

# Características de forma y contenido que predicen la capacidad discriminativa en ítems de personalidad: un análisis basado en la Teoría de Respuesta a los Ítems

Pere Joan Ferrando Piera y Josep Demestre Viladevall  
Universidad Rovira i Virgili

En dos conjuntos de ítems de personalidad que miden extraversión y neuroticismo se estudian las relaciones entre tres grupos de características itemétricas y los índices de discriminación estimados de acuerdo con el modelo TRI de dos parámetros. Las características itemétricas, que se conceptualizan como dimensiones, son: complejidad verbal, ambigüedad e indicatividad, y las hipótesis con respecto a su impacto sobre la capacidad discriminativa se derivan desde un modelo de estadios de procesamiento. La dimensión de indicatividad es la que muestra relaciones más claras con la capacidad discriminativa en la dirección esperada. Se discuten las implicaciones de los resultados cara a la elaboración de medidas en personalidad.

*Content and form characteristics that predict discriminating power in personality items: An Item Response Theory-based analysis.* In two personality item pools that measure extraversion and neuroticism, the relations between three groups of itemmetric properties and the estimated discrimination indexes according to the two-parameter model are assessed. The itemmetric properties, which are conceptualized as dimensions, are: verbal complexity, ambiguity, and trait indicativity, and the hypotheses concerning their impact on the discriminating power are derived from a processing-stage model. The dimension of trait indicativity shows the clearest relations with the discriminating power in the expected direction. The implications of the results in the construction of personality measures are discussed.

En análisis de ítems, el índice de discriminación indica conceptualmente la calidad del ítem como medida (Lord y Novick, 1968). Comparados con los ítems de rendimiento máximo (capacidad, aptitudes, etc.), los ítems de personalidad tienen generalmente una capacidad discriminativa bastante menor, y la necesidad de mejorarla es un objetivo admitido entre los investigadores en este dominio (Burisch, 1984; Meehl, 1945; Loevinger, 1957; Taylor, 1977). Para conseguir esta mejora, sin embargo, parece necesario ir más allá del análisis puramente estadístico, y evaluar cuáles son algunas de las características de un ítem de personalidad que determinan (o simplemente predicen) su capacidad discriminativa.

En personalidad existe una línea de investigación denominada 'itemétrica' (Wiggins y Goldberg, 1965) en la que algunas características de forma y contenido del ítem (itemétricas) son relacionadas con diversos índices psicométricos. Las investigaciones específicas más relevantes cara a la presente investigación se revisan más adelante. En líneas generales, sin embargo, y de acuerdo con Johnson (2004), la investigación itemétrica típica suele adolecer

de una serie de problemas comunes, principalmente: (a) no se basa en modelos psicométricos falseables; (b) utiliza una amalgama de ítems que miden distintas dimensiones, por lo que las estimaciones de discriminación pueden estar influidas por múltiples determinantes; (c) utiliza muestras pequeñas que pueden dar lugar a resultados inestables; y (d) es totalmente exploratoria y no presenta una justificación explícita para la selección de las características a estudiar.

El presente artículo describe dos estudios en los que ciertas características itemétricas se relacionan con los índices de discriminación. En relación con la discusión previa, nuestra investigación pretende superar algunas de las limitaciones citadas. Así, en primer lugar, se basa en un modelo falseable basado en la teoría de respuesta a los ítems (TRI). Respecto al segundo punto, utiliza escalas teóricamente unidimensionales, pudiendo también verificarse el supuesto de unidimensionalidad. Con respecto al punto (c), trabaja con muestras relativamente grandes. Por último, la investigación se basa en un marco teórico que es el que guía la selección de características, agrupadas 'a priori' en una serie de dimensiones.

