Psicothema
2000. Vol. 12, Supl. nº 2, pp. 358-362

ISSN 0214 - 9915 CODEN PSOTEG

Efecto del entrenamiento sobre las propiedades psicométricas de los tests

Julia Martínez-Cardeñoso, Eduardo García-Cueto y José Muñiz
Universidad de Oviedo

Se ha comprobado que algunos programas de entrenamiento influyen en las puntuaciones que las personas obtienen en los tests, si bien no hay unanimidad entre los investigadores acerca de la cuantía de esta influencia. Partiendo de este hecho, es razonable pensar que también las propiedades psicométricas de los tests vengan afectadas por los entrenamientos. En este trabajo se analiza el efecto de dos tipos de entrenamiento sobre las propiedades psicométicas del test. Uno de los entrenamientos se centra en la instrucción académica de la materia a evaluar, mientras que el otro hace hincapié en los aspectos estratégicos (*test wiseness)* implicados en la respuesta a tests de elección múltiple. Se utilizaron respectivamente dos muestras de 152 universitarios y 120 trabajadores de la administración. En el primer caso se estudió la influencia del entrenamiento sobre la fiabilidad, la validez y el funcionamiento diferencial de los ítems. Los resultados indican que el entrenamiento no origina diferencias significativas en la fiabilidad del test, ni parece generar un funcionamiento diferencial de los ítems. En cuanto a la validez de constructo, la estructura factorial se vería ligeramente afectada. Los resultados del segundo estudio indican que el entrenamiento en estrategias para responder a los tests no influye en la fiabilidad, pero sí en la validez de constructo de las pruebas.

Coaching effects on the psychometric properties of tests. In this research the effect of two different coaching programs on the psychometric properties of tests was studied. The first program was designed to coach tests in which academic knowledge is the construct assessed. In this program the training was focused on the academic contents to be assessed. The second program is focused in the strategies and test wiseness used by the subjects in order to answer multiple choice items. A sample of 152 university students was used for the first program, and a sample of 120 workers from the public administration was used for the second program. Test reliability is not affected by the use of coaching programs. Differential Item Functioning was not detected after the application of the coaching programs. However, slightly differences in construct validity (assessed through factor analysis) were observed as a consequence of the implementation of coaching programs. Theoretical and practical implications of the results obtained are discussed.

La mayoría de los resultados obtenidos hasta la fecha permiten afirmar con una seguridad razonable que los entrenamientos sistemáticos mejoran las puntuaciones de las personas en los tests (Anastasi, 1981; Bond, 1989; Messick y Jungeblut, 1981; Powers y Leung, 1995; Powers y Swinton, 1984), lo cual ha llevado a algunos autores a plantearse las posibles implicaciones de los entrenamientos en las propiedades psicométricas de los tests (Allalouf y Shakhar, 1998; Anastasi, 1981; Jones, 1986 a y b; Kelderman y Macready, 1990; Messick, 1982; Ortar, 1960; Powers, 1985). En línea con estos trabajos, el objetivo central de la presente investigación es el análisis del efecto que sobre las propiedades psicométricas de los tests puedan ejercer dos tipos de entrenamiento. El primero de ellos se basa en la instrucción académica de la materia a evaluar, mientras que el segundo hace hincapié en los aspectos estratégicos (*test wiseness)* implicados en la respuesta a tests de elección múltiple.

Correspondencia: Julia Martínez Cardeñoso
Facultad de Psicología
Universidad de Oviedo
33003 Oviedo (Spain)
E-mail: juliamc@sci.cpd.uniovi.es

Estudio 1

El objetivo de este estudio fue analizar el efecto del primero de los entrenamientos (instrucción académica de la materia a evaluar) sobre tres propiedades psicométricas fundamentales del test: fiabilidad, validez y funcionamiento diferencial de los ítems.

Método

Participantes

Se utilizó una muestra de 152 alumnos voluntarios de segundo curso de la Facultad de Psicología de la Universidad de Oviedo.

