DIFERENCIAS ENTRE LOS TEST INFORMATIZADOS DE PRIMERA GENERACIÓN Y LOS TEST EN PAPEL Y LÁPIZ: INFLUENCIA DE LA VELOCIDAD Y EL NIVEL DE DESTREZA INFORMÁTICA.

DIFFERENCES BETWEEN FIRST GENERATION COMPUTER BASED TESTS AND PAPER AND PENCIL TESTS: INFLUENCE OF SPEED AND LEVEL OF COMPUTER SKILLS.

DAVID ARRIBAS ÁGUILA*

Recibido 15-01-04 Aceptado 18-03-04

Resumen

Existen resultados contradictorios respecto a la equivalencia de los tests en papel y lápiz y los tests informatizados de primera generación. Las explicaciones a esta falta de homogeneidad han sido principalmente las diferencias entre los sistemas informáticos usados como punto de referencia (formato de presentación de los elementos, posibilidad de repasar elementos omitidos o previamente contestados) y las diferentes demandas perceptivas o motoras de cada uno de ellos. En este artículo se analizan otras variables que pueden resultar de gran importancia en la explicación de esta heterogeneidad, como pueden ser la destreza informática y el grado de velocidad de las pruebas psicológicas. En función de estos parámetros se analiza la equivalencia entre ambos tipos de soportes evaluativos y se proponen ciertas pautas a tener en cuenta de cara al diseño y evaluación de tests informatizados de primera generación.

Palabras clave

Test informatizados, test de papel y lápiz, evaluación.

Abstract

There are contradictory results about the equivalence of paper and pencil tests and computer based tests. The reasons for this lack of homogeneity have been the differences between the computer systems used as point of reference (item presentation, the possibility to view, respond to, skip or review items not answered or answered) and their different perceptive or motor demands. Other variables, as the computer skills and the speed level of the test, that can be very important in the explanation of this heterogeneity, are analyzed in this paper. According to these parameters, the equivalence between both assessment supports is analyzed and certain standards concerning the design of Computer Based Tests are proposed.

Key words

Computer based tests, paper and pencil tests, assessment.

* Departamento de I+D de TEA Ediciones.

Introducción

Son múltiples las ventajas empíricamente contrastadas de los tests informatizados de primera generación en comparación con sus más próximos antecesores en papel y lápiz, como pueden ser, entre otras, el menor tiempo de aplicación (Bounderson, Inouye y Olsen, 1989), la mayor precisión y rapidez de obtención de las puntuaciones (Pryor, 1989) y de devolución de resultados (Wise y Plake, 1989) o la mayor seguridad del test y el control del tiempo de respuesta (Olsen, Maynes, Slawson y Ho, 1989). Sin embargo, fuera de los Estados Unidos, el test informatizado en cualquiera de sus versiones no forma parte del uso habitual que los profesionales hacen de las evaluaciones psicológicas. Esta falta de aplicación práctica puede deberse principalmente a dos factores: la falta de un sistema científico, y a la vez fácil de usar, que sea accesible al profesional y la falta de confianza en los sistemas informatizados.

Respecto al primero de estos factores, en España recientemente se lanzó al mercado el primer sistema europeo de evaluación por Internet formado por tests informatizados de primera generación: e-teaediciones (TEA Ediciones, 2002). Este sistema cuenta con información exhaustiva de carácter técnico y se apoya en miles de aplicaciones y estudios científicos. En esta solución están implementados los tests más conocidos y usados en sus versiones de papel y lápiz, para poder ser aplicados por ordenador y obtener de manera automática los resultados e informes interpretativos.

Por otra parte, no es de extrañar la desconfianza que pueda crear en el profesional el uso de los tests informatizados, ya que los resultados encontrados respecto a la equivalencia con los tests en papel y lápiz son de por sí contradictorios. En el caso de las pruebas de personalidad y de actitudes parece consolidada la idea de que las diferencias son pequeñas o nulas (Rolls y Feltham, 1993). Sin embargo, los resultados con los tests de aptitudes se podrían dividir en tres grandes grupos:

