anales de psicología
2005, vol. 21, nº 2 (diciembre), 339-344

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

Ítems politómicos *vs.* dicotómicos: Un estudio metodológico

José Antonio López Pina*

Universidad de Murcia

Resumen: En este estudio se comparan cinco formatos de respuesta para ítems politómicos, desde el formato original que consta de cuatro categorías a un formato dicotómico (sólo dos categorías de respuesta). La comparación de distintos análisis psicométricos (análisis de ítems, estudio de la fiabilidad, análisis de componentes principales y estudio de la habilidad) entre los cinco formatos prueba que un formato dicotómico aporta casi tanta información sobre la calidad del tests y las puntuaciones observadas en depresión como un formato politómico de 4 categorías, por lo que el número de categorías de los ítems debe ser estudiado cuidadosamente y probado experimentalmente antes de aplicar el test.
Palabras clave: Formato de los ítems; análisis de ítems; fiabilidad; análisis de componentes principales.

Title: Politomous *vs.* dichotomous items: A methodological study.
Abstract: In this report we compare five response formats for politomous items, from the original format (four categories) to dichotomous format. Different psychometric analysis (item analysis, study of the reliability, principal component analysis and ability analysis) have been done on the five different formats. These psychometric analyses show that two categories can obtain as information as four. Then it is necessary to study the appropriate categories in the items before to apply the test.
Key words: Item format; item analysis; reliability; principal component analysis.

Desde que Likert (1932) introdujo la posibilidad de que las respuestas a los ítems podrían ordenarse en función de categorías a las que se asignaba un número entero, se ha dado por llamar a los ítems con categorías ordenadas o politómicos como ítems tipo Likert, y a los tests construidos con este tipo de ítems como escalas tipo Likert. Los ítems tipo Likert, politómicos en adelante, se hicieron muy populares en el ámbito de la evaluación de las actitudes, intereses y personalidad; estos ítems presentan algún tipo de escalamiento de la respuesta, normalmente de 1 a 5, donde 1 suele reflejar la categoría más baja, con alguna etiqueta adosada tal como Muy en Desacuerdo, y 5 representa la categoría más alta, con una etiqueta semejante como Muy de Acuerdo. La popularidad de este tipo de politomía ha sido tan elevada que prácticamente cualquier investigador que se plantea construir un test de actitudes, motivación, intereses o personalidad asume, sin reflexión, que la respuesta a los ítems de ese test debe presentar una estructura politómica (tipo Likert) de 1 a 5, u otra parecida; otros esquemas válidos pero menos utilizados son de 1 a 3 ó 1 a 7 ó 1 a 9 ó 1 a 11, generalmente impar. Pero, ¿por qué algunos de los tests más famosos en el contexto de la personalidad (MMPI ó 16PF) emplean ítems dicotómicos y no politómicos? ¿Es que un ítem cuyas respuestas se recogen bajo un formato dicotómico evalúa el atributo peor que si el formato de respuesta es politómico? Cualquier investigador responderá, seguramente sin demasiada reflexión, que sí, ya que el formato politómico permitirá recoger todas las posibilidades de respuesta, y esta riqueza permitirá una mayor variabilidad de las puntuaciones, y por ende una mejora de su discriminabilidad (Janda, 1998) y de la fiabilidad y validez del test (Nunnally y Bernstein, 1995). En este estudio utilizaremos la palabra test para representar el conjunto de ítems de actitudes, intereses o personalidad y no la palabra escala, como parece ser costumbre entre los investigadores de este campo de investigación; así, un test de habilidad o capacidad en nada se distingue de una escala de personalidad, en tanto que ambos siguen el mismo procedimiento de evaluación. En la práctica, la palabra test se debe utilizar cuando se emplea algún procedimiento de asignación de números a los atributos de los objetos y/o sujetos en función de una regla (Stevens, 1954), generalmente desconocida. Dawes (1972) llamó a este tipo de medición, medida índice; Torgerson (1958) la denominó medida por "fiat". Un conjunto de ítems formarán una escala si se formula explícitamente la regla de asignación de los números a las propiedades del atributo medido (Michell, 1999).