Tipo de entrenamiento

Se trataba de entrenar a los estudiantes para que mejorasen sus puntuaciones en una prueba académica de la asignatura de Psicometría. Para ello se les entrenó en: estrategias de resolución de tests, revisión de la materia, clases sobre la materia, horas de tutoría, horas de estudio de los alumnos etc. Desde la perspectiva de Messick y Jungesblut (1981), este tipo de entrenamientos se situaría en uno de los extremos del continuo de tipos de preparación cuyo objetivo es la instrucción intensiva enfocada a desarrollar el constructo y los conocimientos.

Procedimiento

Para llevar a cabo los entrenamientos citados se procedió de la siguiente manera: se aplicó a los alumnos un pretest de 50 ítems de elección múltiple (tres alternativas) de conocimientos psicométricos, posteriormente se les aplicó una prueba similar (examen oficial de la asignatura de Psicometría) como postest, llevando a cabo los entrenamientos entre ambas aplicaciones. En el entrenamiento se incluyeron diversos aspectos (Estrategias de reducción de ansiedad, familiarización con los tests y situación de examen, feedback, enseñanza y práctica de estrategias de distribución eficaz del tiempo del examen, estrategias para evitar errores, estrategias para responder las preguntas por azar, estrategias de razonamiento deductivo), si bien el componente esencial fue la enseñanza y estudio de la materia objeto del entrenamiento, a saber, los conocimientos psicométricos. Se realizaron cuatro sesiones de entrenamiento de una hora cada una, aparte de las horas de estudio empleadas por los alumnos para la preparación del examen oficial que se utilizó como postest.

Debido a que la asistencia de los alumnos a las sesiones de entrenamiento era voluntaria, se dispuso de las siguientes situaciones: a) 35 alumnos no realizaron el pretest, sólo el postest, b) 17 alumnos realizaron el pretest y postest y ninguna sesión de entrenamiento, c) 49 alumnos realizaron el pretest, postest y una sesión de entrenamiento, d) 23 alumnos realizaron el pretest, postest y dos sesiones de entrenamiento, e) 13 alumnos realizaron el pretest, postest y tres sesiones de entrenamiento, f) 15 alumnos realizaron el pretest, postest y cuatro sesiones de entrenamiento.

Resultados

Fiabilidad

Dado el escaso número de alumnos por cada nivel de entrenamiento (número de sesiones), se consideró como pertenecientes al grupo experimental a todas aquellas personas que hubiesen hecho alguna sesión de entrenamiento, además del pretest y el postest, en total 100; y como grupo control aquellos alumnos que realizaron pretest y postest o sólo postest, en total 52.

Debido a la diferencia numérica entre el grupo control y el experimental, se calculó alfa en un segundo grupo experimental (grupo experimental I) constituido por los alumnos que habían asistido a dos, tres o cuatro sesiones de entrenamiento, es decir, se eliminaban los resultados de las personas que habían asistido a una sesión de entrenamiento quedando un grupo de 51 personas, de tamaño semejante al grupo control.

Los resultados muestran que en el grupo control el coeficiente alfa es 0,79, mientras que en los grupos experimentales es 0.82.

Tabla 1
Coeficiente Alfa en el grupo control y grupos experimentales

	Coeficiente alfa	**N**
Grupo control	0,79	52
Grupo experimental	0,82	100
Grupo experimental I	0,82	51

Las diferencias obtenidas en el coeficiente alfa fueron analizadas mediante el estadístico de contraste de Feldt (1969) para muestras independientes, no hallándose diferencias estadísticamente significativas ($p<0,05$).

Validez de constructo

Para estudiar el efecto del entrenamiento en la validez de constructo se realizó una comparación entre las estructuras factoriales del postest para el grupo control y el grupo experimental.

En el análisis factorial del grupo control se obtuvieron 18 factores con valor propio superior a uno, que explicaban el 80% de la varianza. En el análisis factorial del grupo experimental se obtuvieron 19 factores con valor propio superior a uno que explicaban el 72,5% de la varianza. Debido al elevado número de factores extraídos por este criterio, se realizaron análisis factoriales en ambos grupos, extrayendo 2, 3, 4 y 5 factores sucesivamente. Posteriormente, se realizó la comparación entre las estructuras factoriales del grupo control y experimental, utilizando el índice de Wrigley y Neuhans (1955) para el estudio de la congruencia entre dos soluciones factoriales obtenidas con las mismas variables y distintas muestras.

