

Características del ítem que predicen el tiempo de respuesta en medidas de personalidad

Pere J. Ferrando y Josep Demestre
Universidad Rovira i Virgili

Se evalúan algunas características de los ítems de personalidad que se espera afecten al tiempo de respuesta obtenido en una administración informatizada. En primer lugar, se discuten los resultados de estudios previos y se propone un marco teórico general desde el que establecer una serie de predicciones. En segundo lugar, estas predicciones se contrastan empíricamente utilizando dos conjuntos de datos, uno de ellos basado en ítems binarios y el otro en ítems de respuesta graduada. Algunos de los resultados obtenidos son nuevos, y el resto tiende a concordar con los resultados obtenidos en estudios previos. Se discuten las implicaciones de los resultados para la utilización de las latencias en estudios de tipo aplicado.

Item characteristics that predict response latency in personality measures. This study assesses some item characteristics that are expected to affect response latencies in the computerized administration of a personality questionnaire. First, some previous empirical results are discussed, and a theoretical framework from which predictions can be made is proposed. Second, the predictions are empirically assessed using two data-sets, one of them based on binary items, and the other one based on graded-response items. Some of the results obtained are new, whereas the remaining agree with previous empirical results. The implications of the present results in applied research, and particularly for the use of latencies, is discussed.

Hambleton (2004) predijo que el mayor cambio en el uso de los tests en las próximas décadas sería la sustitución del formato en papel y lápiz por la administración informatizada. Esta predicción iba referida, sobre todo, a tests educativos y de capacidad. Sin embargo, también la administración informatizada de cuestionarios de personalidad ha aumentado notablemente en los últimos años. Entre otras ventajas, dicha administración permite registrar con precisión el tiempo que media entre la presentación del ítem y la respuesta (tiempo de respuesta o latencia), lo cual aporta un nuevo tipo de información complementaria. Sin embargo, como señala Stout (1981), no está claro cuál es el tipo de información específica que aportan las latencias y la forma en que dicha información se puede integrar en el análisis de ítems.

La mayor parte de estudios empíricos sobre la latencia en personalidad adoptan un enfoque común. El tiempo de respuesta se considera una variable dependiente que, en parte, viene determinada por ciertas: (a) características del ítem; (b) características del individuo; y (c) efectos interactivos entre el ítem y el individuo. Algunos estudios se han concentrado en las características de los ítems (e.g., Dunn, Lushene, y O'Neil, 1972; Rogers, 1973, 1974; Tetrick, 1989), mientras que otros han enfatizado las diferencias individuales y los efectos de interacción (e.g., Popham y Holden, 1990).

El presente estudio pretende ser una extensión de los estudios previos centrados en las características de los ítems. Pretendemos ir más allá de listar un catálogo de características y evaluarlas de forma exploratoria, por lo que el estudio se estructura en dos partes. En primer lugar, se delimita un marco teórico desde el que establecer predicciones. En segundo lugar, estas predicciones se contrastan en dos conjuntos de datos reales. En última instancia, la finalidad de nuestro estudio sería la de evaluar si la información que las latencias aportan con respecto al ítem puede ser de utilidad en el análisis de dicho ítem. Más en detalle, si esta información podría llevar a estimaciones más precisas de los índices psicométricos de dicho ítem.

Nuestro estudio extiende la investigación existente en distintos aspectos. Los estudios previos se basaban generalmente en muestras pequeñas que raramente excedían los 100 sujetos y en ítems binarios. Aquí hemos utilizado muestras considerablemente mayores y evaluado tanto ítems binarios como ítems de respuesta graduada. Además hemos considerado características y medidas que no se habían tenido en cuenta anteriormente. Así, se ha estudiado el impacto de las características del ítem sobre la varianza de las latencias.

Predicciones y revisión de estudios previos

Como esquema del proceso de respuesta a un ítem de personalidad adoptaremos el modelo de Rogers (1974) basado en tres estadios seriales de procesamiento: (a) la codificación del estímulo, (b) la comprensión del enunciado, y (c) la toma de decisión y respuesta. El modelo es muy simple, ya que el procesamiento podría tener componentes en paralelo. Sin embargo, creemos que es suficiente para el tipo de predicciones que haremos.

Las características del ítem se clasifican en tres categorías: (a) de superficie, (b) psicométricas, y (c) de orden. «Características de superficie» (Helfrich, 1986) es un término que refiere a características lingüísticas de forma (por ejemplo, longitud o estructura sintáctica) como oposición a las de contenido.