Como marco teórico de base se utiliza un modelo simple y plausible, basado en 4 estadios de procesamiento y propuesto por Rogers (1974). Este modelo se utiliza tan sólo como un marco de referencia para entender el proceso de responder a un ítem de personalidad y, en principio, no es falseable. Los estadios son: (a) comprensión del estímulo; (b) recuperación en memoria de la información elicitada por el estímulo; (c) integración de la informa-

ción; y (d) toma de decisión y selección de respuesta. A lo largo de esta secuencia el respondiente: (a) interpreta y escala el contenido del ítem en algún continuo psicológico latente; (b) sitúa su propia posición percibida en el mismo continuo; (c) compara ambas posiciones; y (d) emite un juicio y transforma el resultado de la comparación en la escala de respuesta al ítem. Supondremos, además, que la comparación (c) tiene lugar directamente en el continuo del rasgo que se pretende medir, y que la posición del ítem funciona como un umbral en el mecanismo de selección de respuesta. Con estos supuestos y algunas restricciones estadísticas adicionales (Torgerson, 1958) resulta que las respuestas al ítem se comportan de acuerdo con el modelo de ojiva normal de dos parámetros (MON2P) de la TRI, que será el que utilizaremos.

De este planteamiento del MON2P se deduce que la capacidad discriminativa de un ítem está inversamente relacionada con la magnitud de la dispersión discriminativa o variación en su posición percibida (Torgerson, 1958). Conceptualmente esto significa que los ítems con baja capacidad discriminativa son aquellos cuya posición no está claramente definida y que, por tanto, constituyen referencias imprecisas para comparar con la posición que el respondiente percibe de sí mismo (Johnson, 2004). Cara a la presente investigación, se espera que aquellas características de contenido y forma que puedan afectar a la claridad o precisión con la que el ítem se sitúa en el continuo del rasgo tendrán influencia sobre su capacidad discriminativa.

En la presente investigación se han considerado tres grandes grupos de características que se han conceptualizado como dimensiones: (a) Características de superficie; (b) ambigüedad-vaguedad; y (c) indicatividad del rasgo. Desde la teoría, todas ellas pueden relacionarse directamente con la claridad/precisión con la que se determina la posición del ítem. La elección de estas características viene en parte determinada por los resultados de estudios previos que se discuten a continuación.

#### *Características de superficie*

Por características de superficie (Helfrich, 1986) se entienden las características lingüísticas formales (como oposición a las de contenido) del enunciado del ítem. Las características consideradas de mayor interés en este grupo son las de longitud del ítem y las de complejidad verbal (Ferrando y Demestre, 2007; Helfrich, 1986).

Se espera que las características de superficie afecten al primer estadio del modelo. Una mayor longitud y/o complejidad dificultarían la comprensión, provocando una mayor vaguedad en la determinación de la posición del ítem. Por tanto, se esperan encontrar relaciones negativas entre longitud/complejidad y capacidad discriminativa. En general, éstos son los resultados obtenidos en estudios previos (Angleitner, John y Lörh, 1986; Zickar y Ury, 2002). Cabe notar, sin embargo, que algunas de las medidas de complejidad verbal utilizadas aquí (Imaginabilidad, y Legibilidad) no se han empleado en estudios previos.

#### *Ambigüedad-vaguedad*

El tipo de ambigüedad considerado es la ambigüedad psicológica, un concepto diferente al de ambigüedad lingüística, y se entendería como la incertidumbre acerca del significado denotativo de un ítem. Esta incertidumbre puede deberse a distintas causas, tales como falta de precisión en el enunciado (vaguedad) o múltiples interpretaciones posibles del mismo.

De acuerdo con nuestro modelo, la ambigüedad afectaría al estadio de interpretación del estímulo. A mayor ambigüedad, mayor dispersión discriminativa se espera que tenga la posición del ítem y, por tanto, menor capacidad discriminativa. La literatura, sin embargo, indica que esta relación es muy controvertida. Un planteamiento alternativo considera que los ítems de personalidad funcionan como pequeños tests proyectivos que la persona interpreta a su manera. Desde esta postura se ha planteado que la ambigüedad es posiblemente irrelevante para predecir la capacidad discriminativa (Meehl, 1945). Otros autores van más allá y consideran que la forma en que una persona interpreta un ítem es un buen indicador de personalidad en sí mismo, de manera que los ítems ambiguos pueden ser a veces más discriminativos que los claros (Johnson, 2004).

Los resultados empíricos existentes son dispares. Algunos estudios han encontrado relaciones positivas y significativas entre ambigüedad y capacidad discriminativa que apoyarían la interpretación 'proyectiva' (Gordon, 1953; Isard, 1956; Johnson, 2004). Otros han encontrado relaciones negativas y significativas, que apoyarían la presente hipótesis (Angleitner, John y Lörh, 1986; Strong, 1962). Finalmente, en otros estudios no se han encontrado relaciones significativas (Goldberg, 1968; Taylor, 1977; Zickar y Ury, 2002).