1. *No existen diferencias entre el rendimiento en los tests informatizados y en los tests de papel y lápiz.* Algunos estudios han comprobado que no existen diferencias entre ambos procedimientos en los llamados tests de ejecución (Wise y Wise, 1987), en los tests de elección múltiple (Ward, Hooper y Hannafin, 1989) o en los tests puros de potencia (Mean y Drasgow, 1993). En esta misma línea estarían algunas conclusiones derivadas de la revisión de Mazzeo y Harvey (1988), donde se afirma que existen pequeñas diferencias pero de poca significación práctica.
2. *La ejecución en los tests informatizados es superior a la de los tests en papel y lápiz.* En este apartado estarían algunos estudios como los de Chin, Donn y Conry (1991) respecto a los tests de ciencias o los de Greaud y Green (1986) o Van der Vijer y Harsveld (1994) con los tests de velocidad.
3. *La ejecución en papel y lápiz es superior a la de los tests informatizados*, como demuestran algunos estudios con tests de aritmética como los de Lee, Moreno y Sympson (1986), la revisión de Wise y Plake (1989) o los trabajos de Goldberg y Pedulla (2002) con el *Graduate Record Exam*.

Las explicaciones más solventes sobre esta falta de acuerdo en los resultados obtenidos han sido la heterogeneidad de los diseños de las *interfaces* (Bartram, 1993; Bunderson y cols., 1989) y las diferentes demandas motoras o perceptivas de ambos tipos de procedimientos (Brand y Houx, 1992). La diversidad de programas informáticos para la aplicación de pruebas pone de relieve algunas variables de gran importancia a la hora de explicar las diferencias. La respuesta motora que demanda una aplicación en la cual haya que contestar con el teclado no es la misma que la de una prueba a contestar con el ratón, o incluso con ambos mecanismos conjuntamente. En esta misma línea, las demandas motoras (ratón, papel y ordenador, teclado...) y perceptivas (tamaño de la pantalla de presentación de estímulos, resolución gráfica...) pueden suponer un marco heterogéneo de cara a establecer y comparar los procesos que se ponen en juego en las aplicaciones informatizadas.

El objetivo del presente estudio es poner de manifiesto la influencia de otra serie de variables que, relacionadas con las explicacio-

nes anteriores, pueden ser igualmente determinantes en las diferencias de ejecución entre ambos procedimientos. En el marco de un sistema de aplicación informática común e implantado en el mercado (e-teaediciones), se parte de las siguientes hipótesis:

— La ejecución en papel y lápiz es superior a la de los tests informatizados en pruebas con una fuerte carga de velocidad.

— El grado de velocidad de los tests afecta significativamente a las diferencias entre ambos soportes de evaluación.

— A menor velocidad (o mayor componente de potencia) del test, menores diferencias existirán entre las puntuaciones derivadas de la aplicación en papel y lápiz y la informatizada.

— Existen diferencias significativas en las aplicaciones por ordenador en función del nivel de destreza informática de los sujetos.

Método

Muestra

En el estudio participaron un total de 108 sujetos (25 varones y 83 mujeres) con edades comprendidas entre los 22 y los 58 años (Media = 23,60; Dt = 4,18). La mayoría de la muestra (94 de los 108) estaba constituida por estudiantes de último curso de la licenciatura de Psicología del año académico 2002/2003 (Universidad Complutense de Madrid), divididos en 4 grupos de prácticas pertenecientes a dos asignaturas de la especialidad de Psicología del Trabajo. Al tratarse de una muestra incidental, se aprovechó esta agrupación para asignar al azar cada grupo a un tratamiento experimental.

Material e instrumentos

Para estudiar el rendimiento en ambos tipos de soportes evaluativos se diseñaron 2 pruebas con un grado máximo de velocidad (pruebas V) y 2 pruebas con escasa dificultad pero suficiente como para dotarlas de una carga sustancialmente menor de velocidad que las anteriores (pruebas D).

Cada una de las dos pruebas V estaba compuesta por 60 elementos a contestar en 60 segundos. En estas pruebas, al sujeto se le presentaba una letra como enunciado o estímulo y su tarea consistía en buscar, entre 4 opciones de respuesta, la única opción que contenía esa letra y marcarla. Aunque las letras que funcionaban como estímulo fueron elegidas aleatoriamente, la plantilla o disposición de las opciones correctas era la misma en las dos pruebas, para conseguir así un grado elevado de equivalencia o paralelismo entre ellas.