Los resultados muestran que el índice de congruencia sólo es significativo en el primer factor de todos los análisis factoriales realizados (2, 3, 4 y 5 factores), lo que apunta a que el entrenamiento podría estar influyendo en la estructura factorial de la prueba, afectando a los factores de menor importancia.

Funcionamiento diferencial de los ítems

El posible funcionamiento diferencial de los ítems debido al entrenamiento fue analizado por los métodos de regresión logística y Mantel-Haenszel. Estos métodos fueron aplicados a cada uno de los 50 ítems del postest.

Los resultados obtenidos en la regresión logística muestran que en la mayoría de los ítems el acierto de los mismos depende de la puntuación final de las personas por lo que no se puede considerar que haya funcionamiento diferencial. Tan sólo en tres ítems el acierto de los mismos influye, además de la puntuación, el nivel de entrenamiento o número de sesiones en las que participaron los alumnos por lo que estos ítems si parecen funcionar diferencialmente.

Para la aplicación del procedimiento Mantel-Haenszel se utilizó como grupo focal el grupo de las personas que habían realizado sólo el postest y las que habían realizado el pretest y postest y

ninguna sesión de entrenamiento, en total 52 personas. El grupo de referencia lo constituyen todas aquellas personas que habían realizado alguna sesión de entrenamiento, en total 100 personas.

Dicho análisis se realizó sobre dos, tres y cuatro categorías, detectándose tres ítems con funcionamiento diferencial en el caso de dos y tres categorías y dos ítems cuando se utilizaron cuatro categorías ($p<0,05$).

Discusión y conclusiones

Con relación a la fiabilidad, se detectan ligeras diferencias entre el grupo control y los grupos experimentales, si bien como ocurría en el trabajo de Powers (1985), estas diferencias no resultan estadísticamente significativas. No parece, por tanto, que este tipo de entrenamientos afecten de forma relevante a la fiabilidad del test.

En cuanto al efecto del entrenamiento en la validez de constructo los resultados indican que se produce variación en los factores secundarios, como lo refleja el hecho de que el índice de congruencia sea significativo sólo en el primer factor de todos los análisis factoriales realizados (2, 3, 4 y 5 factores). Estos resultados pueden apoyar la hipótesis de Messick (1982) y Jones (1986 a y b) según la cual, el aumento en la puntuación debida a un aumento en la aptitud no produce cambios en la validez del test (primer factor), pero cuando el aumento en la puntuación es debido a estrategias y trucos para seleccionar las respuestas, la validez puede verse afectada (factores secundarios).

Sin embargo, hay que tomar estos resultados con suma precaución, dado que en esta investigación el número de personas en cada grupo es bajo, concretamente en el grupo control el número de personas es prácticamente igual al número de ítems, cuando lo que se aconseja es que haya al menos cinco personas por variable (Gorsuch, 1983, 1988; Stevens, 1996). Una posibilidad de considerar los factores fiables sería que nuestros datos se ajustasen a la regla propuesta por Guadagnoli y Velicer (1988), según la cual: a) Los factores con cuatro o más saturaciones por encima de 0,60 pueden considerarse fiables, independientemente del tamaño muestral. b) Factores con diez o más saturaciónes bajas (en torno a 0,40) son fiables siempre que el tamaño muestral sea superior a 150. c) Factores con saturaciones factoriales bajas no serán interpretados a menos que el tamaño muestral sea 300 o más.