La evidencia empírica sugiere que las características de superficie son las más relacionadas con la latencia, y, entre ellas, la principal es la longitud del ítem (Bartram y Power, 1983; Dunn et al., 1972; Hanley, 1962; Rogers, 1973, 1974; Tetrick, 1989). El primer estadio de codificación incluye la actividad de leer el ítem, y la longitud del ítem es un factor obvio que se espera afecte a dicha actividad (Rogers, 1974). Dunn et al. (1972) y Tetrick (1989) encontraron que la longitud del ítem explicaba entre el 45 y el 65% de la varianza de las latencias. Además de la longitud, se han considerado otras características de superficie tales como: dirección positiva o negativa, tiempo verbal, uso de pasiva, sujeto, etc. En conjunto, éstas no parecen predecir varianza adicional más allá de lo que permite predecir la longitud (Dunn et al., 1972; Tetrick, 1989).

En lugar de explorar un listado de características, aquí hemos adaptado la idea de Helfrich (1986) de que, aparte de la longitud, la complejidad del ítem es la principal característica de superficie, y que dicha complejidad no viene explicada necesariamente por una serie de características consideradas por separado. Helfrich consideraba que la evaluación de la complejidad requería un análisis psicolingüístico. Desde la psicolingüística (e.g., Ferreira, 1991) se ha sugerido que para medir la complejidad de una oración o enunciado lingüístico el mejor índice es el número total de nodos presentes en la representación sintáctica (i. e. arbórea) del enunciado. Desde el modelo de procesamiento adoptado, la longitud y la complejidad afectarían a estadios diferentes: la longitud afectaría al estadio de codificación, la complejidad al de comprensión.

No tenemos noticia de estudios previos que hayan considerado el impacto de la longitud y complejidad del ítem sobre la variabilidad de las latencias. Sin embargo, la administración informatizada de un ítem es, esencialmente, una tarea de elección (Rogers, 1974). Un resultado bastante consistente en este tipo de tareas es que, a medida que aumenta la complejidad, se incrementan tanto los tiempos medios como las varianzas interindividuales (Jensen, 1982). En otras palabras, la mayor complejidad de la tarea permite que aparezcan más claramente las diferencias entre los individuos. Por esta razón, esperamos encontrar que la longitud y complejidad del ítem estarán positivamente relacionadas no sólo con los tiempos medios, sino también con las varianzas.

La característica psicométrica que más se ha relacionado con la latencia es la denominada 'controversialidad' del ítem (Fricke, 1957), y está relacionada con el índice de dificultad o de posición. Se ha encontrado generalmente que los ítems extremos (muy 'fáciles' o muy 'difíciles') muestran tiempos medios de respuesta menores que los ítems con posiciones intermedias (Dunn et al., 1972; Hanley, 1962; Kuncel, 1973; Rogers, 1973). Sin embargo, como indica Kuncel (1973), no se ha propuesto una explicación convincente para este resultado.

Los autores consideramos que el efecto de controversialidad es un resultado marginal que surge a partir de un efecto interactivo denominado 'la hipótesis de cercanía' (Kuncel, 1973, 1977). Si el nivel del individuo en el rasgo y la posición del ítem pueden representarse en un continuo o dimensión común, entonces la dificultad para responder al ítem aumentará a medida que ambas posiciones se aproximen. La lógica es que la posición del ítem funciona como un umbral que marca la transición entre admitir el

ítem y rechazarlo. Cuando el nivel de un sujeto está muy cerca del umbral, este sujeto sentirá que ambas tendencias de respuesta (aceptar o rechazar) pueden ser apropiadas y su dificultad en responder será muy alta. Esta dificultad actúa en el estadio de decisión y se espera se traduzca en una mayor latencia.

La hipótesis de cercanía es relativa, y nada dice respecto a que el ítem sea o no extremo. El efecto marginal puede deducirse si se asume que la distribución de los niveles en el rasgo es unimodal, con la mayor parte de los casos concentrados cerca de la media y con frecuencias que decrecen en las colas. Si es así, un ítem extremo está 'lejos' de la mayor parte de los sujetos y, por tanto, su tiempo medio de respuesta esperado es menor que el de un ítem de dificultad media, que queda 'cerca' de la mayor parte de los individuos.