Las pruebas D estaban compuestas por 30 elementos de respuesta múltiple (dos opciones correctas) cada una a contestar en 60 segundos. La tarea a realizar era buscar, entre 4 opciones de respuesta, las letras anterior y posterior del abecedario a una dada y marcarlas. Exactamente igual que en las pruebas V, la disposición de las opciones correctas era la misma en los dos tests y las opciones seguían el orden alfabético desde la primera a la cuarta (p. ej., K L M N). Dado el formato de respuesta y el número de los elementos, cada respuesta a un elemento en las pruebas D guardaba un grado elevado de equivalencia con cada una de las respuestas a las pruebas V (60 elementos cada una)

En las 4 pruebas el sujeto debía rellenar unos datos de identificación (apellidos, nombre, edad...) e informar del número de horas semanales en las que actuaba con el ordenador en función de 4 niveles: de 0 a 10 horas semanales, de 11 a 20 horas semanales, de 21 a 30 horas semanales y más de 30 horas semanales.

En las aplicaciones en papel y lápiz se proporcionó a cada sujeto dos hojas tamaño DIN A-4. La primera era una hoja impresa por las dos caras, por un lado con las instrucciones y dos ejemplos y, en la parte posterior, con los elementos de la tarea a realizar. La otra era la hoja de respuestas en la cual los sujetos debían marcar con lápiz la respuesta a cada ítem.

En las aplicaciones informáticas se contó con un total de 58 equipos informáticos con conexión a Internet y con el sistema de evaluación e-teaediciones activo y preparado para la interacción con las pruebas. Este sistema requiere de la conexión a Internet únicamente en los momentos previos a la evaluación (carga

de las instrucciones y de los estímulos del test) y una vez finalizada ésta (para enviar las respuestas y corregirlas). Además, el programa encargado de presentar los estímulos y recoger las respuestas es residente en el ordenador, por lo que la velocidad y continuidad de la conexión en ningún caso pone en peligro los resultados de la evaluación.

Como se ha indicado anteriormente, el diseño del soporte informático en el que se realizan las evaluaciones puede ser determinante en la ejecución (Bartram, 1993; Bunderson y cols., 1989). Por esta razón, se describe brevemente el diseño del programa y el proceso que permite realizar las evaluaciones. La ventana en la cual se desarrolla la aplicación mide 800 x 600 píxeles. En primer lugar se proporcionan las instrucciones necesarias, de tal modo que una vez que el sujeto ha comprendido la tarea y ha resuelto dos ejemplos, puede comenzar a contestar al test. A continuación aparece en la parte superior de la ventana el enunciado de la tarea (en una franja de 800 x 113 píxeles), en la parte inferior el estímulo y, unas debajo de otras, las opciones de respuesta. Éstas deberán ser marcadas por el sujeto utilizando el ratón o podrán ser omitidas. Para pasar de un estímulo a otro existen dos botones en la parte inferior de la pantalla que pueden ser seleccionados y accionados con el ratón: uno para pasar al elemento siguiente y otro para retroceder y contestar a elementos omitidos o previamente contestados por el sujeto. Entre estos dos botones se sitúa el contador de tiempo, visible durante toda la aplicación. En la figura 1 se presenta un ejemplo del sistema.

Diseño, variables y condiciones de control

El diseño del estudio fue un diseño factorial 2x2 de medidas repetidas. Las variables que formaban parte del diseño fueron las siguientes:

- *Rendimiento*: variable dependiente operativizada como cantidad de aciertos en cada una de las pruebas, considerando que cada acierto suma un punto a la puntuación directa total.
- *Tarea* (Factor A): variable independiente que representa el nivel de destreza requerido

