relación entre el funcionamiento diferencial de los ítems y la presencia de PRA, y ahondar más en las peculiaridades de un PRA en relación con la variable psicológica evaluada y con el tipo de test.

Referencias

Angoff, W.H. (1982). Uses of difficulty and discrimination indices for detecting item bias. En R.A. Berk (Ed.), *Handbook of methods for detecting item bias* (pp. 96-116). Baltimore, MD: Johns Hopkins University Press.

Baillie, R. y Tatsuoka, K.K. (1982). *TCC3* [Computer program]. Urbana, IL: University of Illinois at Urbana-Champaign, Computer-based Education Research Laboratory.

Birenbaum, M. (1986). Effect of dissimulation motivation and anxiety on response pattern appropriateness measures. *Applied Psychological Measurement, 10*, 167-174.

Bock, R.D. (1972). Estimating item parameters and latent ability when responses are scored in two or more latent categories. *Psychometrika, 37*, 29-51.

Bradlow, E.T. y Weiss, R.E. (2001). Outlier measures and norming methods for computerized adaptive tests. *Journal of Educational and Behavioral Statistics, 26*, 85-104.

Bradlow, E.T., Weiss, R.E. y Cho, M. (1998). Bayesian identification of outliers in computerized adaptive tests. *Journal of the American Statistical Association, 93*, 910-919.

Cronbach, L.J., Gleser, G.C., Nanda, H. y Rajaratnam, N. (1972). *The dependability of behavioral measurements: Theory of generalizability for scores and profiles*. New York: Wiley.

Donlon, T.F. y Fischer, F.E. (1968). An index of an individual's agreement with group-determined item difficulties. *Educational and Psychological Measurement, 28*, 105-113.

Doval, E., Núñez, M.I., Renom, J. y Solanas, A. (2001). *X-PAT: Un explorador de patrones de respuestas.* Comunicación presentada en el VII Congreso de Metodología de las Ciencias Sociales y de la Salud, Madrid.

Drasgow, F. (1982). Choice of test models for appropriateness measurement. *Applied Psychological Measurement, 6*, 297-308.

Drasgow, F. y Levine, M.V. (1986). Optimal detection of certain forms of inappropriate test scores. *Applied Psychological Measurement, 10*, 59-67.

Drasgow, F., Levine, M.V. y McLaughlin, M.E. (1987). Detecting inappropriate test scores with optimal and practical appropriateness indices. *Applied Psychological Measurement, 11*, 59-79.

Drasgow, F., Levine, M.V. y McLaughlin, M.E. (1991). Appropriateness for some multidimensional test batteries. *Applied Psychological Measurement, 15*, 171-191.

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

Drasgow, F., Levine, M.V. y Zickar, M.J. (1996). Optimal identification of mismeasured individuals. *Applied Measurement in Education, 9*, 47-64.

Emons, W.H.M., Sijtsma, K. y Meijer, R.R. (2004). Testing hypotheses about the person-response function in person-fit analysis. *Multivariate Behavioral Research, 39,* 1-35.

Emons, W.H.M., Glas, C.A.W., Meijer, R.R. y Sijtsma, K. (2003). Person fit in order-restricted latent class models. *Applied Psychological Measurement, 27*, 459-478.

Frary, R.B., Tideman, T.N. y Watts, T.M. (1977). Indices of cheating on multiple-choice tests. *Journal of Educational Statistics, 2*, 235-256.

Ferrando, P.J. (2004). Person reliability in personality measurement: an item response theory analysis. *Applied Psychological Measurement, 28*, 126-140.

Ferrando, P.J. y Lorenzo, U. (2000). WPERFIT: A program for computing parametric person-fit statistics and plotting person response curves. *Educational and Psychological Measurement, 60*, 479-487.

Glas, C.A.W. y Ellis, J. (1994). Computer programs: RSP. *Rasch Measurement Transactions, 8*, 339-340.

Guttman L. (1944). A basis for scaling qualitative data. *American Sociological Review, 9*, 139-150.

Guttman L. (1950). The basis for scalogram analysis. En S.A. Stouffer, L. Guttman, E.A. Suchman, P.F. Lazarsfeld, S.A. Star y J.A. Clausen (Eds.), *Measurement and prediction. Studies in social psychology in World War II* (Vol. 4) (pp. 60-90). Princeton, NJ: Princeton University Press.