La literatura acerca de las relaciones entre el índice de discriminación y la latencia es muy escasa. Tetrick (1989) obtuvo una relación negativa, pero no trató de interpretar el resultado. Aquí utilizaremos una interpretación relacionada con la hipótesis de cercanía. Si el umbral del ítem indica el punto de transición entre la tendencia a admitir el ítem y la tendencia a rechazarlo, y si consideramos una región de incertidumbre en torno a este punto, entonces el índice de discriminación determina la 'rapidez' con la que el sujeto se alejará de dicha región. Así, cuando el ítem es muy discriminativo, pequeños cambios en el nivel del rasgo se traducirán en grandes cambios en la respuesta esperada en torno a este punto. El efecto marginal esperado es que los ítems con mayor capacidad discriminativa mostrarán tiempos medios de respuesta menores que los ítems con poca capacidad.

Pasamos, por último, a los efectos de orden. Como en cualquier tarea cognitiva, en la administración informatizada de un test se espera encontrar un efecto de práctica que hace que los tiempos tiendan a disminuir a lo largo de la tarea (Fazio, 1990; Stout, 1991; Tetrick, 1989). Sin embargo, este potencial efecto no parece haberse estudiado en detalle. La idea predominante es la de que el efecto se limita a los dos o tres primeros ítems y después se estabiliza. La hipótesis que adoptaremos aquí es la de una relación no lineal entre el orden de presentación y el tiempo medio, que acaba haciéndose asintótica a partir de un determinado número de presentaciones.

Algunas consideraciones metodológicas

Como variable dependiente, el tiempo de respuesta tiene dos problemas importantes (Fazio, 1990). En primer lugar, es muy 'ruidoso', está sujeto a muchas fuentes de variabilidad aleatorias. En segundo lugar, tiene una distribución con fuerte sesgo positivo por mucho que se controle la situación de recogida de datos (lo que no es el caso en la administración convencional de un test).

La mayor parte del 'ruido' en las latencias parece deberse a las diferencias individuales en factores tales como velocidad de lectura, capacidad de concentración, edad, etc. Una segunda fuente importante de error es la variabilidad momentánea debido a confusión en la respuesta, déficits momentáneos de atención, etc. La magnitud de estos tipos de error puede mitigarse en parte mediante instrucciones apropiadas que enfatizan a la vez la precisión y la velocidad (Fazio, 1990).

En nuestro estudio las unidades básicas de análisis son las medias por ítem obtenidas a través de los individuos. Por tanto, cabe esperar que las diferencias individuales se controlarán sin necesidad de manipulaciones estadísticas adicionales siempre que se utilice una muestra suficientemente grande.

Respecto al segundo punto, la distribución característica de los tiempos de respuesta presenta una cola hacia la derecha considerable, en algunos casos con valores muy extremos (Fazio, 1990; Jensen, 1982; Radcliff, 1993), lo que sugiere la necesidad de transformar los tiempos directos. La práctica habitual consiste en truncar primero los valores más extremos y después transformar los datos mediante la transformación raíz cuadrada, logarítmica o inversa. La literatura (Fazio, 1990; Radcliff, 1993), junto a nuestra experiencia, indica que la transformación inversa ($Y = 1/X$) es la que da mejores resultados.

Estudios empíricos

En esta sección se discuten dos estudios basados en distintos tests y muestras también distintas. El procedimiento de administración informatizada es muy similar en ambos, y las medidas que se analizan son esencialmente las mismas. Para evitar repeticiones innecesarias, los dos estudios se presentarán conjuntamente.

Método

Instrumentos

En el primer estudio se utilizó un cuestionario de 22 ítems con formato de respuesta binario (SÍ/NO) tomados de los cuestionarios EPI-A (Eysenck y Eysenck, 1963) y EPQ-R (Eysenck, Eysenck, y Barrett, 1985). El cuestionario se componía de 2 escalas con 11 ítems cada una que medían las dimensiones de Extraversión (E) y Neuroticismo (N). Todos los ítems estaban redactados en positivo y medían en la misma dirección.

