Psicothema, 1998. Vol. 10, nº 1, pp. 197-208
ISSN 0214 - 9915 CODEN PSOTEG

EFECTOS DE FORMATO DE RESPUESTA Y MÉTODO DE ESTIMACIÓN EN ANÁLISIS FACTORIAL CONFIRMATORIO

José Manuel Tomás y Amparo Oliver
Universitat de València

Se compara el funcionamiento del formato continuo frente al Likert de cinco puntos a partir de la administración de una batería de cuestionarios sobre autoestima a 640 estudiantes de enseñanza secundaria. El trabajo se centra en el comportamiento de los 20 items de la *State Self-Esteem Scale (SSES)* de Heatherton y Polivy (1991). Se someten a estudio diferentes modelos factoriales coherentes con el estudio de la validez de constructo de la escala (modelo de uno y tres factores y modelo multirrasgo-multimétodo) ofreciéndose las estimaciones por máxima verosimilitud y métodos de distribución libre. Tras evaluar el ajuste global y analítico de los modelos, se abordan cuestiones de carácter estadístico tales como aspectos diferenciales de los efectos de método en las escalas según formato de respuesta y método de estimación utilizados.

Response format and method of estimation effects on confirmatory factor analysis. Five-point Likert scale and continuous response format are compared based on results from a 640 high school students sample. This work focuses in the 20 items *State Self Esteem Scale* (SSES) developed by Heatherton & Polivy (1991). Confirmatory factor structures are tested according to the construct validity models (single factor, three factors and multitrait-multimethod model). For these models, maximum likelihood and distribution free methods solutions are provided. After goodness-of-fit assessment, some psychometric issues are addressed: unequal method effects' performance depending on both, the response format and the estimation method being used.

El estudio de la incidencia del formato de respuesta es clásico y de interés central para la medición psicológica y de las ciencias sociales en general. Entre los métodos de escalamiento para la medición de actitudes y características de personalidad, el de más amplio uso es el desarrollado por Likert, sea con formato binario o con mayor número de anclajes. No obstante, lo que este tipo de escalamiento y formato de respuesta intenta apresar es algún constructo psicológico supuestamente continuo. Desde este punto de vista, parece lógico que el aumentar el número de anclajes -alternativas de respuesta- acercaría las variables, los ítems, hacia la verdadera naturaleza de los constructos, mejorando por tanto la medición.

Existe un amplio debate sobre si el aumento en el número de anclajes produce o no una mejora de las características psicométricas de las escalas. Desde la teoría clá-

Correspondencia: José Manuel Tomás
Depart. de Metodología, Psicobiología y Psicología Social
Facultat de Psicología. Universitat de València
46010 Valencia (Spain)
E-mail: tomasjm@uv.es

sica de test y en particular desde el estudio de la fiabilidad hay trabajos empíricos pioneros de Ruch y Charles (1928), Ruch y Stoddard (1925, 1927), Toops (1921), que optan, considerando aspectos de fiabilidad y de practicidad, por los ítems de dos-tres alternativas, frente a los de cuatro o cinco. Mientras, Williams y Ebel (1957) comparan en su trabajo tests de vocabulario con dos, tres y cuatro alternativas de respuesta, concluyendo que las diferencias en fiabilidad no eran estadísticamente significativas en ningún caso. Mediante derivación matemática, suponiendo igual tiempo de respuesta para todos los ítems del test -supuesto poco razonable para diversos autores (Budesco y Nevo, 1985)-, Tversky (1964), Grier (1975) y Lord (1980) proponen tres como el número de anclajes óptimo en términos de fiabilidad. Algunos trabajos empíricos comparan ítems dicotómicos con politómicos (especialmente de siete alternativas, concluyendo que el formato Likert de más anclajes ofrece mejoras sustanciales en la fiabilidad y una estructura factorial más clara y precisa (Comrey y Montag, 1982; King, King y Klockars, 1983; Oswald y Vellicer, 1980).