Harnisch, D.L. (1983). Item response patterns: Applications for educational practice. *Journal of Educational Measurement, 20*, 191-206.

Harnisch, D.L. y Linn, R.L. (1981). Analysis of item response patterns: Questionable test data and dissimilar curriculum practices. *Journal of Educational Measurement, 18*, 133-146.

Harnisch, D.L. y Tatsuoka, K.K. (1983). A comparison of appropriateness indices based on item response theory. En R.K. Hambleton (Ed.), *Applications of item response theory* (pp. 104-122). Vancouver, Canada: Kluwer-Nijhoff Publishing.

Hendrawan, I., Glas, C.A.W. y Meijer, R.R. (2005). The effect of person misfit on classification decisions. *Applied Psychological Measurement, 29,* 26-44.

Kane, M.T. y Brennan, R.L. (1980). Agreement coefficients as indices of dependability for domain-referenced tests. *Applied Psychological Measurement, 4*, 105-126.

Karabatsos, G. (2003). Comparing the aberrant response detection performance of thirty-six person-fit statistics. *Applied Measurement in Education, 16*, 277-298.

Klauer, K.C. (1995). The assessment of person fit. En G.H. Fischer y I.W. Molenaar (Eds.), *Rasch models: Foundations, recent developments and applications* (pp. 97-110). New York: Springer-Verlag.

Klauer, K.C. y Rettig, K. (1990). An approximately standardized person test for assessing consistency with a latent trait model. *British Journal of Mathematical and Statistical Psychology, 43*, 193-206.

Kogut, J. (1988). *Asymptotic distribution of a person-fit statistic* (Research Report No. 88-13). Enschede, The Netherlands: University of Twente.

Levine, M.V. y Drasgow, F. (1982). Appropriateness measurement: Review, critique and validating studies. *British Journal of Mathematical and Statistical Psychology, 35*, 42-56.

Levine, M.V. y Drasgow, F. (1983a). Appropriateness measurement: Validating studies and variable ability models. En D.J. Weiss (Ed.), *New horizons in testing: Latent trait test theory and computerized adaptive testing* (pp. 109-131). New York: Academic Press.

Levine, M.V. y Drasgow, F. (1983b). The relation between incorrect option choice and estimated ability. *Educational and Psychological Measurement, 43*, 675-685.

Levine, M.V. y Drasgow, F. (1984). *Performance envelops and optimal appropriateness measurement* (Report No. 84-5). Champaign, IL: University of Illinois, Department of Educational Psychology, Model-based Measurement Laboratory. (ERIC Document Reproduction Service No. ED 263 126).

Levine, M.V. y Rubin, B.D. (1979). Measuring the appropriateness of multiple-choice test scores. *Journal of Educational Statistics, 4*, 269-290.

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

Loevinger, J. (1947). A systematic approach to the construction and evaluation of tests of ability. *Psychological Monograph, 61* (No. 4).

Loevinger, J. (1948). The technique of homogeneous tests compared with some aspects of scale analysis and factor analysis. *Psychological Bulletin, 45*, 507-530.

Lord, F.M. (1968). An analysis of the Verbal Scholastic Aptitude Test using Birnbaum's three-parameter logistic model. *Educational and Psychological Measurement, 28*, 989-1020.

McLeod, L.D. y Lewis, C. (1999). Detecting item memorization in the CAT environment. *Applied Psychological Measurement, 23*, 147-160.

McLeod, L.D., Lewis, C. y Thissen, D. (2003). A bayesian method for the detection of item preknowledge in computerized adaptive testing. *Applied Psychological Measurement, 27*, 121-137.

Meijer, R.R. (1994). The number of Guttman errors as a simple and powerful person-fit statistic. *Applied Psychological Measurement, 18*, 311-314.

Meijer, R.R. (1995). A supplement to "The number of Guttman errors as a simple and powerful person-fit statistic". *Applied Psychological Measurement, 19*, 166.

Meijer, R.R. (2002). Outlier detection in hihg-stakes certification testing. *Journal of Educational Measurement, 39,* 219-233.

Meijer, R.R. (2003). Diagnosing item score patterns on a test using item response theory-based person-fit statistics. *Psychological Methods, 8,* 72-87.

Meijer, R.R., Muijtjens, M.M. y van der Vleuten, C.P.M. (1996). Nonparametric person-fit research: Some theoretical issues and an empirical example. *Applied Measurement in Education, 9*, 77-89.