#### *Indicatividad*

En esta dimensión se pretende evaluar el grado de relación existente entre la conducta que describe el ítem y la dimensión que pretende medir, así como la centralidad de dicha conducta con respecto al rasgo (hasta qué punto el ítem describe aspectos esenciales del rasgo). Para entender esta dimensión puede ser útil considerar la distinción que hace Burisch (1984) entre ítems «definitorios» o «prototípicos» e ítems «correlacionales». Para cualquier rasgo de personalidad habría una serie de conductas básicas que definen el rasgo (las que miden los ítems definitorios) y una serie de conductas asociadas que tienden a ir relacionadas con aquellas (y que miden los ítems correlacionales). La relación ítem-rasgo es, generalmente, mucho más clara y directa en los ítems prototípicos que en los correlacionales.

De acuerdo con nuestro modelo, cuanto más clara y central sea la relación ítem-rasgo, más precisa se espera que sea la localización del ítem en el continuo del rasgo, y, por tanto, mayor su capacidad discriminativa. Sin embargo, al igual que pasaba con la ambigüedad, aquí existe también controversia teórica. Algunos autores consideran que las relaciones ítem-rasgo muy claras puede elicitar sesgos de respuesta o tendencia al falseamiento y, posiblemente, disminuir la calidad del ítem como medida (Zickar y Ury, 2002). Otros autores (Burisch, 1984; Johnson, 2004; Loevinger, 1957) admiten este riesgo, pero consideran que los ítems indicativos son preferibles y que 'a priori' implican un mayor cociente señal/ruido.

Los resultados existentes no dan apoyo claro a ninguna de las posturas. En una revisión Adams y Butler (1967) concluyeron que el contenido del ítem era esencialmente irrelevante para predecir su capacidad discriminativa. En la misma línea, la revisión de Burisch (1984) concluía que, en la práctica, no parecía haber diferencias en la capacidad discriminativa entre ítems prototípicos e ítems correlacionales. Por otra parte, en algunos estudios se obtuvieron relaciones positivas y significativas entre indicatividad y discriminación (Johnson, 2004; Taylor, 1977; Zickar y Ury, 2003), tal como se espera encontrar aquí.

ESTUDIOS EMPÍRICOS

Se presentan dos estudios empíricos, cada uno de los cuales se basa en un test y una muestra distintos a los del otro. Sin embargo, el método de análisis es común a ambos y, además, algunas de las medidas requirieron el uso conjunto de los dos tests. Para evitar repeticiones innecesarias, los dos estudios se presentarán simultáneamente.

Método

Instrumentos

En el primer estudio se utilizó una escala de Extraversión (E) formada por 100 ítems en formato de respuesta binario. Los ítems se obtuvieron de los siguientes cuestionarios o bancos de ítems: EPQ-R (Eysenck, Eysenck y Barrett, 1985), EPI (Eysenck y Eysenck, 1963) e IPIP (Goldberg, 1999). En el segundo estudio se utilizó una escala de neuroticismo (N) compuesta por 60 ítems también binarios. En este caso todos los ítems procedían de los diferentes cuestionarios de Eysenck. Las características ítemétricas (las mismas en ambos estudios) se evaluaron mediante tres tipos de medidas: objetivas, directas e indirectas.

Las medidas de superficie eran todas ellas objetivas. Se utilizaron las siguientes: número de sílabas; número de letras y espacios en blanco; frecuencia de uso; imaginabilidad y legibilidad. Las cuatro primeras se consideran las medidas de superficie más relevantes de acuerdo con los estudios previos (Angleitner, John y Lörh, 1986; Zickar y Ury, 2002). Helfrich (1986) considera a las otras dos características como potencialmente relevantes, pero no parecen haber sido utilizadas previamente quizá por falta de medidas adecuadas.