En el segundo estudio se utilizaron dos escalas del cuestionario de personalidad FFPI (Rodríguez-Fornells, Lorenzo-Seva, y Andrés-Pueyo 2001). El cuestionario utilizado es la versión española de un test holandés de 100 ítems desarrollado para medir el modelo de personalidad de los «cinco grandes» (e.g., Hendriks, Hofstee, y De Raad, 1999; Carrasco, Holgado, y del Barrio, 2005). Los ítems del FFPI utilizan un formato de respuesta graduada en 5 puntos (desde 'no aplicable en absoluto' hasta 'totalmente aplicable'). Cada escala se compone de 20 ítems, todos ellos redactados en positivo, pero midiendo la mitad de ellos en una dirección del rasgo y la otra mitad en dirección contraria (para controlar los efectos de aquiescencia). Para el presente estudio utilizamos los 40 ítems que correspondían a las dimensiones de E y N, las mismas que en el test binario.

Participantes

En el primer estudio el cuestionario fue administrado a una muestra de 750 estudiantes de Psicología y Pedagogía. Aproximadamente el 80% eran mujeres. El rango de edad iba de 17 a 45 años con una media de 22 años y una desviación típica de 6.10. En el segundo estudio el FFPI fue administrado a una muestra de 262 estudiantes de una escuela de graduados sociales. Aproximadamente el 68% eran mujeres. El rango de edad iba de 17 a 51 años, con media de 21 años y desviación típica de 5.39. En ambos casos el test se administró en grupo, por clases. La administración era voluntaria y anónima, y los únicos datos que se pedían al sujeto eran sexo y fecha de nacimiento.

Procedimiento

El procedimiento de administración informatizada fue el mismo en ambos estudios, diferenciándose tan sólo en las opciones de respuesta. Se utilizaron los lenguajes HTML y Java Script, y el diseño seguía las directrices dadas por Green, Bock, Humphreys y Reckase (1984), y Kyllonen (1991). La pantalla se dividía en dos zonas: una zona superior en la que se mostraba el enunciado del ítem, y una zona inferior en la que aparecían los botones de respuesta en el centro (dos botones en el primer estudio y cinco en el segundo). Primero se daban las instrucciones que también aparecían en pantalla (tomadas de Fazio, 1990, y discutidas arriba). Después el participante respondía a un ítem de muestra. Finalmente se presentaba el test. Los tres primeros ítems eran de práctica, los restantes eran los ítems del cuestionario que se presentaban de forma mezclada. Para cada ítem se registraba en milisegundos el tiempo que mediaba entre la presentación en pantalla y la respuesta.

Los tiempos menores de 0.5 segundos se truncaron a 0.5 segundos y los mayores a 40 segundos se truncaron a 40. Después, los tiempos se transformaron utilizando la transformación inversa y se reflejaron a fin de que las relaciones encontradas tuviesen la misma dirección que con los tiempos originales.

Medidas

Las medidas correspondientes a las variables predictoras fueron:

- Longitud del ítem. El número de letras incluyendo los espacios en blanco entre letras (Rogers, 1974).
- Complejidad del enunciado. El número de nodos del árbol sintáctico del enunciado (Ferreira, 1991). Cada ítem se analizó mediante el analizador lingüístico automático FreeLing 1.5 (Carreras, Chao, Padró, y Padró, 2004), obteniéndose como output la representación arbórea de la estructura sintáctica del ítem. Después se procedió al recuento manual del número de nodos.
- Orden de presentación. De 1 a 22 (binario), de 1 a 40 (respuesta graduada).
- Controversialidad. En el caso binario la medida fue la distancia en valor absoluto entre la media del ítem (p) y 0.5. En el caso de respuesta graduada, primero las respuestas se escalaron entre 0 y 1 y después se utilizó la misma medida.
- Índice de discriminación. En el caso binario se utilizó el índice estimado según el modelo logístico de dos parámetros. En el caso de respuesta graduada se usó el peso factorial estimado.

Análisis

En primer lugar se evaluó el ajuste de las respuestas al modelo psicométrico correspondiente. En ambos estudios las dimensiones evaluadas son teóricamente independientes, por lo que todas las escalas se analizaron por separado. En el primer estudio las escalas se calibraron mediante el modelo logístico de dos parámetros. En el segundo estudio los ítems se calibraron mediante el modelo de análisis factorial lineal, utilizándose además el procedimiento propuesto por Ferrando, Lorenzo-Seva y Chico (2003) para eliminar los efectos de aquiescencia.