Meijer, R.R. y Nering, M.L. (1997). Trait level estimation for nonfitting response vectors. *Applied Psychological Measurement, 21*, 321-336.

Meijer, R.R. y Sijtsma, K. (1999). *A review of methods for evaluating the fit of item score patterns on a test* (Research Report No. 99-01). Twente, The Netherlands: University of Twente, Department of Educational Measurement and Data Analysis.

Meijer, R.R. y Sijtsma, K. (2001). Methodology review: Evaluating person fit. *Applied Psychological Measurement, 25*, 107-135.

Meijer, R.R., Sijtsma, K. y Smid, N.G. (1990). Theoretical and empirical comparison of the Mokken and the Rasch approach to IRT. *Applied Psychological Measurement, 14*, 283-298.

Miller, M.D. (1986). Time allocation and patterns of item response. *Journal of Educational Measurement, 23*, 147-156.

Mokken, R.J. (1971). *A theory and procedure of scale analysis*. The Netherlands: Mounton.

Molenaar, I.W. y Hoijtink, H. (1990). The many null distributions of person fit indices. *Psychometrika, 55*, 75-106.

Molenaar, I.W. y Hoijtink, H. (1996). Person-fit and the Rasch model, with an application to knowledge of logical quantors. *Applied Measurement in Education, 9*, 27-45.

Nelson, R.B. y Chatman, S.P. (1985). RASCH/ECIZ: A SAS PROC MATRIX program for Rasch analysis and person-fit statistics. *Applied Psychological Measurement, 9*, 325.

Nering, M.L. (1995). The distribution of person fit using true and estimated person parameters. *Applied Psychological Measurement, 19*, 121-129.

Nering, M.L. (1997). The distribution of indexes of person fit within the computerized adaptive testing environment. *Applied Psychological Measurement, 21*, 115-127.

Nering, M.L. y Meijer, R.R. (1998). A comparison of the person response function and the l_z person-fit statistic. *Applied Psychological Measurement, 22*, 53-69.

Noonan, B.W., Boss, M.W. y Gessaroli, M.E. (1992). The effect of test length and IRT model on the distribution and stability of three appropriateness indexes. *Applied Psychological Measurement, 16*, 345-352.

Núñez, R.M. (2002). *Propiedades distribucionales de un estadístico de medición apropiada.* Tesis doctoral no publicada. Universidad de Murcia.

Page, E.S. (1954). Continuous inspection schemes. *Biometrika, 41*, 100-115.

Parsons, C.K. (1983). The identification of people for whom job descriptive index scores are inappropriate. *Organizational Behaviour and Human Performance, 33*, 365-393.

Reise, S.P. (1990). A comparison of item- and person-fit methods of assessing model-data fit in IRT. *Applied Psychological Measurement, 14*, 127-137.

Reise, S.P. (1995). Scoring method and the detection of person misfit in a personality assessment context. *Applied Psychological Measurement, 19*, 213-229.

Reise, S.P. y Flannery, Wm. P. (1996). Assessing person-fit on measures of typical performance. *Applied Measurement in Education, 9*, 9-26.

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

Reise, S.P. y Widaman, K.F. (1999). Assessing the fit of measurement models at the individual level: A comparison of item response theory and covariance structure approaches. *Psychological Methods, 4*, 3-21.

Rogers, H.J. y Hattie, J.A. (1987). A Monte Carlo investigation of several person and item fit statistics for item response models. *Applied Psychological Measurement, 11*, 47-57.

Rosenbaum, P.R. (1987). Probability inequalities for latent scales. *British Journal of Mathematical and Statistical Psychology, 40*, 157-168.

Rudner, L.M. (1983). Individual assessment accuracy. *Journal of Educational Measurement, 20*, 207-219.

Rudner, L.M., Bracey, G. y Skaggs, G. (1996). The use of a person-fit statistic with one high-quality achievement test. *Applied Measurement in Education, 9*, 91-109.

Sato, T. (1975). *The construction and interpretation of S-P tables*. Tokyo: Meiji Tosho. (En japonés).

Schmitt, N., Chan, D., Sacco, J.M., McFarland, L.A. y Jennings, D. (1999). Correlates of person fit and effect of person fit on test validity. *Applied Psychological Measurement, 23*, 41-53.

Sijtsma, K. (1986). A coefficient of deviance of response patterns. *Kwantitatieve Methoden, 7*, 131-145.