Para responder a las preguntas arriba formuladas, sería conveniente conocer como se postula el funcionamiento del proceso de respuesta a un ítem, ya sea dicotómico o politómico, bajo el modelo clásico de tests (MCT). En este modelo, y otros procedimientos de medida índice como la técnica Likert, no se hace ninguna previsión sobre la estructura de las respuestas que se recogerán de los ítems. Se asume, sin prueba, que las categorías están igualmente espaciadas no sólo físicamente en el papel, sino psicológicamente, en la mente de los evaluados, tal que la distancia psicológica que existe entre Muy en Desacuerdo y En desacuerdo es equivalente a la distancia que existe entre De Acuerdo y Muy de Acuerdo. Así que, utilizar cinco o más categorías igualmente espaciadas, sin probar experimentalmente que los sujetos evaluados funcionan bajo un esquema psicológico similar, se convierte en un mero artefacto estadístico cuyo objetivo es incrementar la variabilidad de las puntuaciones, y por ende el rango de los valores posibles del test. El investigador sabe que un aumento del rango de puntuaciones en la escala puede llevar a que la ordenación de las puntuaciones de los sujetos evaluados sea más estable de una ocasión a otra, y como consecuencia el coeficiente de fiabilidad sea mayor. Pero no conviene olvidar que el coeficiente de fiabilidad es una propiedad de la muestra (Thompson, 1994; Thompson y Vacha-Haase, 2000) y no del test, por lo que el aumento de la varianza de las puntuaciones generará artificialmente un aumento del coeficiente de fiabilidad, sin que realmente sepa-

*** Dirección para correspondencia** [Correspondence address]: José A. López Pina. Facultad de Psicología. Universidad de Murcia (Campus de Espinardo). Apdo. 4021, 30080 Murcia (España). E-mail: jlpina@um.es

mos la calidad métrica de las puntuaciones que se han obtenido con ese test.

La Teoría de la Respuesta al Ítem (TRI) es bastante más explícita sobre el proceso de respuesta que subyace a un ítem bajo cualquier tipo de formato. Lord y Novick (1968) asumen que para cada ítem existe una variable aleatoria Γ_j que representa la propensión de los sujetos a responder al ítem en una categoría. La distribución de esta variable aleatoria puede ser normal o logística. La propensidad de un sujeto a situarse en una categoría dada es desconocida, por lo que Lord y Novick (1968) asumieron la existencia de un umbral γ_k que controla la respuesta de cada sujeto. Si la propensión Γ_j a responder el ítem es menor o igual que el umbral γ_k, la respuesta, por ejemplo en un ítem dicotómico, se situará en una categoría 0, mientras que la si propensión $\Gamma_j > \gamma_k$, entonces la respuesta se situará en la categoría 1. Si el formato del ítem tiene un formato de respuesta dicotómico existirá un solo umbral (*k*). Sin embargo, si el ítem es politómico, existirán tantos umbrales como categorías se hayan establecido. En realidad, existirán *k* - 1 umbrales de respuesta. Así, resulta que cada ítem politómico tiene una distribución de propensidad (normal o logística) y tantos umbrales como categorías (*k* – 1) presentan las posibles respuestas. Es decir, un ítem politómico de 5 categorías, tendrá 4 (5-1) umbrales de respuesta. Estos umbrales deben determinarse en todos los ítems del test; si son iguales para todos los ítems, tenemos el modelo de escalas de tasación de Andrich (1978), mientras que si los umbrales cambian de ítem a ítem, tenemos el modelo de crédito parcial (Wright y Masters, 1982). Así que el primer problema que nos encontramos desde la perspectiva de la TRI es estadístico. Para un ítem con formato dicotómico es necesario estimar sólo un umbral, el que corresponde al cambio del sujeto de la categoría 0 (fallo, No o cualquier otra etiqueta) a 1 (Acierto, Sí o cualquier otra etiqueta), mientras que un ítem con formato politómico requiere estimar tantos umbrales como categorías menos una se hayan formulado. Por ejemplo, un ítem de cinco categorías, donde los números asignados a las categorías pueden ser: 0, 1, 2, 3 y 4, requiere que se estimen cuatro umbrales: el umbral de paso de 0 a 1, el umbral de 1 a 2, de 2 a 3, y de 3 a 4. Nótese que utilizamos en la primera categoría el número 0, y no el número 1, como es costumbre en el formato tipo Likert; no existe ningún problema en ello, ya que los números ejercen su función sólo nominalmente, y además el número 0 indica mejor la ausencia/disconformidad total del atributo que cualquier otra etiqueta (Bond y Fox, 2001).