Una revisión de los pesos factoriales de los análisis factoriales realizados en esta investigación permite comprobar que el primer factor del grupo control de dichos análisis factoriales tenía tres saturaciones superiores a 0.60 y entre ocho y doce saturaciones comprendidas en un entorno $\varepsilon(0.50, 0.10)$. En el primer factor del grupo experimental, las saturaciones resultaron más bajas, dependiendo del número de factores solicitados en el análisis factorial, se definieron entre 12 y 14 saturaciones comprendidas en un entorno $\varepsilon(.50, .10)$. En el resto de los factores tanto del grupo control como experimental las saturaciones eran menores. Si bien la regla de Guadagnoli y Velicer (1988) no se cumple *sensu stricto*, acogiéndonos al criterio de Stevens (1996), según el cual esta regla no debe ser tomada de forma estricta sino como base de interpretación, se podría considerar fiable el primer factor, tanto del grupo control como del grupo experimental de todos los análisis factoriales realizados, a pesar del reducido número de personas por grupo; sin embargo, el resto de los factores no se pueden considerar fiables, pues debido a que sus saturaciones son bajas, el número de personas por grupo resulta muy pequeño.

El estudio del funcionamiento diferencial de los ítems se realizó por los métodos de Regresión Logística y Mantel-Haenszel. Si bien ninguno de estos dos métodos funciona bien con muestras menores de 200 personas en cada grupo (Swaminathan y Rogers, 1990); Mazor, Clauser y Hambleton (1992) consideran que el método Mantel-Haenszel puede ser útil en muestras de 200 personas para identificar ítems con un funcionamiento diferencial claro. Sin embargo, dado que el funcionamiento del Mantel-Haenszel no es muy diferente entre las muestras 50/50, 100/50, 200/50 y 100/100 (Muñiz, Hambleton y Xing, 1997), cabe esperar que en nuestra muestra 100/52 se hayan detectado aquellos ítems que tenían un marcado funcionamiento diferencial.

Otra dificultad en la detección de ítems con funcionamiento diferencial con el método Mantel-Haenszel en muestras pequeñas es el número de categorías. El número ideal de categorías que se aconseja realizar son tantas como número de ítems del test más una, a medida que se reducen las categorías de este número aumentan las posibilidades de detectar falsos positivos o error tipo I (Muñiz, 1998). En nuestro caso, el número de categorías realizadas 2, 3 y 4 es muy inferior a las 51 categorías que se deberían de realizar, por lo que es posible que algunos de los ítems identificados con funcionamiento diferencial sean falsos positivos. Dadas las características de los análisis realizados para la detección del funcionamiento diferencial de los ítems y considerando el reducido número de ítems en los que se detectó funcionamiento diferencial, puede concluirse que el entrenamiento genera funcionamiento diferencial en algunos ítems, pero su incidencia es ciertamente baja, por lo que se puede considerar que estos resultados concuerdan con los obtenidos por Allalouf y Shakhar (1998), introduciendo dudas en las críticas realizadas por algunos autores referidas a que el entrenamiento generaría sesgos en los tests, otra cosa bien distinta sería el impacto introducido.

Estudio 2

Como ya se ha señalado, en este estudio se analiza el efecto que los entrenamientos de estrategias tienen sobre las propiedades psicométricas de los tests de elección múltiple. Se analizan estos efectos sobre la fiabilidad y la validez.

Método

Participantes

En este programa de entrenamiento participaron 120 personas empleadas en los servicios de limpieza de la administración asturiana.

Entrenamiento

El programa de entrenamiento aplicado se desarrolló en 21 horas, tres horas cada día. Durante los siete días que duró se les entrenó en siete técnicas fundamentales: estrategias de reducción de ansiedad, familiarización con los tests y situación de examen, Feedback, enseñanza y práctica de estrategias de distribución del tiempo del examen eficazmente, aprendizaje de las estrategias para evitar errores, aprendizaje de estrategias para responder las preguntas al azar y uso de estrategias de razonamiento deductivo. Nótese que aquí, al contrario que en el estudio 1, todo el entrenamiento es independiente de los contenidos de las pruebas, no en-

trenando directamente sobre ellos, no se entrenaba ningún contenido en especial, únicamente estrategias para responder a los tests de elección múltiple.