Las predicciones planteadas en la parte teórica se evaluaron mediante procedimientos correlacionales: correlaciones producto-momento entre los predictores y los tiempos, y análisis de regre-

sión múltiple. Para evaluar la significación de las correlaciones bivariadas se obtuvieron los intervalos de confianza al 90% mediante 200 replicaciones autodocimantes (Bootstrap) utilizando los valores percentiles de la distribución empírica. Estos intervalos corresponderían a un test formal al 95% a una cola. La significación del ajuste de regresión múltiple se evaluó mediante el intervalo de confianza al 90% para el coeficiente de determinación múltiple corregido. En todos los casos en que se utilizó el modelo lineal, se verificó el supuesto de linealidad mediante inspección de los gráficos bivariados. Finalmente, las relaciones orden de presentación-tiempos se evaluaron mediante regresión no paramétrica con función de suavizado Gaussiana.

Resultados

Con respecto a los análisis previos, en ambos estudios el ajuste de las respuestas al modelo correspondiente podía considerarse razonablemente bueno. Los autores podemos proporcionar más detalles al lector interesado.

La tabla 1 presenta los descriptivos de los predictores y las correlaciones entre las características del ítem y las medias de los tiempos transformados. Los resultados pueden resumirse como sigue. Primero, todas las relaciones empíricas se comportan de acuerdo con las predicciones. Segundo, tal como se esperaba, la longitud del ítem es el factor con mayor impacto. Tercero, existen diferencias importantes en cuanto a la magnitud de las relaciones en ambos formatos. En el caso binario el orden de presentación no muestra una relación significativa con el tiempo. En cambio, en el caso de respuesta graduada, las características de superficie parecían tener una menor importancia relativa y el orden tenía un impacto considerable. Por último, en ambos formatos, las características psicométricas se relacionan con el tiempo en la forma esperada, pero el impacto, sobre todo en el caso de la controversialidad, parece ser bastante débil.

La figura 1 muestra la relación entre los tiempos y el orden de presentación en el caso de respuesta graduada. La tendencia decreciente es clara, y es especialmente acusada en los 7 u 8 primeros ítems, pero se mantiene a lo largo de los 40. Este resultado contrasta con el supuesto de que el efecto de orden se limita a los dos primeros ítems. Contrasta también con los resultados para el caso binario, en que no parece haber ningún efecto significativo de orden. La diferencia entre formatos está de acuerdo con un principio establecido en tareas de tipo cognitivo: el efecto de práctica sobre el tiempo de respuesta es más acusado cuanto más compleja es la tarea (Welford, 1980).

La tabla 2 muestra las correlaciones entre las características de longitud y complejidad y las varianzas de los tiempos transforma-

dos. Las relaciones van en la dirección predicha: a mayor longitud y complejidad, mayor variabilidad interindividual. Además, los efectos son considerables: en el caso de respuesta graduada las relaciones con las varianzas son más fuertes que con las medias.

La tabla 3 muestra los resultados del análisis de regresión múltiple para las medias. En negrita se muestran los coeficientes que pueden considerarse como estadísticamente significativos al nivel de 0.05.

Existen diferencias importantes en los resultados obtenidos con ambos tipos de ítems. En el caso binario, tan sólo la longitud del enunciado tiene un impacto significativo y esta característica, por sí sola, explica casi un 80% de la varianza de los tiempos. Este resultado apoyaría la conjetura, bastante pesimista, de Bartram y Powell (1983) de que el tiempo de respuesta refleja casi exclusivamente tiempo de lectura, y que al parcializar este tiempo de lectura no hay otra característica adicional que tenga un impacto significativo.

En el caso de respuesta graduada existen varias características con un impacto significativo. La longitud del enunciado sigue siendo el principal predictor, pero también cuentan la complejidad, el orden y la capacidad discriminativa del ítem. Todos ellos influyen en la dirección esperada desde la teoría. Globalmente, la proporción de varianza explicada puede considerarse alta, pero queda algo por debajo de la del caso binario.

La tabla 4 resume el análisis de regresión sobre las varianzas. En ambos formatos tan sólo la longitud tiene una pendiente significativa. La proporción de varianza explicada, sin embargo, es considerable.