Sijtsma, K. y Meijer, R.R. (1992). A method for investigating the intersection of the item response functions in Mokken's nonparametric IRT model. *Applied Psychological Measurement, 16*, 149-157.

Sijtsma, K. y Meijer, R.R. (2001). The person response function as a tool in person-fit research. *Psychometrika, 66*, 191-208.

Smith, R.M. (1985). A comparison of Rasch person analysis and robust estimators. *Educational and Psychological Measurement, 45*, 433-444.

Smith, R.M. (1991). *IPARM: Item and person analysis with the Rasch model*. Chicago: MESA Press.

Snijders, T.A.B. (2001). Asymptotic null distribution of person fit statistics with estimated person parameter. *Psychometrika, 66*, 331-342.

Tatsuoka, K.K. (1984). Caution indices based on item response theory. *Psychometrika, 49*, 95-110.

Tatsuoka, K.K. (1996). Use of generalized person-fit indexes, zetas for statistical pattern classification. *Applied Measurement in Education, 9*, 65-75.

Tatsuoka, K.K. y Linn, R.L. (1983). Indices for detecting unusual patterns: Links between two general approaches and potential applications. *Applied Psychological Measurement, 7*, 81-96.

Tatsuoka, K.K. y Tatsuoka, M.M. (1982). Detection of aberrant response patterns and their effect on dimensionality. *Journal of Educational Statistics, 7*, 215-231.

Tatsuoka, K.K. y Tatsuoka, M.M. (1983). Spotting erroneous rules of operation by the individual consistency index. *Journal of Educational Measurement, 7*, 215-231.

Trabin, T.E. y Weiss, D.J. (1979). *The person response curve: Fit of individuals to item characteristic curve models* (Research Report No. 79-7). Minneapolis,

MN: University of Minnesota, Department of Psychology, Psychometric Methods Program.

Trabin, T.E. y Weiss, D.J. (1983). The person response curve: Fit of individuals to item response theory models. En D.J. Weiss (Ed.), *New horizons in testing: Latent trait test theory and computerized adaptive testing* (pp. 83-108). New York: Academic Press.

van der Flier, H. (1977). Environmental factors and deviant response patterns. En Y.H. Poortinga (Ed.), *Basic problems in Cross-Cultural Psychology*. Amsterdam: Swets and Zeitlinger.

van der Flier, H. (1980). *Vergelijkbaarheid van individuele testprestaties* [Comparability of individual test performance]. Lisse, The Netherlands: Swets and Zeitlinger.

van der Flier, H. (1982). Deviant response patterns and comparability of test scores. *Journal of Cross-Cultural Psychology, 13*, 267-298.

van Krimpen-Stoop, E.M.L.A. y Meijer, R.R. (1999). The null distribution of person-fit statistics for conventional and adaptive tests. *Applied Psychological Measurement, 23*, 327-345.

van Krimpen-Stoop, E.M.L.A. y Meijer, R.R. (2000). Detecting person-misfit in adaptive testing using statistical process control techniques. En W.J. van der Linden y C.A.W. Glas (Eds.), *Computerized adaptive testing: Theory and practice* (pp. 201-219). Boston: Kluwer-Nijhoff Publishing.

van Krimpen-Stoop, E.M.L.A. y Meijer, R.R. (2001). CUSUM-based person-fit statistics for adaptive tests with polytomous items. *Journal of Educational and Behavioral Statistics, 26*, 199-218.

van Krimpen-Stoop, E.M.L.A. y Meijer, R.R. (2002). Detection of person misfit in computerized adaptive testing. *Applied Psychological Measurement, 26*, 164-180.

Wollack, J.A. (1997). A nominal response model approach for detecting answer copying. *Applied Psychological Measurement, 21*, 307-320.

Wollack, J.A., Cohen, A.S. y Serlin, R.C. (2001). Defining error rates and power for detecting answer copying. *Applied Psychological Measurement, 25*, 385-404.

Wright, B.D. y Masters, G.N. (1982). *Rating scale analysis*. Chicago: MESA Press.

Wright, B.D. y Stone, M.H. (1979). *Best test design. Rasch measurement*. Chicago: MESA Press.

Zickar, M.J. y Drasgow, F. (1996). Detecting faking on a personality instrument using appropriateness measurement. *Applied Psychological Measurement, 20*, 71-87.

(Artículo recibido: 12-4-05; aceptado: 24-4-06)