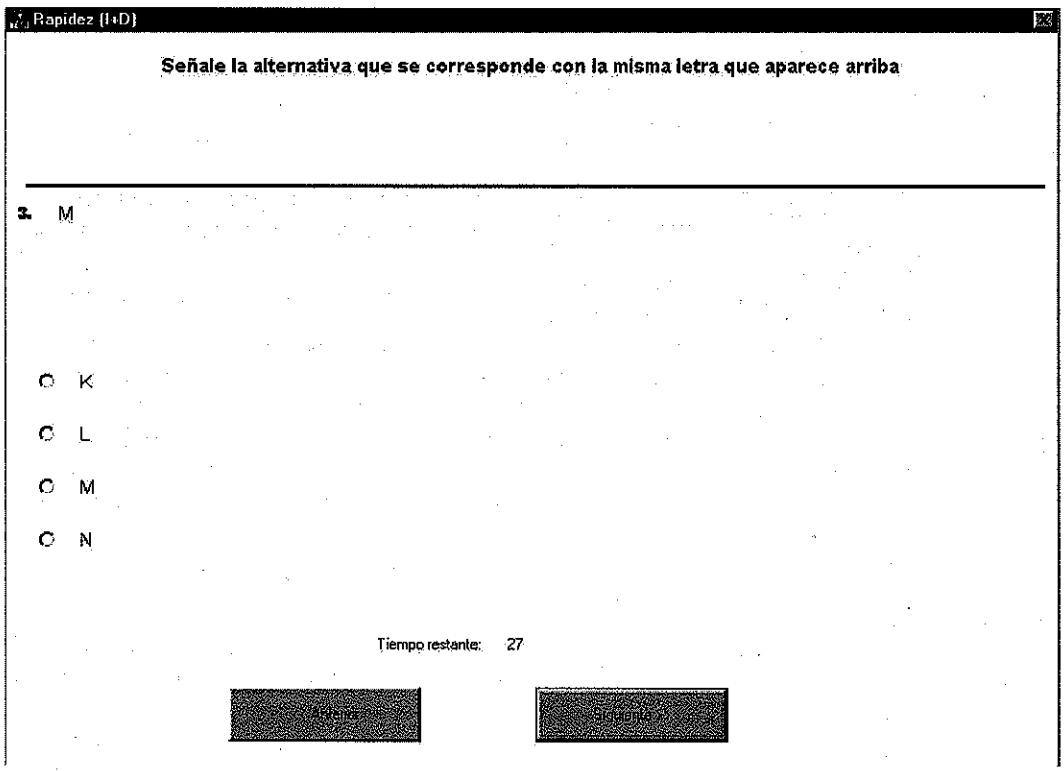


Figura 1. Evaluación mediante e-teaediciones

para acertar un conjunto de respuestas o grado de dificultad del test, con dos niveles:

— **TV.** Test V (a_1): prueba con mayor grado de velocidad (o menor grado de dificultad).

— **TD.** Test D (a_2): prueba con mayor grado de dificultad (o menor grado de velocidad).

- *Soporte de evaluación* (Factor B): variable independiente referida al tipo de formato utilizado para recoger la medida, con dos niveles:

— **PL.** Papel y lápiz (b_1): formato clásico de aplicación de pruebas.

— **ET.** Ordenador (b_2): formato informático mediante la utilización del sistema e-teaediciones.

La combinación de los diferentes niveles de las variables independientes dio lugar a los 4 tratamientos experimentales que aparecen en la tabla 1.

Por otra parte, se controlaron los efectos de otras variables que podían afectar a los resultados. Estas variables fueron:

— 0 a 10 horas

— 11 a 20 horas

— 21 a 30 horas

— Más de 30 horas

- *Fátiga / aprendizaje*: variable extraña controlada mediante el método de la equiparación parcial sistemática. De esta forma se aseguró que las unidades de errores progresivos fuesen iguales en cada una de las secuencias de los tratamientos experimentales y de los niveles de las variables independientes en su conjunto. Las secuencias experimentales fueron las siguientes:

— 1 2 3 4 — 2 3 4 1
— 3 4 1 2 — 4 1 2 3

- *Sexo y edad*: variables extrañas controladas mediante el método de la fluctuación aleatoria entre los distintos grupos.

Procedimiento

Las pruebas se aplicaron de forma colectiva en 4 sesiones en las que se informaba, al terminar la aplicación, de la finalidad del estudio.

Tabla 1. Tratamientos experimentales

		Soporte de evaluación	
		Papel y lápiz (b1)	**e-teaediciones (b2)**
Tarea	**Prueba V (a1)**	*Tratamiento 1* (a1b1)	*Tratamiento 2* (a1b2)
	Prueba D (a2)	*Tratamiento 3* (a2b1)	*Tratamiento 4* (a2b2)

- **OR.** *Nivel de destreza informática*: variable extraña controlada mediante la medición y categorización en 4 niveles dependiendo del número de horas semanales trabajando con el ordenador:

Las dos aulas en las que se llevaron a cabo las aplicaciones estaban formadas por 30 y 28 equipos informáticos respectivamente, con una separación aproximada entre puestos de un metro lateralmente y dos metros frontalmente.