La frecuencia de uso de las palabras de un ítem se calculó mediante el programa LEXESP (Sebastián-Gallés et al., 2000), que proporciona el número de veces que aparece una palabra particular en un texto de un millón de palabras. La medida utilizada para cada ítem fue el valor de frecuencia de la palabra menos frecuente. La imaginabilidad se calculó mediante las normas de Valle (1998) utilizando para cada palabra una escala de 1 a 7, siendo 7 la puntuación que indica el máximo de imaginabilidad (es muy fácil imaginar el significado de la palabra). La medida para cada ítem fue la mediana de los valores de las distintas palabras. Finalmente, la medida de legibilidad (o facilidad de lectura) fue la adaptación española del índice de Flesch (1948). Cuanto mayor es el valor del índice, mayor es la facilidad con la que se lee el enunciado.

Todas las medidas de ambigüedad eran directas. Se utilizaron cuatro escalas de clasificación en formato de tres puntos que fueron evaluadas por expertos. Estas medidas se presentan en la tabla 1.

Para medir la indicatividad se utilizaron tres medidas directas y una medida indirecta. Las directas eran escalas de clasificación como las descritas arriba y se presentan también en la tabla 1 (IND1-IND3). Estas medidas se adaptaron de las utilizadas anteriormente por Johnson (2004) y Zickar y Ury (2002). La medida indirecta (IND-IDENT) se derivó de un procedimiento utilizado por Holden y Jackson (1979). Se facilitaba a los 5 evaluadores (véase abajo) un listado con los 160 ítems de E y N presentados al azar. Para cada ítem el evaluador debía indicar si creía que medía E, N, o ninguno de los dos. La puntuación de IND-IDENT (de 0 a 5) era el número de evaluaciones que coincidían con la dimensión que dicho ítem pretendía medir.

Tabla 1  
Medidas ítemétricas directas

Ambigüedad-vaguedad
<p>AMB1: El enunciado de este ítem es: 1) Claro; 2) Término medio (TM); 3) Ambiguo</p> <p>AMB2: El enunciado de este ítem es: 1) Se entiende a la primera lectura; 2) TM; 3) No se entiende a la primera lectura</p> <p>AMB3: El enunciado de este ítem es: 1) Preciso; 2) TM; 3) Vago</p> <p>AMB4: El enunciado de este ítem es: 1) Tiene una única interpretación; 2) TM; 3) Puede interpretarse de diferentes maneras</p>
Indicatividad
<p>IND1: ¿Hasta qué punto es fácil-difícil ver la conexión entre la respuesta a este ítem y el rasgo que pretende medir? 1) Fácil; 2) TM; 3) Difícil</p> <p>IND2: ¿Hasta qué punto la conducta o sentimiento que describe este ítem es típica o representativa del rasgo que pretende medir? 1) Representativa; 2) TM; 3) Poco representativa</p> <p>IND3: De acuerdo con la definición del rasgo, ¿hasta qué punto el ítem describe aspectos primarios o centrales del rasgo que pretende medir? 1) Mucho; 2) TM; 3) Poco</p>

Participantes

Los participantes en el primer estudio (E) fueron 531 estudiantes universitarios de ciencias sociales de la Universidad Rovira i Virgili (Tarragona). Los datos se recogieron en el curso 2004-2005. Los participantes en el segundo estudio fueron 436 estudiantes con la misma procedencia y, en este caso, los datos se recogieron durante el curso 2003-2004. En ambos estudios la media de edad era de unos 21 años y el 80% eran mujeres.

Procedimientos

En ambos estudios los cuestionarios se administraron en grupos de clase de unos 60 alumnos.

Todas las escalas de clasificación (tabla 1) fueron respondidas por los mismos tres evaluadores que eran profesores de Psicología general en nuestra Universidad. En la medida indirecta de indicatividad se incluyeron dos evaluadores adicionales que eran becarios predoctorales. En todas las medidas de indicatividad los evaluadores debían establecer la relación entre la respuesta al ítem y la dimensión que este ítem pretende medir. Para que la evaluación fuese lo más objetiva posible se facilitó a los evaluadores la descripción de los rasgos E y N que se incluye en los cuadernillos del EPI y del EPQ. Se les instruyó para que respondiesen de acuerdo con esta descripción.

Diseño y análisis

Ambos estudios utilizan un diseño correlacional común estructurado en tres fases de análisis. En la primera fase se obtienen las medidas y se evalúan sus propiedades básicas. Por una parte, se calibran los ítems de acuerdo con el MON2P y se evalúa la adecuación del

modelo. Por otra parte, se obtienen las medidas ítemétricas y se evalúa la fiabilidad de las puntuaciones en las escalas de clasificación.