Un segundo problema en los ítems con formato politómico es el tamaño muestral necesario para encontrar frecuencias significativas en todas las categorías propuestas. Bond y Fox (2001) aconsejan que se debe colapsar aquellas categorías con frecuencias muy bajas ó 0, hasta obtener un número suficiente de sujetos en cada categoría. Pero también es posible, un tercer problema, tener un tamaño muestral elevado y, sin embargo, encontrar que algunas categorías tienen frecuencias muy bajas o incluso que nadie las haya contestado. ¿Qué ha sucedido? Simplemente que las categorías propuestas no han funcionado como se esperaban o también que los sujetos que han contestado ese ítem no han funcionado psicológicamente con los umbrales propuestos por el investigador sino, generalmente, con un número menor. Por ejemplo, es corriente encontrar en ítems con formato politómico de cinco categorías una doble joroba, donde los sujetos se sitúan mayoritariamente en las categorías 1 y 2, y 4 y 5, mientras que la categoría 3 tiene una frecuencia muy baja. En este caso, aunque el investigador planteó el ítem como una politomía, en realidad los sujetos han manejado psicológicamente sólo un umbral de paso, el que diferencia a las categorías 1 y 2 de las categorías 4 y 5, lo que es un ejemplo claro de formato dicotómico. Es frecuente, pues, que el investigador plantee el formato politómico sin un planteamiento experimental previo donde se estudie detenidamente en un grupo piloto el número de categorías apropiado cuando se desea un formato politómico; investigar el número de categorías que funciona adecuadamente para cada tipo de atributo que se quiere medir es, pues, un prerrequisito antes de aplicar el test a la población general (Bond y Fox, 2001).

Un tamaño muestral suficiente para poder interpretar adecuadamente las categorías de los ítems con formato politómico y una investigación detenida del número de categorías apropiado en los ítems, son situaciones que se deben plantear previamente a la construcción de cualquier test psicométrico, sea bajo el Modelo Clásico de Test o la Teoría de la Respuesta al Ítem, y en cualquier ámbito aplicado: capacidades/habilidades, intereses, motivación, actitudes y personalidad.

En este estudio se estudian las conclusiones estadísticas y psicométricas a las que se llegó con una escala de depresión (CES-D), comparando distintos esquemas de puntuación después de colapsar las categorías propuestas en el test original. La hipótesis de trabajo es que si un número menor de categorías ofrece los mismos resultados interpretativos que un número de categorías mayor, se debe escoger la solución más parsimoniosa, es decir, la de menor número de categorías.

Método

El test de depresión CES-D (*Center for Epidemiologic Studies-Depression Scale*) (Radloff, 1977) es un instrumento ampliamente utilizado en USA para evaluar la depresión. Consta de 20 ítems con formato politómico de 4 categorías (0, 1, 2 y 3), donde se le pide a cada sujeto que describa cuan a menudo se ha sentido de esta forma durante la última semana. Las etiquetas de las categorías son: Categoría 0: raramente (menos de un día); 1: algunas o pocas veces (1 – 2 días); 2: ocasional o moderadamente (3 – 4 días); 3: muchos días (5 – 7 días). La escala original aparece en la Figura 1. El orden de los ítems no es el del test original, sino el resultante después de haber realizado el análisis de componentes principales.

1. I felt that I could not shake off the blues even with the help of my family or friends.
2. I felt depressed.
3. I felt lonely.
4. I had crying spells.
5. I felt sad.
6. I felt fearful.
7. I thought my life had been a failure.
8. I felt that I was as good as other people.
9. I felt hopeful about the future.
10. I was happy.
11. I enjoyed life.
12. I was bothered by things that usually don't bother me.
13. I did not feel like eating; my appetite was poor.
14. I felt that everything was an effort.
15. My sleep was restless.
16. I could not "get going".
17. I had trouble keeping my mind on what I was doing.
18. I talked less than usual.
19. People were unfriendly.
20. I felt that people disliked me.

Figura 1: Test CES-D.