Para evaluar el efecto del entrenamiento diseñado se utilizó como pretest y postest una misma prueba psicotécnica de aptitudes de 100 ítems de elección múltiple con cinco alternativas. La prueba se aplicó el primer y último día de entrenamiento, con un tiempo máximo para realizarlo de 75 minutos. Dada la naturaleza de la prueba, el recuerdo de la primera aplicación a la segunda resulta prácticamente imposible.

Resultados

Fiabilidad

Para estudiar el efecto del entrenamiento sobre la fiabilidad del test, se calculó la fiabilidad por el método test-retest, dado que se había aplicado el mismo test a las mismas personas en dos ocasiones. Se obtuvo un coeficiente de fiabilidad de 0,50, que puede considerarse como bajo. Parece evidente que el entrenamiento tiende a reordenar las puntuaciones obtenidas por las personas en el pretest, rebajando la correlación. Esta explicación es coherente con los valores encontrados para el coeficiente alfa, bastante más elevados: 0,75 en el pretest y 0,80 en el postest.

Tabla 2
Coeficientes de fiabilidad

	Fiabilidad	**N**
Test-Retest	0,51	120
Alfa (pretest)	0,75	120
Alfa (postest)	0,80	120

Para determinar si la diferencia entre el coeficiente alfa del pretest y el del postest es estadísticamente significativa, se empleó el estadístico de contraste de Feldt (1980) para muestras dependientes. Las diferencias no resultaron estadísticamente significativas ($p<0,05$).

Validez

En el estudio del efecto del entrenamiento en la validez de constructo se realizó una comparación entre las estructuras factoriales en el pretest y en el postest. Para ello se realizó un análisis factorial con rotación oblicua del pretest y del postest. En dichos análisis factoriales se obtuvieron 33 factores con valor propio superior a uno en el pretest (que explicaban el 78% de la varianza total) y 34 factores en el postest (que explicaban 77,4 % de la varianza total). Debido al elevado número de factores se realizó un análisis factorial de segundo orden. Se obtuvieron 14 factores con valor propio mayor que 1 en el pretest, cuyo porcentaje de varianza explicada resultó ser de 57,9%. En el postest se obtuvieron 15 factores con valor propio mayor que 1 y porcentaje de varianza explicada del 59%. Tanto el análisis descriptivo de los pesos factoriales, como los coeficientes de congruencia calculados, indican que, a excepción del primer factor, ambas estructuras factoriales (pretest y postest) son bastante diferentes.

Discusión y conclusiones

Puesto que no se dispone de grupo control es imposible saber con certeza si la baja estabilidad de las puntuaciones (0,50) se debe en su totalidad al entrenamiento, o qué parte es achacable al propio test. Por otra parte, los resultados del coeficiente alfa son superiores al coeficiente de fiabilidad de estabilidad, lo que indica que la consistencia interna de la prueba resulta más elevada que su estabilidad temporal. Cabe subrayar la ligera variación positiva que se produce en el coeficiente alfa del examen pretest y postest (de 0.75 a 0.80), por lo que el entrenamiento en estrategias para la resolución de tests no parece influir negativamente en la fiabilidad del test evaluada a través de la consistencia interna.

Con relación a la validez de constructo, los resultados parecen indicar que cuando se realiza un entrenamiento potente de estrategias éste puede alterar la estructura factorial de la prueba. Como ocurría con la fiabilidad test-retest, la explicación más plausible sería el fuerte impacto del entrenamiento sobre las puntuaciones, lo que podría explicar también porqué en el estudio 1 el entrenamiento influye en las estructuras factoriales secundarias.

Los resultados obtenidos en la validez de constructo, según los cuales el aumento de las puntuaciones produce una variación en la estructura factorial, no concuerdan con los estudios de simulación realizados por Baydar (1990), según los cuales el entrenamiento de estrategias puede aumentar la puntuación en los tests, pero la validez no se ve afectada por ellos dado que no se produce aumento en la adquisición de habilidades. De nuevo se impone la cautela a la hora de generalizar los resultados, puesto que el número de ítems del test (100) es prácticamente coincidente con el número de personas de la muestra (120) y las saturaciones de los factores son bajas. Además, también hay que tener en cuenta que el test resultó excesivamente difícil para los participantes, con lo que se pudo añadir un factor distorsionante adicional.