En la segunda fase las medidas no objetivas se analizan factorialmente. De las 8 medidas no objetivas, cuatro pretenden medir una dimensión general de ambigüedad y las cuatro restantes una dimensión general de indicatividad. Estas medidas se tratan como si fuesen ítems de un instrumento psicométrico bidimensional. Si la estructura es plausible, entonces pueden obtenerse puntuaciones totales en las dos «escalas generales» de ambigüedad e indicatividad, que pueden después utilizarse en la tercera fase del análisis para obtener relaciones más simples y fiables con los índices de discriminación. Sin embargo, para que esto sea correcto, los datos deben apoyar esta supuesta estructura bidimensional. De aquí el análisis factorial.

En la tercera fase se evalúan las relaciones entre las características ítemétricas y los índices de discriminación estimados. En principio se obtienen las correlaciones bivariadas correspondientes a cada característica por separado. Si el análisis factorial apoya la estructura propuesta, entonces se obtendrán también las correlaciones correspondientes a las puntuaciones totales en ambigüedad y en indicatividad. En conjunto, esta tercera fase correspondería al diseño típico utilizado en investigación ítemétrica.

## Resultados

### Primera fase

Ambos conjuntos de ítems fueron calibrados separadamente de acuerdo con el MON2P mediante el programa NOHARM (Fraser y McDonald, 1988). En ambos estudios el ajuste del modelo fue aceptable. Para los ítems E, la media cuadrática residual fue 0.014 y el índice gamma-GFI 0.83. Para los ítems N los valores correspondientes fueron 0.013 y 0.81. En el conjunto E, el recorrido de los índices de discriminación era de 0.10 a 1.90, con una media de 0.72 y una varianza de 0.13. En el conjunto N era de 0.05 a 1.20, con media 0.66 y varianza 0.06. En la métrica utilizada, los valores medios obtenidos son los habituales en medidas de personalidad.

Las fiabilidades correspondientes a cada una de las escalas de clasificación se calcularon mediante la correlación media entre los tres evaluadores. Para las medidas de ambigüedad, la fiabilidad de cada una de las escalas era aproximadamente de 0.30. Para las medidas de indicatividad era aproximadamente de 0.60. La baja fiabilidad de las evaluaciones de ambigüedad es un resultado típico (Taylor, 1977). Los resultados justifican el enfoque de tratar a cada escala como un ítem psicométrico con un considerable error de medida. Si las medidas pudiesen combinarse en escalas totales, las puntuaciones totales tendrían fiabilidades estimadas de 0.63 (ambigüedad) y 0.86 (indicatividad).

### Segunda fase

En ambos estudios las observaciones son los ítems y, por tanto, se analizan factorialmente conjuntos de 8 variables medidas en grupos de 100 y 60 observaciones, respectivamente. El pequeño número de observaciones apunta la necesidad de utilizar un procedimiento factorial lo más simple y robusto posible, por lo que se eligió el procedimiento MINRES. Se utilizó el programa FACTOR (Lorenzo y Ferrando, 2006) y se especificó una solución inicial no restringida en dos factores que fue después transformada mediante rotación oblimin directa. En ambos estudios el modelo bifactorial proporcionó un excelente ajuste. La raíz media cuadrática residual y el índice

gamma-GFI fueron de 0.03 y 0.99 (E ítems) y 0.04 y 0.99 (N ítems). Los patrones rotados se presentan en la tabla 2. Las correlaciones estimadas entre factores fueron 0.06 (E ítems) y 0.25 (N ítems). Además de una estructura muy clara, del tipo de conglomerados independientes, la solución factorial sugiere que las dos dimensiones son esencialmente independientes. Las soluciones obtenidas justifican claramente el tratamiento de las escalas como ítems y la obtención de puntuaciones totales de ambigüedad y de indicatividad.

### Tercera fase

Para las medidas ítemétricas compuestas (las puntuaciones totales en ambigüedad y en indicatividad) se transformaron las correspondientes medidas simples a la misma escala y se obtuvo la puntuación total mediante suma simple. El pequeño número de observaciones no justifica el uso de combinaciones más complejas.