Los datos analizados corresponden a una muestra americana a la que se administró el test CES-D, y que aparece como una base de datos genérica para aplicar técnicas multivariantes en Afifi y Clark (1984). La muestra está compuesta por 294 casos a los que se administró el test CES-D. Algunas características descriptivas de esta muestra son las siguientes: Un 38.1% eran varones, mientras que un 61.9% fueron mujeres; la edad media del grupo fue de 44,4 años con una desviación típica de 18.1, variando entre 18 y 89 años. Con respecto al estado civil, un 25.2% eran solteros, un 43,2% casados, un 12,9% eran viudos, un 14,3% divorciados y el 4.4% restante separados. Un 9.1% de los componentes del grupo eran amas de casa, un 4.7% estaban en situación de desempleo, un 0.7% eran estudiantes, un 56.4% estaban trabajando a tiempo completo, un 14.3% estaban trabajando a tiempo parcial, un 13.3% eran jubilados, el 1.3% restante estaba en otras situaciones. El nivel de ingresos medio fue de 20514 dólares con una desviación típica de 15.3, variando entre 2000 y 65000 dólares anuales. Por último, un 10.6% profesaba la religión judía, un 53.1% era protestante, un 17.5% se declaraban católicos y un 18.8% declararon no profesar ninguna religión.

Un análisis de componentes principales (ACP) (Afifi y Clark, 1984) sobre la muestra de 294 casos reveló la existencia de cuatro componentes: Afecto Negativo (ítems 1 al 7); Afecto Positivo (ítems 8 al 11); Actividad retardada y somática (ítems 12 al 18) y un componente interpersonal (ítems 19 y 20).

A partir del formato original de 4 categorías se construyeron otros cuatro formatos que aparecen en la Tabla 1. Aunque existen otras combinaciones posibles, estas cuatro combinaciones de las categorías parecen las más adecuadas. El formato 1 es el formato original; el formato 2 representa la posibilidad de que las categorías 0 y 1 se colapsen en una única, representando la dificultad para establecer subjetivamente el estado de depresión y la diferencia entre nada (situación poco probable) y algo. El formato 3 ejemplifica justamente lo contrario, las dos categorías más elevadas (2 y 3) de depresión no reciben las frecuencias suficientes para asumir que existe un umbral claro entre estas dos categorías. El formato 4 representa a un ítem dicotómico donde la categoría 1 se colapsa con la categoría 0 (ausencia de depresión) y las dos finales (2 y 3) se colapsan en una sola, indicando que el umbral de paso entre 2 y 3 no se puede obtener claramente. Por último, el formato 5 representa un ítem dicotómico, donde la categoría 0 representa ausencia del atributo, mientras que cualquier otra (1, 2 y 3) representa presencia del atributo, sin poder establecer claramente los umbrales de paso entre las tres categorías.

Tabla 1: Formatos para colapsar las categorías de los ítems del test CES-D.

Formato	*Categorías / Codificación*			
1	0	1	2	3
2	0	0	1	2
3	0	1	2	2
4	0	0	1	1
5	0	1	1	1

Cada uno de estos cuatro formatos más el original dio lugar a una matriz de datos diferente. A estas matrices de datos se les aplicaron los análisis tradicionalmente empleados en el contexto del modelo clásico de tests. En primer lugar, se realizó un análisis de ítems donde se obtuvieron los índices de dificultad y homogeneidad de todos los ítems, después se obtuvieron los coeficientes de fiabilidad por el procedimiento de las dos mitades y el coeficiente alfa de Cronbach (Crocker y Algina, 1986) y, por último, se realizó un análisis de componentes principales de todas las matrices para comparar los resultados con la solución original. Para completar el análisis psicométrico se estudiaron las distribuciones de puntuaciones observadas resultantes de aplicar cada uno de los cinco formatos de respuesta. Todos los análisis se realizaron con el paquete estadístico SYSTAT (v. 7.0).

Resultados

Análisis de ítems

La selección de ítems con el MCT supone realizar un estudio de la dificultad y homogeneidad de los ítems incluidos en el test (Crocker y Algina, 1986). En el caso de un ítem politómico, el índice de dificultad es la media obtenida en la muestra a través de las cuatro categorías. Aunque se suele asumir que este indicador no es importante en los tests de personalidad, actitudes o intereses, su interpretación es clara y no ofrece confusión. Si la media del ítem está cercana a la categoría 0 (categoría más baja), supone que gran parte de los sujetos consideran ese ítem como un indicador bajo del atributo medido, mientras que si la media está cercana a 4 (categoría más elevada), supone que una mayoría de sujetos lo considera un indicador elevado del atributo medido.