Discusión y conclusiones

Consideramos que el presente estudio amplía el conocimiento de un tema que irá adquiriendo mayor importancia en un futuro próximo. Con respecto a los estudios anteriores, la nueva información se centra en dos grupos de resultados. En primer lugar, la dificultad y complejidad del ítem se relacionan no sólo con los tiempos medios, sino también con las variabilidades de los ítems, y el impacto sobre la variabilidad es casi tan importante como el impacto sobre la media. En segundo lugar, todas las relaciones encontradas están de acuerdo con las predicciones, pero parecen existir efectos diferenciales importantes en la magnitud de dichas relaciones según el tipo de formato. En el caso binario, la longitud del ítem tiene un impacto muy fuerte, y el impacto adicional del resto de características no llega a ser significativo. En el caso de respuesta graduada, la longitud, la complejidad, los efectos del orden y la capacidad discriminativa del ítem, contribuyen significativamente a la predicción de los tiempos medios.

Tabla 1
Descriptivos y correlación entre los predictores y el tiempo medio de respuesta por ítem

Característica	Formato binario			Formato rta. graduada		
	$\bar{X}; s_x$	r	I.C. 90%	$\bar{X}; s_x$	r	I.C. 90%
Longitud	49.31; 14.84	0.89	(0.80;0.94)	25.72; 8.41	0.61	(0.44;0.77)
Complejidad	18.68; 5.89	0.75	(0.57;0.85)	10.35; 3.60	0.36	(0.16;0.58)
Orden pres.	-; -	-0.12	(-0.42;0.24)	-; -	-0.49	(-0.63;-0.28)
Controv.	0.18; 0.11	-0.13	(-0.48;0.32)	0.15; 0.09	-0.20	(-0.50;0.03)
Discrim.	0.91; 0.41	-0.44	(-0.64;-0.17)	0.51; 0.15	-0.17	(-0.43;0.04)

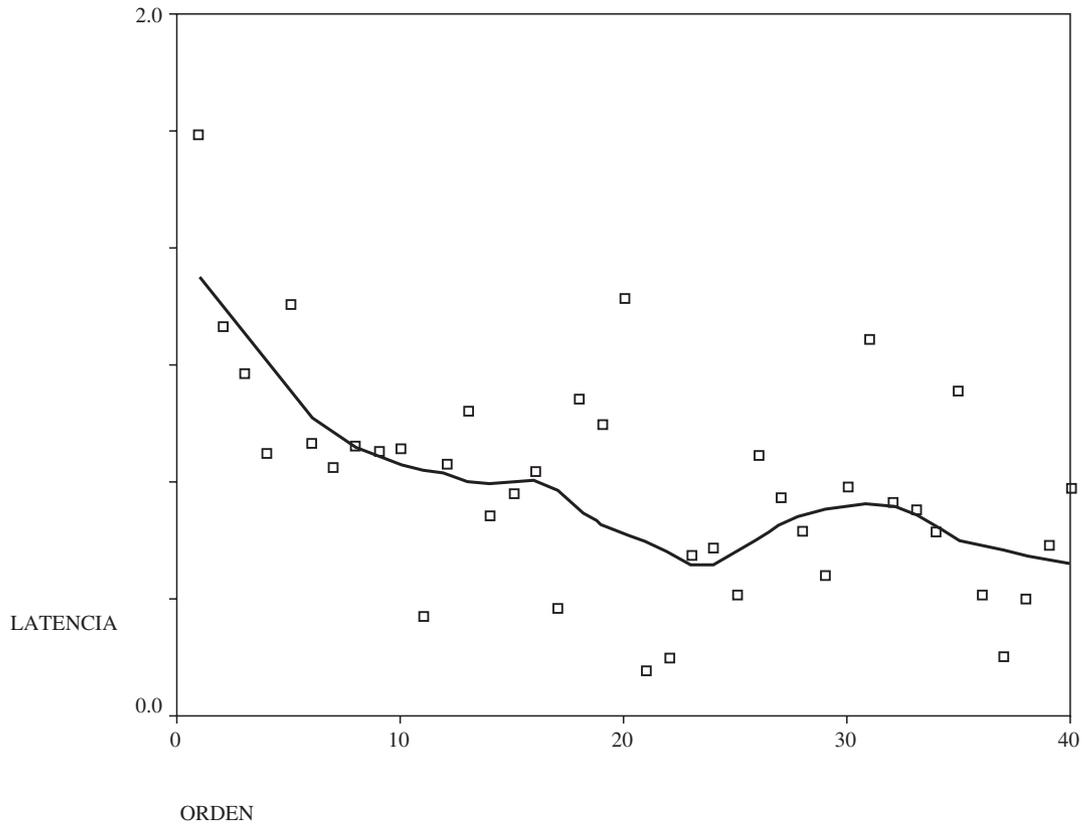


Figura 1. Relación entre el orden de presentación y la latencia media. Items de respuesta graduada