Se examinaron en primer lugar todos los gráficos bivariados a fin de detectar posibles relaciones no-lineales o bien valores extremos que pudiesen distorsionar las correlaciones estimadas. No se detectaron problemas de este tipo y, por tanto, se obtuvieron las estimaciones puntuales y los intervalos de confianza al nivel del 95%. Los intervalos de confianza se obtuvieron mediante autodocimasia (Bootstrap) utilizando 100 replicaciones en cada caso. Los resultados se presentan en la tabla 3. Los intervalos de confianza se presentan entre paréntesis tras el estimador puntual correspondiente.

Tabla 2  
Solución bifactorial para las medidas de ambigüedad e indicatividad

Medida	Ítems E		Ítems N	
AMB1	<b>0.91</b>	-0.03	<b>0.84</b>	0.16
AMB2	<b>0.81</b>	-0.05	<b>0.69</b>	-0.05
AMB3	<b>0.83</b>	0.05	<b>0.79</b>	-0.13
AMB4	<b>0.72</b>	0.07	<b>0.70</b>	0.02
IND1	0.01	<b>0.97</b>	0.09	<b>0.90</b>
IND2	0.02	<b>0.97</b>	-0.03	<b>0.95</b>
IND3	-0.01	<b>0.95</b>	0.03	<b>0.95</b>
INDIDENT	-0.03	<b>0.69</b>	-0.23	<b>0.59</b>

Tabla 3  
Correlaciones entre las medidas ítemétricas y los índices de discriminación

Medida	Ítems E	Ítems N
AMB1	0.12 (-0.08; 0.35)	-0.30 (-0.53; -0.12)*
AMB2	0.02 (-0.19; 0.23)	-0.16 (-0.38; 0.05)
AMB3	0.18 (-0.04; 0.31)	-0.06 (-0.28; 0.24)
AMB4	0.14 (-0.04; 0.32)	-0.16 (-0.44; 0.15)
Total AMB	-	-0.20 (-0.50; 0.08)
IND1	0.37 (0.25; 0.49)*	0.60 (0.41; 0.65)*
IND2	0.43 (0.32; 0.57)*	0.70 (0.43; 0.76)*
IND3	0.41 (0.29; 0.52)*	0.72 (0.47; 0.74)*
INDIDENT	0.28 (0.16; 0.42)*	0.22 (0.02; 0.42)*
Total IND	0.41 (0.29; 0.52)*	0.59 (0.40; 0.65)*
Nº sílabas	-0.25 (-0.30; -0.05)*	-0.03 (-0.23; 0.16)
Nº letras	-0.25 (-0.39; -0.07)*	-0.04 (-0.26; 0.16)
Frecuencia uso	0.11 (-0.12; 0.21)	0.18 (0.00; 0.38)*
Imaginabilidad	0.04 (-0.13; 0.09)	-0.17 (-0.37; 0.00)*
Legibilidad	-0.01 (-0.19; 0.17)	0.19 (-0.06; 0.40)

Los resultados más claros en la tabla 3 corresponden a las relaciones entre las medidas de indicatividad y los índices de discriminación. En ambos estudios las relaciones van en la dirección prevista: cuanto más clara es la relación entre el ítem y el rasgo que pretende medir, y cuanto más relevante es la conducta descrita con respecto a dicho rasgo, mayores tienden a ser los índices de discriminación. Las relaciones obtenidas apoyan la distinción teórica de Burisch (1984). En un conjunto de ítems amplio y bien calibrado parece que los ítems prototípicos tienden a mostrar mayor capacidad discriminativa que los ítems correlacionales. Por otra parte, las medidas basadas en evaluaciones de jueces permiten distinguir bastante bien entre ambos tipos de ítems y, sobre todo, predecir con relativa precisión su capacidad discriminativa.