En una revisión, Morales (1988) concluye que no hay una relación clara entre número de alternativas y fiabilidad; si bien algunos proponen cinco alternativas como un óptimo, y además la mayoría de trabajos encuentran un límite de siete alternativas a partir del cual la fiabilidad no aumenta o incluso disminuye. Por su parte, Sancerni, Meliá y González-Romá (1990) comparan un test con dos formatos de respuesta, dicotómico frente a cuatro anclajes, analizando desde una perspectiva de teoría clásica de tests el efecto sobre la fiabilidad y la validez. Los resultados muestran que respecto a la fiabilidad no hay diferencias, siendo muy similares los coeficientes de validez criterial. Es de destacar que la validez criterial más alta aparecía sistemáticamente cuando se producía congruencia de formato entre la escala y el criterio. En cualquier caso, los coeficientes de validez criterial de uno y otro formato no eran estadísticamente diferentes salvo en unos pocos casos, e incluso en estos, en ocasiones a favor del formato binario y en otras del de cuatro alternativas.

También desde el modelo psicométrico de la teoría de respuesta al ítem se ha estudiado el efecto del número de anclajes sobre diversos parámetros de los ítems, como la discriminación, dificultad, etc. (Gómez, Artés y Deumal, 1989; Bock, 1972), y sobre la función de información (Lord, 1980). Así, este último autor encuentra que aumentar el número de anclajes incrementa la eficiencia del test para sujetos de alta capacidad -alto nivel en el constructo medido-, mientras que la disminuye para los de baja capacidad -bajo nivel en el constructo-. Vale y Weiss (1977) encuentran que se produce un aumento en las funciones de información al aumentar el número de anclajes.

En el límite, si aumentáramos los anclajes podríamos trazar una línea recta donde cada sujeto que responde la prueba marca en función del nivel que presenta él mismo (u otros) en la característica señalada por el ítem. A priori, este tipo de respuesta se aproximaría más al tipo de variables continuas que debieran ser el objeto de los modelos psicométricos lineales en general, y en particular análisis factorial o componentes principales, ampliamente utilizados en estudios de validez y fiabilidad de las medidas. Las datos ajustarían mejor a los modelos matemáticos de muchas técnicas estadísticas y psicométricas usuales y producirían diferencias en las características psicométricas de las variables.

Diversos trabajos han comparado el funcionamiento de variables medidas de forma continua y medidas en formato Likert. Se ha estudiado, por ejemplo, los efectos de escala de medida al incluir variables moderadoras en estudios de regresión (Russell y Bobko, 1992). Los resultados apuntan a que

los efectos moderadores, de interacción, en regresión pueden verse atenuados cuando se utiliza el formato tipo Likert frente al continuo en la variable dependiente. También Rasmussen (1989) revisa los efectos encontrados en estadísticos calculados en escalas medidas ordinalmente y de forma continua, con la conclusión de que cuando se puntúan con cinco anclajes o más las diferencias no son relevantes. Gregoire y Driver (1987) y el propio Rasmussen (1989) atienden al efecto de la escala ordinal en los errores tipo I y II de diversos contrastes paramétricos y no paramétricos, no encontrando distorsiones de relevancia práctica.

De particular importancia por su proximidad al objetivo del presente trabajo son las investigaciones centradas en el efecto del formato Likert y continuo sobre los resultados del análisis factorial y de componentes principales. Bernstein y Eveland (1982) y Gorsuch (1983) encontraron a través de datos simulados que la categorización podía producir en análisis factorial exploratorio un efecto espúreo de multidimensionalidad. Por su parte, Bernstein y Teng (1989) utilizan datos simulados en formato de respuesta continua, dicotómicos y de cuatro anclajes, con diversos puntos de corte para las alternativas, y considerando tres niveles de intercorrelaciones entre los 20 items de la escala simulada, 0.25, 0.5 y 0.75. Estudian el efecto de estas condiciones sobre la fiabilidad, análisis de componentes principales, análisis factorial exploratorio de máxima verosimilitud y análisis factorial confirmatorio. Varias son las conclusiones del trabajo para las condiciones simuladas: 1) la categorización puede producir evidencia falsa de multidimensionalidad; 2) componentes principales y algunos índices de ajuste descriptivo aplicados al análisis confirmatorio, como el índice de Tucker-Lewis, son más sensibles a efectos del formato de respuesta conforme la fiabilidad de la escala disminuye; 3) el estadístico de ji-cuadrado para el ajuste se ve contrariamente más afectado por la categorización si la fiabilidad de la escala aumenta; 4) los efectos sobre la ji-cuadrado son esencialmente independientes del tamaño muestral; 5) el criterio de Kaiser-Guttmann para la extracción de factores se muestra más afectado por la categorización que el 'screetest' de Cattell.