El índice de homogeneidad de un ítems politómico se evalúa a través de la correlación producto-momento de Pearson (en el formato dicotómico se utiliza la correlación biserial-puntual o la correlación biserial) entre el patrón de respuestas obtenidas por el ítem y la puntuación total resultante de sumar las respuestas a todos los ítems. Este índice, generalmente, es positivo, varía entre 0 y 1, donde 0 es un indicador de una relación nula del ítem con el test, y 1 indica máxima relación del ítem con el total del test. Un índice de homogeneidad alto indica que el ítem mide lo mismo que el resto de ítems del test, pero no es un indicador de unidimensionalidad del test (McDonald, 1999).

La Tabla 2 presenta las correlaciones entre las medias de los ítems en los diferentes formatos de los ítems empleados en este estudio. Todas las correlaciones están por encima de .90, pero es destacable que estos resultados apuntan a que el formato original de cuatro puntos podría reducirse al formato 3 (sólo tres categorías) manteniendo un 99.6% de varianza explicada, y si se redujera al formato 5 (dicotómico), la proporción de varianza explicada sería aún del 96.6%.

Tabla 2: Correlaciones entre las medias de los ítems en los cinco formatos.

	Formato 1	Formato 2	Formato 3	Formato 4	Formato 5
Formato 1	--				
Formato 2	.954	--			
Formato 3	.998	.934	--		
Formato 4	.969	.995	.955	--	
Formato 5	.983	.881	.992	.909	--

La Tabla 3 presenta las correlaciones entre los índices de homogeneidad obtenidos a partir de los cinco formatos empleados en este estudio. De nuevo, la reducción de 4 a 3 categorías (formato 3) supone un porcentaje de varianza explicada del 98.8%, y la reducción a 2 categorías (formato 5) supone un porcentaje de varianza explicada del 90.4%.

Tabla 3: Correlaciones entre los índices de discriminación de los ítems en los cinco formatos.

	Formato 1	Formato 2	Formato 3	Formato 4	Formato 5
Formato 1	--				
Formato 2	.962	--			
Formato 3	.994	.937	--		
Formato 4	.957	.984	.942	--	
Formato 5	.951	.856	.971	.852	--

Análisis de la fiabilidad

La Tabla 4 presenta los coeficientes de fiabilidad obtenidos a través del procedimiento de las dos mitades, aplicando después la ecuación de Spearman-Brown para el caso de longitud doble (Gulliksen, 1950; Lord y Novick, 1968), y el coeficiente alfa de Cronbach (Crocker y Algina, 1986). Todos los coeficientes de fiabilidad obtenidos a través del procedimiento de las dos mitades se mantienen en torno a .90, siendo la diferencia entre el formato original (formato 1) y el formato 3 (tres categorías) de sólo .003 unidades, mientras que la diferencia entre el formato original (formato 1) y el formato 5 (dicotómico) es de .011 unidades. En el caso del coeficiente alfa, la primera diferencia (formato 1 vs. formato 3) es de .005 unidades, mientras que la segunda diferencia (formato 1 vs. formato 5) es de .023 unidades. Diferencias bastante escasas en el coeficiente de fiabilidad que permiten aventurar algunas hipótesis de cómo el formato de puntuación de los ítems puede dar indicios del modo de funcionamiento de los sujetos a la hora de contestarlos.

Tabla 4: Coeficiente de Spearman-Brown y coeficiente alfa para los cinco formatos.

	Coeficiente Spearman-Brown	Coeficiente alfa
Formato 1	.909	.893
Formato 2	.887	.871
Formato 3	.906	.888
Formato 4	.861	.853
Formato 5	.898	.870

Análisis de componentes principales

Afifi y Clark (1984) presentaron un ACP del test de depresión. Con el formato politómico original de 0 a 4 puntos resultaron los cuatro componentes descritos en el método. El mismo procedimiento de ACP fue aplicado a cada uno de los cuatro formatos restantes de este estudio. Ya que los tres primeros componentes explican entre un 45.3% (formato 4) y un 49.1% (formato 1), comentaremos los resultados con sólo estos tres componentes. La Tabla 5 presenta las correlaciones de Pearson entre los cinco formatos en cada uno de los tres primeros componentes una vez rotados.

Tabla 5: Correlaciones entre las cargas factoriales rotadas en los tres componentes principales (C. I a C. III) en los diferentes formatos experimentales (F. 1 a 5).