En cada una de las dos aplicaciones en papel y lápiz, que se realizaban en la misma sesión a todos los sujetos, el examinador leía las instrucciones y resolvía los dos ejemplos en voz alta para que los sujetos comprendiesen la tarea a realizar. Una vez entendida la tarea se les indicaba el tiempo del que iban a disponer para contestar, se les daba la instrucción de comenzar la prueba y se ponía en marcha el cronómetro. Transcurridos los 60 segundos se les indicaba que dejaran de contestar. En las dos aplicaciones informáticas era el propio sujeto el que leía las instrucciones en la pantalla, planteaba las dudas al examinador y se resolvían en alto los dos ejemplos presentados. Igualmente, el tiempo era controlado por el sistema, de tal modo que transcurridos los 60 segundos el sujeto no podía interactuar con el ordenador.

Resultados

Análisis de la fiabilidad y velocidad de las pruebas

En primer lugar se estudió la consistencia interna de las pruebas (TV y TD) en los dos tipos de soporte de evaluación (PL y ET). Los resultados utilizando el coeficiente alfa de Cronbach son los siguientes:

— TV_PL: 0,9456
— TV_ET: 0,9360
— TD_PL: 0,9169
— TD_ET: 0,9227

Por otro lado, el análisis del grado de velocidad de las pruebas es uno de los elementos básicos del estudio, ya que se parte de la potencia como elemento que reduce las diferencias entre ambos tipos de soportes de evaluación. Por esta razón, se analizó el grado de velocidad de cada prueba independientemente del soporte de evaluación utilizado. Para ello, se calculó el índice de velocidad de Gulliksen (1950), la media de errores de cada prueba y se estimó el tiempo medio utilizado para contestar a cada elemento. Los resultados se presentan en la tabla 2.

Tabla 2. Estudio de la velocidad de las pruebas

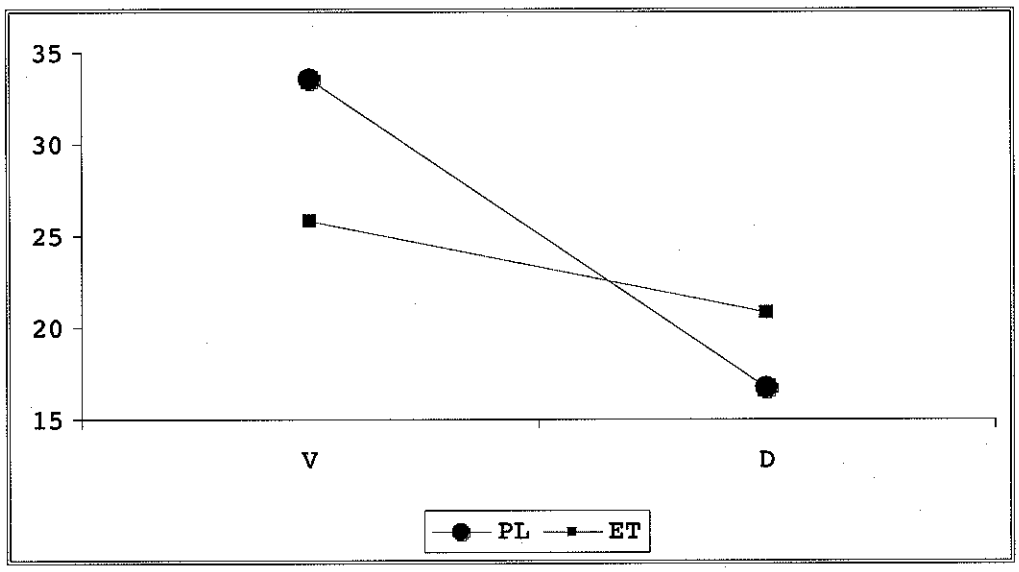
	Prueba V	Prueba D
IV Gulliksen	0,000	0,157
Media de errores	0,051	1,094
Tiempo medio por elemento (en seg.)	1,534	3,450

Los resultados indican que la prueba V tiene un grado máximo de velocidad, puesto que el índice de Gulliksen es 0. Esta fuerte carga se traduce en un segundo y medio por elemento al contestar y en una práctica ausencia de errores. En la prueba D existe igualmente una fuerte carga de velocidad (si el índice de Gulliksen va de 0 a 1 podríamos hablar de un 15% de dificultad añadida respecto a la prueba V), pero con una diferencia en dificultad óptima para la finalidad del estudio respecto a la prueba V.