Los resultados de Bernstein y Teng (1989) con datos simulados arrojan luz sobre muchos aspectos relevantes, como que los efectos pueden ser diferentes en función de diversas condiciones y técnicas de reducción de datos aplicadas, e incluso de los diversos índices o estadísticos aplicados. Los datos simulados pueden mimetizar con precisión diversas condiciones psicométricas y estadísticas relevantes. Desgraciadamente no resuelven el problema de cómo los sujetos responden el test. O dicho de otra forma, no analizan las limitaciones inherentes de los sujetos al responder a lo largo de un continuo (Garner, 1960, 1962). Por contra, Ferrando (1995) administra un mismo cuestionario, la subescala de impulsividad del EPI de Eysenck, con escala de respuesta continua y también con escala de respuesta de tipo Likert de cinco puntos. Las respuestas a ambas escalas ajustaban adecuadamente al modelo unifactorial propuesto, si bien en el caso del estadístico χ^2 mostraba mejor ajuste para el formato continuo. Por su parte, las saturaciones factoriales eran muy similares en ambos formatos, pero siempre menores en el formato Likert. Esta disimilitud era mínima y estadísticamente no significativa, con coeficientes de proporcionalidad cercanos a 0.95. Podría decirse, no obstante, que se encontró, aunque mínimo, un efecto de atenuación de la saturación en el formato ordinal frente al continuo.

A la vista de los estudios realizados, el presente trabajo pretende ofrecer un análisis de la incidencia del formato de respuesta y del método de estimación utilizado sobre las

soluciones factoriales confirmatorias para datos reales. Específicamente, se pretende medir el efecto diferencial sobre distintos índices de ajuste global y analítico y sobre la aparición de efectos de método en las escalas.

Método

Muestra. La muestra la componen 640 estudiantes de bachillerato que contestaron una batería de tests en sus aulas, en pases colectivos, a mediados del curso 1994-95, y fuera del período temporal de realización de los exámenes. Son estudiantes de los diferentes cursos de Bachillerato y Curso de Orientación Universitaria. La edad oscila de 14 hasta 20 años, con una media de 15.8 y una desviación típica de 1.32. Un 55.47% son varones y un 43.75% mujeres.

Instrumentos. La escala objeto de estudio es la State Self Esteem Scale (SSES) compuesta por 20 ítems (Heatherton y Polivy, 1991). Se han utilizado dos formatos de respuesta: una escala tipo Likert de cinco puntos, con anclajes desde 'en absoluto' hasta 'totalmente' y una escala continua que se detalla más adelante. Está construida para medir tres factores de autoestima estado: *autoestima social*, *autoestima de desempeño* y *autoestima de apariencia física.* El posible efecto de ubicación dentro del conjunto total de aquellos ítems con formato de respuesta continuo y Likert fue balanceado. En la mitad de los cuestionarios administrados se hallaban los ítems con respuesta continua al principio y tipo Likert al final, y de forma inversa en la otra mitad. En ningún caso, distintos formatos de respuesta de una misma escala se hallaban contiguos. El procedimiento exacto de recogida de las respuestas en escala continua consistió en medir la distancia en milímetros del extremo izquierdo a la marca realizada por los sujetos sobre la línea. Las respuestas pues, oscilaron entre 0 y 65, longitud total de la línea en milímetros.

Además de la escala SSES ya comentada, con dos formatos distintos de respuesta, a los sujetos se les administró otro grupo de tres cuestionarios de medida de variables relacionadas y aspectos demográficos. Una escala de *autoestima de apariencia física* de 6 ítems con escala de respuesta Likert 5 puntos desde 'nunca' a 'siempre'; la escala de autoestima desarrollada por Rosenberg, de 10 ítems con formato Likert 4 puntos desde 'muy en desacuerdo' a 'muy de acuerdo' y el STAI rasgo, 20 ítems con respuesta Likert de 4 puntos desde 'casi nunca' a 'casi siempre'.

Análisis. Para asistir en la decisión sobre los métodos de estimación a emplear, se aporta el coeficiente multivariado de Mardia, que permite conocer el ajuste de los datos a la normalidad multivariada. Los análisis factoriales confirmatorios fueron realizados con el programa EQS 3.0 (Bentler, 1989) sobre 579 casos completos para la escala en formato continuo y 608 en Likert. Los métodos de estimación utilizados fueron máxima verosimilitud (ML) y métodos de distribución libre o métodos arbitrarios (AGLS). El método de máxima verosimilitud asume la normalidad multivariada, mientras que los métodos de distribución libre son válidos para cualquier tipo de distribución de las variables, pero exigen un alto tamaño muestral (Jöreskog y Sörbom, 1988).