Los resultados con respecto a la ambigüedad son bastante débiles. En el caso de los ítems E ninguna de las medidas individuales muestra relaciones significativas con la capacidad discriminativa. En el caso de los ítems N las relaciones van en la dirección prevista, aunque la intensidad de la relación es bastante baja. Algo parecido sucede con las medidas de superficie. En ambas escalas la longitud se relaciona negativamente con la capacidad discriminativa, pero sólo pueden considerarse significativas las correlaciones obtenidas con los ítems E. Las relaciones con frecuencia e imaginabilidad también van en la dirección prevista: cuanto más frecuentes y fáciles de imaginar las palabras del enunciado mayor tiende a ser el índice de discriminación. Sin embargo, en este caso tan sólo pueden considerarse significativas las relaciones obtenidas en N. El índice de Flesch no muestra relaciones significativas en ninguna de las dos escalas. Globalmente, y desde un punto de vista metodológico, la falta de resultados más importantes aconseja no ir más allá la tabla de correlaciones.

### Discusión

Aceptando provisionalmente los resultados significativos aunque no se repliquen en ambos estudios (los concernientes a algunas características de superficie y, sobre todo, a indicatividad), se podría llegar a la siguiente recomendación. Para aumentar la capacidad discriminativa de los ítems debe procurarse que los enunciados sean cortos y directos, y que el contenido se dirija a los aspectos centrales o definitorios del rasgo. Estas recomendaciones podrían parecer bastante obvias para un psicómetra no especiali-

zado en personalidad. Sin embargo, como se ha discutido antes, no lo son tanto, y existen aún posiciones teóricas enfrentadas.

La presente investigación intenta superar limitaciones importantes de la investigación ítemétrica previa, pero también tiene, desde luego, sus propias e importantes limitaciones. Una de ellas sería un problema potencial de doble atenuación por restricción de rango. En primer lugar, en ambos estudios los participantes son estudiantes universitarios, con un nivel cultural presumiblemente alto. Bock, Dicken y van Pelt (1969) conjeturaron que los universitarios tienden a esforzarse más y a responder de manera más meticulosa al ítem que la población general. Es posible que las relaciones negativas entre las medidas de superficie y la capacidad discriminativa hubiesen sido más importantes en grupos de participantes con menor riqueza de vocabulario, menor hábito de lectura, o menor interés. En segundo lugar, en ambos estudios se han utilizado ítems pertenecientes a cuestionarios que se consideran buenas medidas del rasgo. Estos ítems han sido obtenidos tras un largo proceso de refinamiento en el que, posiblemente, se fueron eliminando los ítems menos claros. Quizá las relaciones ambigüedad-capacidad discriminativa serían más sustanciales en las etapas iniciales, cuando se trabaja con el banco de ítems inicial. En todo caso estas interpretaciones son sólo conjeturas y el tema requiere investigación.

Desde un punto de vista teórico, algunos de los presentes resultados pueden ayudar a entender qué es lo que realmente indican los índices de discriminación en ítems de personalidad y qué hace que los ítems sean más o menos precisos como medidas del rasgo. Desde un punto de vista práctico, la utilización de un sistema de jueces en las primeras etapas de desarrollo de un test podría permitir al constructor desarrollar conjuntos de ítems de una manera más racional, o bien obtener conjuntos con propiedades psicométricas similares. Se podría ir aún más allá y empezar a plantear un tipo de investigación como la llevada a cabo por Embretson y colaboradores (e.g., Embretson y Wetzel, 1987) en el dominio cognitivo. Así, se podría plantear un modelo de TRI que incorporase las evaluaciones de indicatividad como información auxiliar para modelar el índice de discriminación. Los resultados del presente análisis factorial sugieren además la posibilidad de emplear una única medida de indicatividad obtenida como combinación lineal. Un modelo de este tipo podría utilizarse en los estadios iniciales de selección de ítems, y también en la calibración, para conseguir estimaciones más precisas del índice de discriminación.