Para conocer mejor estos aspectos, en el gráfico 1 se representa el porcentaje de sujetos que ha omitido o errado cada elemento en ambas pruebas.

A la luz de los datos, hay algunos aspectos importantes a destacar como son la ausencia de errores u omisiones hasta el elemento 17 de la prueba V en contraste con el 10% aproximado de la prueba D en los 12 primeros elementos. Además, el porcentaje de omisiones o errores aumenta gradualmente en la prueba V, de tal modo que el 50% de los sujetos no contestan al

Gráfico 1. Proporción de sujetos que no alcanzan u omiten cada elemento



elemento 33 y el 100% al 51. En la prueba D esos límites se encuentran en los elementos 18 y 33.

Por tanto, los tests V y D son dos pruebas de velocidad: una de ellas de velocidad máxima en la cual no existe dificultad al contestar a los elementos y otra con una fuerte carga de velocidad pero con un grado de dificultad superior en un 15% al de la anterior.

Diferencias derivadas del uso de un soporte u otro de evaluación e influencia del nivel de destreza con el ordenador

Se analizaron conjuntamente las diferencias o influencias existentes a partir de los tratamientos que formaban parte del estudio. El análisis llevado a cabo fue un Anova de medidas repetidas, previo contraste de esfericidad de Mauchly (Chi-cuadrado=0,000). Hay que decir que en ninguno de los contrastes fue necesario corregir los grados de libertad y que en todos ellos se estimó el tamaño del efecto y la potencia observada (alfa=0,05). Además, se decidió incluir en el análisis la variable OR (Nivel de destreza con el ordenador) para estudiar los posibles efectos intersujetos mediante un contraste F y sus correspondientes pruebas a posteriori. Los estadísticos descriptivos en función de estas variables se presentan en la tabla 3.

Los resultados del análisis fueron los siguientes:

— **Efectos principales**: existen diferencias significativas entre el rendimiento en los tests V y en los tests D ($F = 394,69$; $p < 0,001$; Eta cuad. parcial = 0,804). Estas diferencias se producen por una mayor puntuación global en las pruebas V (Media = 29,72) en comparación con las pruebas D (Media = 18,78). Igualmente, existen diferencias entre las aplicaciones en papel y lápiz y en e-teaediciones ($F = 9,17$; $p < 0,01$; Eta cuad. parcial = 0,087). En este caso las diferencias se producen por un mejor rendimiento en las aplicaciones en papel y lápiz (Media = 25,12) en contraste con el ordenador (Media = 23,38).

— **Efectos de la interacción**: existen diferencias significativas en la interacción de las variables Tarea y Soporte de evaluación ($F = 156,36$; $p < 0,001$; Eta cuad. parcial = 0,620). No se encontraron diferencias significativas entre la Tarea y OR ($F = 1,86$), el Soporte de evaluación y OR ($F = 1,59$) y entre ambas variables independientes y OR ($F = 0,92$). La interacción entre Tarea y

Tabla 3. Estadísticos descriptivos.

	Nº horas Ordenador	Media	Desv. tip.	N
TV_PLDe 21 a 30	De 0 a 10	29,80	7,23	35
	De 11 a 20	34,86	6,35	29
	34,30	9,05	20	
	Más de 30	38,50	9,28	16
	TOTAL	**33,56**	**8,22**	**100**
TV_ETDe 21 a 30	De 0 a 10	23,11	5,97	35
	De 11 a 20	26,79	6,11	29
	28,55	7,54	20	
	Más de 30	27,00	3,03	16
	TOTAL	**25,89**	**6,30**	**100**
TD_PLDe 21 a 30	De 0 a 10	15,45	6,18	35
	De 11 a 20	17,34	5,30	29
	16,00	4,63	20	
	Más de 30	19,06	5,68	16
	TOTAL	**16,69**	**5,63**	**100**
TD_PLDe 21 a 30	De 0 a 10	18,85	5,85	35
	De 11 a 20	21,82	4,61	29
	22,00	4,21	20	
	Más de 30	22,12	4,42	16
	TOTAL	**20,87**	**5,13**	**100**

Soporte de evaluación se presenta en el gráfico 2.

— **Efectos simples**: existen diferencias significativas en el rendimiento global de los sujetos en función del nivel de destreza con el ordenador ($F = 7,047$; $p < 0,001$; Eta cuad. parcial = 0,180). Estas diferencias, de acuerdo con las comparaciones múltiples de Bonferroni, se dan sólo entre el grupo de 0 a 10 horas y el resto de los grupos ($p < 0,01$).