Tabla 2
Correlación entre los predictores y la varianza de los tiempos de respuesta por ítem

Característica	Formato binario		Formato rta. graduada	
	<i>r</i>	I.C. 90%	<i>r</i>	I.C. 90%
Longitud	0.75	(0.58;0.92)	0.66	(0.52;0.79)
Complejidad	0.59	(0.36;0.82)	0.45	(0.23;0.66)

Tabla 3
Ajuste de regresión múltiple sobre el tiempo medio: coeficientes estandarizados y correlación múltiple

Característica	Beta (I. binarios)	Beta (I. R. grad.)
Longitud	0.97	0.88
Complejidad	0.11	0.33
Orden Pres.	0.02	-0.52
Controversialidad	0.05	0.03
Discriminación	0.02	-0.23
R	0.89	0.84
R ²	0.79	0.71
R ² _c ; 90%IC	0.77(0.59;0.89)	0.66(0.55;0.80)

Una explicación plausible para el segundo grupo de resultados es que la respuesta a un ítem de respuesta graduada es una tarea cognitiva bastante más compleja que la respuesta a un ítem binario (Fazio, 1990). Un resultado que apoya esta conjetura es el claro efecto de orden observado en el caso respuesta graduada: los tiempos medios de respuesta tienden a disminuir a lo largo de los 40 ítems, contradiciendo la creencia de que el efecto se limita sólo a los primeros tres o cuatro. En cambio, en el caso binario no existe prácticamente efecto de orden. Sin embargo, aunque la explicación de los efectos diferenciales basada en la complejidad de la tarea es plausible, no pueden descartarse explicaciones alternativas, ya que los enunciados de los ítems no son los mismos en ambos formatos.

Desde un punto de vista aplicado, los resultados son algo decepcionantes y no permiten generar grandes expectativas respecto al uso de la información proporcionada por las latencias en el aná-

Tabla 4
Ajuste de regresión múltiple sobre la varianza de los tiempos: coeficientes estandarizados y correlación múltiple

Característica	Beta (I. binarios)	Beta (I. R. grad.)
Longitud	1.00	0.81
Complejidad	-0.27	-0.16
R	0.77	0.68
R ²	0.59	0.46
R ² _c ; 90%IC	0.55(0.30;0.77)	0.43(0.25;0.63)

lisis psicométrico de los ítems. Los índices psicométricos parecen relacionarse con las latencias en el sentido esperado, pero el impacto es débil. No parece, por tanto, que la información aportada por los tiempos pueda contribuir a una mayor precisión en la estimación de los parámetros de los ítems. Como ejemplo, Hanley (1962) había sugerido que para medir la controversialidad del ítem el mejor índice era el tiempo medio de respuesta a dicho ítem. Sin embargo, la optimista sugerencia de Hanley era, sobre todo, fruto de la especulación teórica. Los presentes datos, en cambio, sugieren que existe una relación negativa entre controversialidad y tiempo, pero dicha relación es demasiado débil para poder considerar al tiempo medio como un «índice de controversialidad». Por otra parte, un resultado novedoso de este estudio es que los índices de discriminación tienen una relación más importante con la latencia que la controversialidad del ítem, de forma que los ítems con mayor capacidad discriminativa tienden a ir asociados con menores tiempos de respuesta. Sin embargo, también aquí la relación pare-

ce demasiado débil como para utilizar esta información en la estimación del índice de discriminación.

En conclusión, consideramos que aunque los resultados sean poco esperanzadores, es importante darlos a conocer, ya que dan una idea realista de lo que se puede esperar de las latencias en investigación aplicada. Esta idea, como hemos comentado arriba, puede diferir notablemente de algunas expectativas puramente especulativas. Además, los resultados sugieren futuras líneas de investigación. Así, por una parte, creemos que sería necesario desarrollar esquemas de respuesta que permitiesen separar el tiempo de lectura del tiempo de respuesta, manteniendo al mismo tiempo la tarea lo más simple posible. Si la administración de un test informatizado se convierte en una tarea compleja de laboratorio, entonces se pierde la ventaja de recoger el tiempo como complemento a las respuestas sin costo adicional. Una segunda sugerencia sería concentrarse en las diferencias inter e intra individuales y en los efectos interactivos sujeto \times ítem.