En primer lugar, se somete a estudio el modelo de un solo factor por ser el más parsimonioso, y el de tres factores coincidente con los tres aspectos del rasgo recogidos en la escala: *autoestima social* (F1), de *desempeño* (F2) y de *apariencia física* (F3). El siguiente paso será evaluar la tercera de las estructuras que, a juzgar por los resultados de Bagozzi y Heatherton (1994) y Tomás, Oliver y Pastor (1996), se corresponderían con la dimensionalidad de la escala de autoestima que nos ocupa. Se trata de un modelo multirrasgo-multimétodo que define ads-

cripciones de los ítems a los tres factores de rasgo, así como a dos factores de método. Estos últimos están formados por ítems invertidos (F4) apreciable en ítems como 'me siento a disgusto conmigo mismo' o 'me siento inferior a otros', e ítems no invertidos (F5), como ocurre con 'siento confianza en mis capacidades' y 'siento que otros me respetan y admiran', entre otros ítems.

Para la evaluación del ajuste global de los modelos planteados, se consideraron diversos criterios: estadístico χ^2, índice de ajuste comparativo (CFI), valor de la media absoluta de los residuales estandarizados (MRE), el criterio informativo de Akaike (AIC) y el criterio informativo de Akaike corregido (CAIC). El índice de ajuste (FI) y el índice de ajuste ajustado (AFI) -ambos presentes en las soluciones ofrecidas por el programa LISREL- se aportan específicamente para modelos estimados por métodos arbitrarios. Todos estos indicadores de ajuste son de amplio uso, existiendo numerosas referencias donde se trata su interpretación (Oliver y Tomás, 1994). Respecto a algunos menos habituales, valores del MRE por debajo de 0.05 se consideran indicativos de un buen modelo. AIC y CAIC se interpretan de forma comparativa entre modelos, indicando valores bajos un mejor ajuste del modelo a los datos observados. Además del ajuste global, se estudia analíticamente las saturaciones factoriales de cada modelo.

Resultados

Tomando los ítems en su conjunto, puede evaluarse la curtosis multivariada mediante el coeficiente multivariado de Mardia cuyo valor es 91.735, siendo su estimación normalizada 38.125. Estos valores indican que la distribución conjunta de los ítems se aleja respecto a la distribución normal multivariada. En el plano univariado, los ítems que muestran simultáneamente mayor asimetría y curtosis son 18, 19 y 20; aunque de todos los elevados en curtosis, el 7 es el más asimétrico. Los ítems 1, 14 y 16 muestran, en comparación, un comportamiento más acorde a la distribución normal. Desde este punto de vista, y a pesar de que máxima verosimilitud es robusta a desviaciones de la normalidad, el método de estimación que mejor ajustaría a los datos sería el de distribución libre.

Tabla 1
Ajuste global sobre los datos obtenidos para el SSES según formatos de respuesta continuo y Likert. Ambos obtenidos por estimación máximo verosímil (ML) y por métodos de distribución libre (AGLS)

CONTINUO					LIKERT			
ML		AGLS		Criterio	ML		AGLS	
1F	3F	1F	3F	ajuste	1F	3F	1F	3F
1842.81 p<.001	1552.03 p<.001	945.62 p<.001	953.694 p<.001	χ^2 p=	1793.32 p<.001	1524.65 p<.001	1026.87 p<.001	1001.50 p<.001
.067	.088	.111	.120	MRE	.065	.068	.103	.111
1502.81	1218.03	605.62	619.69	AIC	1453.325	1190.65	686.87	667.50
591.39	322.69	-305.80	-275.64	CAIC	533.59	287.15	- 232.85	-.235.99
.613	.679	.399	.390	CFI	.616	.679	.315	.333
		.776	.774	FI			.732	.739
		.723	.716	AFI			.669	.671