		Formato 1	Formato 2	Formato 3	Formato 4	Formato 5
	F. 1	--				
	F. 2	.965	--			
C. I	**F. 3**	.994	.943	--		
	F. 4	.948	.985	.936	--	
	F. 5	.947	.852	.969	.834	--
	F. 1	--				
	F. 2	.902	--			
C. II	**F. 3**	.972	.796	--		
	F. 4	.957	.941	.914	--	
	F. 5	.894	.656	.959	.803	--
	F. 1	--				
	F. 2	.840	--			
C. III	**F. 3**	-.887	-.581	--		
	F. 4	.681	.789	-.569	--	
	F. 5	.712	.384	-.730	.138	--

Como se aprecia en esta tabla, las correlaciones entre las cargas del componente I están por encima de .947, indicando una baja incidencia en la extracción e interpretación de este componente del formato empleado en el test a través de colapsar las categorías. Así, un formato dicotómico (formato 5) ofrece un porcentaje de varianza explicada del 90% contra un formato politómico de cuatro categorías (formato 1). Ese porcentaje se reduce al 80% en el componente II y al 51% en el componente III. De estos resultados se deduce que un formato dicotómico produce, al menos en los componentes I y II, resultados equivalentes al formato politómico original de cuatro puntos.

No obstante, algunas investigaciones (Hattie, 1985; McDonald, 1985) han puesto de manifiesto que el ACP, cuando se aplica sobre una matriz de correlaciones entre ítems dicotómicos tiende a ofrecer un primer componente que correlaciona altamente con el índice de dificultad de los ítems, y en el caso de un ítem politómico con la media de cada ítem. Para comprobar si esta situación se ha reproducido en nuestro análisis, hemos correlacionado las medias de los cinco formatos con las cargas factoriales del primer componente principal. También hemos calculado la correlación entre las cargas factoriales de este componente y los índices de homogeneidad de los ítems obtenidos en cada formato. Los resultados aparecen en la Tabla 6. Obviamente estos resultados no descalifican las investigaciones anteriores sino que muestran que en presencia de ítems cuyas medias se encuentran entre .108 (formato 3) y .445 (formato 1), y desviaciones típicas .049 y .175 respectivamente, las cargas factoriales resultantes en el componente I tienen una relación más elevada y significativa con los índices de homogeneidad de los ítems, tal como se espera, que con la media de los ítems del test en cada formato.

Tabla 6: Correlaciones entre las cargas factoriales no rotadas en el primer componente principal (C. I) y las medias de los ítems y los índices de homogeneidad en cada formato.

	Correlación Media/ Homog.	Correlación Media/ C. I	Correlación Homog/C. I
Formato 1	.155	.045	.991
Formato 2	.173	.055	.990
Formato 3	.247	.121	.987
Formato 4	.136	.022	.991
Formato 5	.176	.033	.985

Análisis de la habilidad

Por último, y no por ello menos importante, presentamos los resultados obtenidos después de puntuar a cada sujeto según el formato de respuesta propuesto. La Tabla 7 recoge las correlaciones entre las puntuaciones obtenidas con cada uno de los cinco formatos de respuesta propuestos.

Los resultados son elocuentes y van en la misma dirección que los presentados hasta ahora. El formato 5 (dicotómico) explica un 85% del formato 1 original (politómico), mientras que el formato 3 (politómico con tres categorías) explica un 97% del formato politómico original de 4 categorías. De ahí que se plantee una duda razonable sobre si sería suficiente en este caso haber empleado un formato dicotómico, o como mucho un formato politómico de tres puntos.

Tabla 7: Correlaciones entre las puntuaciones observadas en cada uno de los formatos.

	Formato 1	Formato 2	Formato 3	Formato 4	Formato 5
Formato 1	--				
Formato 2	.923	--			
Formato 3	.986	.856	--		
Formato 4	.941	.970	.910	--	
Formato 5	.920	.897	.962	.761	--