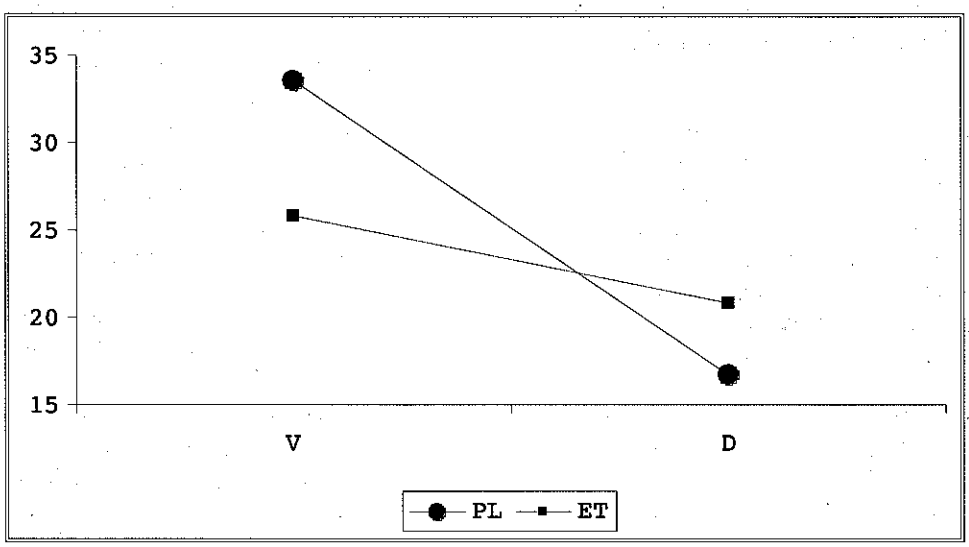


Gráfico 2. Interacción entre Tarea y Soporte de evaluación

Discusión

Los requisitos de la tarea y, en concreto, el grado de velocidad de las pruebas psicológicas parecen ser una causa más de la falta de acuerdo en los estudios que equiparan las aplicaciones de los tests informatizados de primera generación con los tests en papel y lápiz. En primer lugar, las diferencias a favor del papel y lápiz son máximas cuando la prueba requiere un grado extremo de velocidad al responder (derivado de un nivel muy bajo de dificultad de los elementos). Este hecho matiza los resultados derivados de los estudios de Greaud y Green (1986) y Van der Vijer y Harsveld (1994), ya que si la carga de velocidad del test es máxima los datos indican que las diferencias son sustanciales en favor del papel y lápiz. Sin embargo, cuando la carga de velocidad se reduce en un 15% los resultados coinciden con los de estos autores. Por tanto, podemos afirmar que la velocidad de los tests afecta significativamente a las diferencias entre los soportes de evaluación, por lo que parece ser un factor importante a tener en cuenta en los estudios futuros y en los diseños de plataformas informáticas de evaluación psicológica.

Por otro lado, cuando hemos incrementado mínimamente la dificultad de la prueba se han reducido en valores absolutos las diferencias entre ambos procedimientos de evaluación. En este sentido, confimamos parcialmente nuestra hipótesis referida a este aspecto, ya que realmente se han reducido las diferencias pero han variado su sentido (han cambiado de ser a favor del papel y lápiz para ser a favor del soporte informático). Esta circunstancia pone de manifiesto una vez más la importancia de la velocidad en el análisis de las diferencias. No obstante, queda abierta la puerta a futuros trabajos para confirmar esta tendencia a la inversión o incluso a la nulidad de las diferencias al tratar con tests como los que actualmente existen dentro del sistema e-teaediciones, es decir, de mayor dificultad que los de este estudio. Igualmente, sería interesante analizar los posibles efectos derivados de la planificación de la tarea en cuanto al tiempo de trabajo, ya que en la aplicación en papel y lápiz, a diferencia de la aplicación por ordenador, los sujetos no reciben información constante del tiempo que resta para finalizar la prueba.