Referencias

- Bartram, D., y Power, M. (1983). *An examination of response latencies as indices of faking on the Eysenck Personality Inventory*. Ergonomic research group, University of Hull, Report ERG/MS/83/12.
- Carrasco, M.A., Holgado, F.P., y del Barrio, M.V. (2005). Dimensionalidad del cuestionario de los cinco grandes (BFQ-N) en población infantil española. *Psicothema*, 17, 286-291.
- Carreras, X., Chao, I., Padró, L., y Padró, M. (2004). *FreeLing: An open source site of language analyzers*. Proceedings of the 4th International Conference on Language Resources and Evaluation (LREC'04). Lisboa, Portugal.
- Dunn, T.G., Lushene, R.E., y O'Neil, H.F. (1972). Complete automation of the MMPI and a study of its response latencies. *Journal of Consulting and Clinical Psychology*, 39, 381-387.
- 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.
- Fazio, R.H. (1990). A practical guide to the use of response latency in social psychological research. En C. Hendrick y M.S. Clark (eds.): *Research methods in personality and social psychology* (pp. 74-97). London: Sage.
- Ferrando, P.J., Lorenzo-Seva, U., y Chico, E. (2003). Unrestricted factor analytic procedures for assessing acquiescent responding in balanced, theoretically unidimensional personality scales. *Multivariate Behavioral Research*, 38, 353-374.
- Ferreira, F. (1991). Effects of length and syntactic complexity on initiation times for prepared utterances. *Journal of Memory and Language*, 30, 210-233.
- Fricke, B.J. (1957). A response bias (B) scale for the MMPI. *Journal of Counseling Psychology*, 4, 149-153.
- Green, B.F., Bock, R.D., Humphreys, L.G., y Reckase, M.D. (1984). Technical guidelines for assessing computerized adaptive tests. *Journal of Educational Measurement*, 21, 347-360.
- Hambleton, R.K. (2004). Theory, methods and practices in testing for the 21st century. *Psicothema*, 16, 696-701.
- Hanley, C. (1962). The 'difficulty' of a personality inventory item. *Educational and Psychological Measurement*, 22, 577-584.
- Helfrich, H. (1986). On linguistic variables influencing the understanding of questionnaire items. En Angleitner, A., y Wiggins, J.S. (eds.): *Personality assessment via questionnaires* (pp. 61-107). Berlín: Springer.
- Hendriks, A.A.J., Hofstee, W.K.B., y De Raad, B. (1999). The Five-Factor Inventory (FFPI). *Personality and Individual Differences*, 27, 307-325.
- Jensen, A.R. (1982). Reaction time and psychometric g. En H.J. Eysenck (ed.): *A model for intelligence* (pp. 93-132). Berlín: Springer.
- Kuncel, R.B. (1973). Response processes and relative location of subject and item. *Educational and Psychological Measurement*, 33, 545-563.
- Kuncel, R.B. (1977). The subject-item interaction in itemmetric research. *Educational and Psychological Measurement*, 37, 665-678.
- Kyllonen, P.C. (1991). Principles for creating a computerized test battery. *Intelligence*, 15, 1-15.
- Ratcliff, R. (1993). Methods for dealing with reaction time outliers. *Psychological Bulletin*, 114, 510-532.
- Rodríguez-Fornells, A., Lorenzo-Seva, U., y Andrés-Pueyo, A. (2001). Psychometric Properties of the Spanish Adaptation of the Five Factor Personality Inventory. *European Journal of Psychological Assessment*, 17, 133-145.
- Rogers, T.B. (1973). Toward a definition of the difficulty of a personality item. *Psychological Reports*, 33, 159-166.
- Rogers, T.B. (1974). An analysis of the stages underlying the process of responding to personality items. *Acta Psychologica*, 38, 205-213.
- Stout, R.L. (1981). New approaches to the design of computerized interviewing and testing systems. *Behavior Research Methods and Instruments*, 13, 436-442.
- Tetrick, L.E. (1989). An exploratory investigation of response latency in computerized administration of the Marlowe-Crowne social desirability scale. *Personality and Individual Differences*, 10, 1281-1287.
- Welford, A.T. (1980). *Reaction times*. New York: Academic Press.