Discusión

A la vista de estos resultados parece que, en este caso, y con el modelo clásico de tests, emplear un formato politómico no mejora sistemáticamente la medida de depresión sobre un formato dicotómico. Es decir, prácticamente la misma información psicométrica se ha obtenido de un formato politómico que de un formato dicotómico, en sus aspectos esenciales (análisis de ítems, análisis de la fiabilidad y validez factorial), lo que aconseja a los investigadores reducir el número de categorías de respuesta de los ítems, ya que los sujetos de la muestra parece que utilizaron psicológicamente menos categorías de respuesta que las cuatro originalmente propuestas. Esto puede ciertamente sorprender, aunque no se puede generalizar a todos los tests en todas las situaciones experimentales, ya que las características de la muestra empleada en el estudio pone de manifiesto una media (8.908) en depresión muy por debajo de lo esperado (30), dando lugar a una distribución con un fuerte sesgo positivo (1.681). Queda para una futura investigación comprobar qué puede suceder en el caso de que la distribución de habilidad esté centrada en torno a la media esperada en el test, para poder obtener conclusiones más generalizables. Una distribución empírica centrada en la media esperada seguramente elevará la frecuencia de las categorías propuestas y permitirá realizar nuevos planteamientos sobre la conveniencia o no de solapar esas categorías, en función de que ofrezcan más o menos información de la habilidad de los sujetos. Bond y Fox (2001) y Stone, (1998, 2003) presentan sugerencias muy interesantes sobre la construcción de ítems politómicos que seguramente ayudarán a aliviar sus limitaciones en algunos contextos aplicados.

En cualquier caso, el investigador en psicología debe saber que la utilización de un formato politómico bajo el MCT descansa, en primer lugar, no sobre un hecho probado científicamente de que evalúa mejor el atributo que un formato dicotómico, sino sobre un hecho asumido (no investigado generalmente) acerca de su mayor potencialidad como herramienta de asignación de números, que amplia el rango de puntuaciones obtenidas por los sujetos, la discriminabilidad de los ítems, y la fiabilidad del test que los incluye. Y, en segundo lugar, sobre una definición deficiente de medida en psicología. Medir no es asignar numerales a las propiedades de los objetos en función de una regla (Stevens, 1954). Medir supone descubrir la naturaleza cuantitativa de los atributos psicológicos, y para ello es necesario encontrar la estructura ordinal y aditiva de las magnitudes de esos atributos (Michell, 1999).

Referencias

Afifi, A. A. y Clark, V. (1984). *Computer-aided multivariate analysis*. Belmont, CA: Lifetime Learning Pub.

Andrich, D. (1978). Scaling attitude items constructed and scored in the Likert tradition. *Educational and Psychological Measurement*, *38*, 665-680.

Bond, T. G. y Fox, Ch. M. (2001). *Applying the Rasch model: Fundamental measurement in the human sciences*. New Jersey: LEA.

Crocker, L. y Algina, J. (1986). *An introduction to classical and modern test theory*. New York: Holt, Rinehart & Winston.

Dawes, R. M. (1972). *Fundamentals of attitude measurement*. New York: Wiley.

Gulliksen, H. (1950). *Theory of mental test*. New York: Wiley.

Hattie, J. A. (1985). Methodological review: Assessing unidimensionality of test and items. *Applied Psychological Measurement*, *9*, 139-164.

Janda, L. H. (1998). *Psychological testing: Theory and applications*. Boston: Allyn and Bacon.

Likert, R. A. (1932). A technique for the measurement of attitudes. *Archives of Psychology*, *22*, 40-53.

Lord, F. M. y Novick, M. R. (1968). *Statistical theories of mental test scores*. Reading Mass: Addison-Wesley.

McDonald, R. (1999). *Test theory: A unified treatment*. New Jersey**:** LEA.

Michell, J. (1999). *Measurement in Psychology: A critical history of a methodological concept*. Cambridge**:** Cambridge University Press.

Nunnally, J. C. y Bernstein, I. J. (1994). *Psychometric theory*. New York**:** McGraw-Hill.

Radloff, L. S. (1977). The CES-D scale: A self-report depression scale for research in the general population. *Applied Psychological Measurement*, *1*, 385-401.

Stevens, S. S. (1946). On the theory of scales of measurement, *Science*, *103*, 667-680.

Stone, M. H. (1998). Rating scale categories: Dichotomy, double dichotomy and the number two. *Popular Measurement*, *1*, 61-65.

Stone, M. H. (2003). Substantive scale construction. *Journal of Applied Measurement*, *4*(3), 282-297.

SYSTAT (1997) (v. 7.0). *The System for Statistics*. SPSS Inc.

Thompson, B. (1994). Guidelines for authors. *Educational and Psychological Measurement*, *54*, 837-847.

Thompson, B. y Vacha-Haase, T. (2000). Psychometrics is datametrics: The test is not reliable. *Educational and Psychological Measurement*, *60*, 174-195.

Torgerson, W. S. (1958). *Theory and methods of scaling*. New York: Wiley.

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

(Artículo recibido: 12-11-04; aceptado: 19-9-05)