Algunos de los principales puntos de partida a la hora de explicar la causa de las diferencias entre ambos soportes de evaluación eran el tiempo de ejecución de la respuesta motora y el nivel de destreza con el ordenador. Los dos aspectos parecen íntimamente relacionados, ya que a mayor nivel de destreza menor tiempo de respuesta y viceversa. Sin embargo, el nivel de destreza informática en este estudio no influye en las diferencias encontradas entre ambos soportes de evaluación, por lo que parece que esta variable no es determinante a la hora de realizar evaluaciones mediante sistemas informatizados como e-teaediciones. Las diferencias encontradas en este sentido parecen deberse a la heterogeneidad de los grupos del estudio en cuanto a la aptitud demandada por la tarea más que al propio soporte evaluativo, ya que el grupo con peor destreza informática obtiene menores puntuaciones en los dos soportes de evaluación y no únicamente en las aplicaciones informáticas. Por lo tanto, el diseño de la solución informática del sistema utilizado en el estudio parece asegurar la homogeneidad de los resultados derivados de las aplicaciones independientemente del nivel de destreza con el ordenador. Sin embargo, futuros estudios con un muestreo más representativo de cada uno de los niveles de destreza, así como con niveles extremos de destreza informática u operativizaciones distintas de dichos niveles, podrían arrojar relaciones importantes sobre la influencia de este aspecto. Además, sería interesante estudiar más en profundidad el efecto de la respuesta motora unida a la destreza informática con pruebas que evalúen otro tipo de aptitudes.

Referencias bibliográficas

Bartram, D. (1993). Emerging trends in computer-assisted assessment. En H.Schuler, J.L. Farr y M. Smith (Eds.), *Personnel selection and assessment*. Hillsdale, NJ: LEA.

Brand, N. y Houx, P.J. (1992). MINDS: Toward a computerized test battery for use in health psychological and neuropsychological assessment. *Behaviour Research Methods, Instruments and Computers, 24 (2)*, 385-389.

Bunderson, C.V., Inouye, D.K. y Olsen, J.B. (1989). The four generations of computerized educatio-

nal measurement. En R.L. Linn, *Educational measurement*. Nueva York: Macmillan.

Chin, C.H.L., Donn, S. y Conry, R.F. (1991) Effects of computer based tests on the achievement, anxiety and attitudes of grade 10 students. *Educational and Psychological measurement, 51*, 735-745.

Greaud, V.A. y Green, B.F. (1986). Equivalence of conventional and computer presentation of speed tests. *Applied Psychological Measurement, 10*, 23-34.

Gulliksen, H. (1950). *Theory of mental tests*. Nueva York: John Wiley.

Lee, J.A., Moreno, K.E. y Sympson, J.B. (1986). The effects of mode of test administration on test performance. *Educational and Psychological Measurement, 46*, 467-474.

Mazzeo, J. y Harvey, A.L. (1988). *The equivalence of scores from automated and conventional versions of educational and psychological tests: A review of the literature*. Princeton, NJ: Educational Testing Services.

Mead y Drasgow (1993). Equivalence of computerized and paper-and-pencil cognitive ability tests: a meta-analysis. *Psychological Bulletin, 114 (3)*, 449-58.

Olsen, J.B., Maynes, D.D., Slawson, D. y Ho, K. (1989). Comparisons of paper administered, computer administered and computerized adaptative achievement tests. *Journal of Educational Computing Research, 5*, 311-326.

Pryor, R.G.L. (1989). Some ethical implications of computer technology. *Bulletin of the Australian Psychological Society, November*, 164-166.

Rolls, S. y Feltham, R. (199·). Practical and professional issues in computer-based assessment and interpretation. En M. Smith y V. Sutherland (Eds.), *Professional issues in selection and assessment*. Chichester, England: John Willey and Sons.

TEA Ediciones (2002). e-teaediciones. Sistema de evaluación en línea. Web: http://www.e-teaediciones.com

Van der Vijer, F.J.R. y Harsveld, M. (1994). The incomplete equivalence of the paper and pencil and computerized versions of the general aptitude test battery. *Journal of Applied Psychology, 79*, 852-859.

Ward, T.J., Hooper, S.R. y Hannafin, K.M. (1989). The effect of computerized tests on the performance and attitudes of college students. *Journal of Educational and Computing Research, 5*, 327-333.

Wise, S.L. y Plake, B.S. (1989). Research on the effects of administering tests via computers. *Educational Measurement: Issues and Practices, Fall*, 5-10.

Wise, S.L. y Wise, L.A. (1987). Comparison of computer administered and paper administered achievement tests with elementary school children. *Computers in Human Behaviour, 3*, 15-20.