En la tabla 1 se ofrece el ajuste global de los primeros dos modelos planteados. La significatividad y especialmente la cuantía del estadístico χ^2 señalan una gran discrepancia entre la variabilidad de los datos reales y la reproducida a partir de los modelos, tanto de uno como de tres factores. Ninguna media absoluta de residuales estandarizados es menor de 0.05, indicando así mal ajuste. A la misma conclusión se llega por el índice de ajuste comparativo (CFI), índice de ajuste (FI) y su versión ajustada (AFI), puesto que con ellos no se alcanza el valor 0.9 para ningún modelo, bajo ningún método ni formato. A pesar de que el ajuste para estos dos modelos es inadecuado, resulta interesante apreciar posibles diferencias halladas en función del formato, frente a dife-

rencias debidas al método de estimación. Los patrones generales son claros. Por todos los indicadores de ajuste las diferencias entre medidas continuas y medidas tipo Likert son mínimas. Esto es cierto tanto para la estimación máximo verosímil como para la de distribución libre. Por ejemplo, la media de residuales estandarizados, una medida absoluta de ajuste, es 0.067 en el modelo de un factor estimado por máxima verosimilitud en escala continua, mientras el mismo modelo con el mismo tipo de estimación pero sobre los datos tipo Likert es 0.065, idénticos desde el punto de vista práctico. Otro índice, el CFI, alcanza el valor 0.613 en el modelo unifactorial y estimación máximo verosímil para escala continua, y un valor muy similar, 0.616, para el mismo modelo en escala Likert. Sin embargo, y pese a que la conclusión sobre ajuste inadecuado es la misma con ambos tipos de estimación, los valores de los índices de ajuste son bastante diferentes entre sí. La media de residuales estandarizados valía, según hemos visto, 0.067 en el modelo de un factor en escala continua y máxima verosimilitud, y el mismo modelo también en escala continua pero estimado mediante métodos de estimación libre alcanza 0.111, casi el doble. Lo mismo ocurre con el resto de índices. El CFI vale 0.613 al estimarlo por máxima verosimilitud en la escala continua y 0.399 al estimarlo por métodos arbitrarios en el mismo tipo de escala. Como puede verse en la tabla 1, los ejemplos planteados son representativos, la tendencia es estable y ocurre igualmente al comparar los dos tipos de estimación, ML y AGLS, en el caso de escala de respuesta tipo Likert. Resumiendo, para los modelos de inadecuado ajuste, el utilizar un formato u otro de escala de respuesta no supone cambios apreciables en el ajuste global por ningún índice, mientras que el cambio de método de estimación sí produce diferencias cuantitativas fuertes, aunque no en la interpretación del ajuste ya que los modelos están todavía muy alejados de los límites en que se suele considerar ajuste aceptable.

Tabla 2
Saturaciones factoriales obtenidas por máxima verosimilitud y métodos de distribución libre en dos estructuras factoriales planteadas para los dos formatos de respuesta

	CONTINUO				LIKERT			
	ML		Arbitrarios		ML		Arbitrarios	
Item	1F	3F	1F	3F	1F	3F	1F	3F
1	.684	.759 F2	.816	.812 F2	.618	.655 F2	.753	.774 F2
2	.262	.394 F1	.205	.250 F1	.302	348 F1	.279	.272 F1
3	.599	.810 F3	.816	.842 F3	.540	.738 F3	.768	.840 F3
4	.564	.538 F2	.575	.498 F2	.584	.583 F2	.681	.598 F2
5	.399	.563 F2	.502	.579 F2	.311	.453 F2	.382	.496 F2
6	.335	.394 F3	.433	.334 F3	.373	.353 F3	.375	290 F3
7	.323	.398 F3	.386	.354 F3	.425	.514 F3	.589	.603 F3
8	.368	.376 F1	.476	.439 F1	.355	.395 F1	.463	471 F1
9	.539	.462 F2	.628	.580 F2	.461	.433 F2	.606	.571 F2
10	.788	.442 F1	.883	.823 F1	.774	.727 F1	.837	.786 F1
11	.816	.657 F3	.858	.785 F3	.760	.672 F3	.828	.807 F3
12	.641	.849 F3	.821	.850 F3	.643	.810 F3	.827	.837 F3
13	.280	.719 F1	.466	.503 F1	.305	.418 F1	.452	.586 F1
14	.552	.701 F2	.720	.805 F2	.481	.595 F2	.591	.784 F2
15	.655	.458 F1	.712	.686 F1	.671	.694 F1	.780	.763 F1
16	.403	.482 F3	.584	.556 F3	.503	.580 F3	.606	.542 F3
17	.244	.701 F1	.476	.591 F1	.288	396 F1	.376	.519 F1
18	.489	.648 F2	.545	.633 F2	.539	.687 F2	.743	.768 F2
19	.530	.617 F2	.597	.561 F2	.507	.597 F2	.574	.574 F2
20	.325	.735 F1	.589	.655 F1	.353	.451 F1	.450	.529 F1