Referencias

Allalouf, A. & Shakhar, G. B. (1998). The effect of coaching on the predictive validity of scholastic aptitude tests. *Journal of Educational Measurement, 35*(1), 31-47.

Anastasi, A. (1981). Coaching, test sophistication, and developed abilities. *American Psychologist, 36*(10), 1086-1093.

Baydar, N. (1990). Effects of coaching on the validity of the SAT: Results of a simulations study. En Willingham. W. W., Lewis, Ch., Morgan, R. y Ramist, L. (1990). *Predicting College Grades: An analysis of institutional Trends over Two Decades*.(pp.213-224) Princeton, N. J.: Educational Testing Service.

Bond, L. (1989). The effects of special preparation on measures of scholastic ability. En: R. Linn. *Educational Measurement*. Nueva York: MacMillan Publishing Company.

Feldt, L. S. (1969). A test of the hypothesis that Conbach's alpha or Kuder-Richardson coefficient twenty is the same for two test. *Psychometrika, 34*, 363-373.

Feldt, L. S. (1980). A test of the hypothesis that Cronbach Alpha reliability coefficient is the same for two tests administered to the same sample. *Psychometrika, 45*, 99-105.

Gorsuch, R. L. (1983). *Factor analysis*. Hillsdale, NJ: Lawrence Erlbaum.

Gorsuch, R. L. (1988). Explor atory factor analisis. En J. R. Nesselroade y R. B. Cattell, (eds.), *Handbook of multivariate experimental Psycho - logy*. Nueva York: Plenum.

Guadagnoli, E. Y Valicer, W. (1988). Relation of sample size to the stability of component patterns. *Psychological Bulletin, 103*, 265-275.

Jones, R. F. (1986a). A comparison of the predictive validity of the MCAT for coached and uncoached students. *Journal of Medical Education, 61*(4), 335-338.

Jones, R. F. (1986b). The effect of commercial coaching courses on performance on the MCAT. *Journal of Medical Education, 61*(4), 273-284.

Kelderman, H. y Macready, G. (1990). The use of loglinear models for assessing differential item functionning across manifest and latent examine groups. *Journal of Educational Measurement, 27*(4), 307-327.

Mazor, K., Clauser, B. Y Hambleton, R.K. (1992). The effect of sample size on the functioning of the Mantel-Haenszel statistic. *Educational and Psychological Measurement, 52*, 443-451.

Messick, S. (1982). Issues of effectiveness and equity in the coaching controversy: Implications for educational and testing practice. *Educational Psychologist, 17*, 67-91.

Messick, S. y Jungeblut, A. (1981). Time and method in coaching for the SAT. *Psychological Bulletin, 89*(2), 191-216.

Muñiz, J. (1998). *Teoría clásica de los tests.* Madrid: Pirámide.

Muñiz, J., Hambleton, R. K. & Xing, D. (1997). *Small sample studies to detect flaws in test translation.* V Congreso de Metodología de las Ciencias Humanas y Sociales. Sevilla.

Ortar, G. R. (1960). Improving test validity by coaching. *Educational Re - search, 2*, 137-142.

Powers, D. E. (1985). Effects of coaching on GRE aptitude test scores. *Journal of Educational Measurement, 22*(2), 121-136.

Powers, D. E. & Leung, S. W. (1995). Answering the new SAT reading comprehension questions without the passages. *Journal of Educational Measurement, 32*(2), 105-129.

Powers, D. E. & Swinton, S. S. (1984). Effects of self-study for coachable test item types. *Journal of Educational Psychology, 76*, 266-278.

Stevens, J. (1996). *Applied multivariate statistics for the social sciences*. Hillsdale, N. J.: Lawrence Erlbaum.

Swaminathan, H. & Rogers, H. J. (1990). Detecting differential item functioning using logistic regression procedures. *Journal of Educational Measurement, 27*(4), 361-370.

Wrigley, L. & Neuhans, J. O. (1955). *The matching of two sets of factors.* Contract Memorandum Report. University of Illinois.