### Referencias

- Adams, H.E., y Butler, J.R. (1967). The deviation hypothesis: A review of the evidence. En I.A. Berg (ed.): *Response sets in personality assessment* (pp. 191-208). Chicago: Aldine.
- Angleitner, A., John, O.P., y Löhr, F.J. (1986). It's what you ask and how you ask it: An itemmetric analysis of personality questionnaires. En A. Angleitner y J.S. Wiggins (eds.): *Personality assessment via questionnaires* (pp. 61-107). Berlin: Springer-Verlag.
- Bock, R.D., Dicken, C., y van Pelt, J. (1969). Methodological implications of content acquiescence correlation in the MMPI. *Psychological Bulletin*, 71, 127-139.
- Burisch, M. (1984). Approaches to personality inventory construction. *American Psychologist*, 39, 214-227.
- Embretson, S.E., y Wetzel, C.D. (1987). Component latent-trait models for paragraph comprehension tests. *Applied Psychological Measurement*, 11, 175-193.
- Eysenck, H.J., y Eysenck, S.B.G. (1963). *Manual of the Eysenck personality inventory*. London: University of London press.
- Eysenck, S.B.G., Eysenck, H.J., y Barrett, P.T. (1985). A revised version of the Psychoticism scale. *Personality and Individual Differences*, 6, 21-29.
- Ferrando, P.J., y Demestre, J. (2007). Características del ítem que predicen el tiempo de respuesta en medidas de personalidad. *Psicothema*, 19, 337-342.
- Flesch, R. (1948). A new readability yardstick. *Journal of Applied Psychology*, 32, 221-233.
- Fraser, C., y McDonald, R.P. (1988). NOHARM: Least squares item factor analysis. *Multivariate Behavioral Research*, 23, 267-269.
- Goldberg, L.R. (1968). The interrelationships among item characteristics in an adjective checklist: The convergence of different indices of item ambiguity. *Educational and Psychological Measurement*, 28, 273-296.

- Goldberg, L.R. (1999). A broad-bandwidth, public domain, personality inventory measuring the lower-level facets of several five-factor models. En I. Mervielde, I. Deary, F. De Fruyt y F. Ostendorf (eds.): *Personality Psychology in Europe, vol. 7* (pp. 7-28). Tilburg: Tilburg University Press.
- Gordon, L.V. (1953). Some interrelationships among personality item characteristics. *Educational and Psychological Measurement, 13*, 264-272.
- Helfrich, H. (1986). On linguistic variables influencing the understanding of questionnaire items. En A. Angleitner y J.S. Wiggins (eds.): *Personality assessment via questionnaires* (pp. 61-107). Berlín: Springer.
- Holden, R.R., y Jackson, D.N. (1979). Item subtlety and face validity in personality assessment. *Journal of Consulting and Clinical Psychology, 47*, 459-468.
- Isard, E.S. (1956). The relationship between item ambiguity and discriminating power in a forced-choice scale. *Journal of Applied Psychology, 40*, 266-268.
- Johnson, J.A. (2004). The impact of item characteristics on item and scale validity. *Multivariate Behavioral Research, 39*, 273-302.
- Loevinger, J. (1957). Objective tests as instruments of psychological theory. *Psychological Reports, 3*, 635-694.
- Lord, F.M., y Novick, M.R. (1968). *Statistical theories of mental test scores*. Reading: AddisonWesley.
- Lorenzo, U., y Ferrando, P.J. (2006). FACTOR: A computer program to fit the exploratory factor analysis model. *Behavioral Research Methods, Instruments & Computers, 38*, 88-91.
- Meehl, P.A. (1945). The dynamics of «structured» personality tests. *Journal of Clinical Psychology, 56*, 367-374.
- Rogers, T.B. (1974). An analysis of the stages underlying the process of responding to personality items. *Acta Psychologica, 38*, 205-213.
- Sebastián-Gallés, N., Martí, M.A., Carreiras, M., y Cuetos, F. (2000). *LEX-ESP: léxico informatizado del español*. Barcelona: Edicions Universitat de Barcelona.
- Strong, E.K. (1962). Good and poor interest items. *Journal of Applied Psychology, 40*, 266-268.
- Taylor, J.B. (1977). Item homogeneity, scale reliability and the self-concept hypothesis. *Educational and Psychological Measurement, 37*, 349-361.
- Torgerson, W. (1958). *Theory and methods of scaling*. New York: Wiley.
- Valle, F. (1998). *Normas de imaginabilidad*. Oviedo: Servicio de Publicaciones de la Universidad de Oviedo.
- Wiggins, J.S., y Goldberg, L.R. (1965). Interrelationships among MMPI item characteristics. *Educational and Psychological Measurement, 25*, 381-397.
- Zickar, M.J., y Ury, K.L. (2002). Developing an interpretation of item parameters for personality items: Content correlates of parameter estimates. *Educational and Psychological Measurement, 62*, 19-31.