De la misma forma que comparamos los modelos de un factor y de tres factores al respecto del ajuste global, podemos analizar los efectos de método de estimación y escala de medida sobre la saturaciones factoriales atendiendo al ajuste analítico (ver tabla 2). Los resultados en ajuste global se mimetizan al nivel de las saturaciones. Si escogemos como representativo el ítem 1, podemos analizar estas diferencias para el modelo unifactorial. La diferencia entre la saturación por formato Likert y la saturación por continua es 0.06 en el caso de estimación máximo verosímil, y también de 0.06 en el

caso de estimación por métodos de distribución libre. O en términos de porcentaje de varianza explicada, las diferencias entre usar una escala continua frente a una Likert para el ítem 1 suponen una ganancia de 0.36% de varianza explicada. Por otra parte, la diferencia entre las saturaciones por máxima verosimilitud y distribución libre sí son algo mayores. Por ejemplo para la escala continua es de 0.132. Incluso si escogemos el ítem 16, por ser uno de los que más diferencias entre escala continua y Likert manifiestan, los resultados no permiten decantarse por un tipo u otro de formato. Para este ítem, la diferencia entre Likert y continua para máxima verosimilitud es de 0.1, mientras que para métodos de distribución libre es 0.02. Por su parte, y para este ítem, la diferencia entre las saturaciones por máxima verosimilitud y distribución libre en el caso de escala continua es 0.18, mientras en el caso de respuesta Likert es 0.103. Otra vez en términos de porcentaje de varianza explicada, puede decirse que la máxima diferencia entre continua y Likert supone un 1% de mejora en la explicación. Además, esta mejora no es siempre en el sentido de que la escala continua sea la de saturación más alta y por tanto de mayor fiabilidad, si no que a veces es la respuesta tipo Likert la que ofrece una saturación mayor. En resumen, tampoco hay diferencias entre la escala continua y Likert al nivel de las saturaciones, y esta diferencias son en cualquier caso menores que las que se producen en función del método de estimación empleado.

En la tabla 3, se halla un resumen de bondad de ajuste similar al ofrecido para los modelos anteriores, pero esta vez para el modelo que propone tres factores de rasgo junto a dos de método. Este modelo mostró ajuste adecuado en el trabajo de Bagozzi y Heatherton (1994) y en Tomás et al. (1996). Se puede considerar que hay repetida e independiente evidencia empírica al respecto de que es un modelo de ajuste aceptable. Efectivamente, a la luz de los diversos índices de ajuste, puede considerarse que el modelo ajusta razonablemente a los datos, tanto al analizar los datos obtenidos mediante escala continua como los de escala Likert. El ajuste es más deficiente cuando se comparan los métodos de distribución libre con máxima verosimilitud. En lo que respecta a los posibles efectos del cambio de escala de medida sobre los índices de ajuste, son prácticamente inexistentes. Por ejemplo, la media de residuales estandarizados es 0.035 para los datos continuos estimados por máxima verosimilitud, frente a un 0.034 de la escala Likert por el mismo método. Esto es, una diferencia en el error cometido en el modelo de 0.001. Si hacemos la misma comparación para otro índice relevante, el CFI, en este caso la diferencia entre ambas escalas es de 0.04. Estos cambios mínimos no afectan en absoluto a la interpretación del ajuste. Si observamos ahora el efecto que presenta el cambio de método de estimación, la situación es diferente. Así, la media de los residuales estandarizados para los datos continuos al estimar el modelo por máxima verosimilitud vale 0.035, frente a un valor de 0.057 en el caso de estimación por métodos de distribución libre. Este cambio de 0.022 es no sólo mayor que en el caso de cambio de escala continua a Likert, si-

Tabla 3
Ajuste global según métodos de estimación y formatos de respuesta para el modelo multirrasgo-multimétodo

Criterio ajuste	CONTINUA		LIKERT	
	ML	AGLS	ML	AGLS
χ^2 p=	500.73 p<.001	393.66 p<.001	510.39 p<.001	433.71 p<.001
MRE	.035	.057	.034	.060
AIC	208.73	101.66	218.395	141.71
CAIC	-574.02	-681.09	-571.49	-648.17
CFI	.918	.808	.914	.770
FI	–	.907	–	.887
AFI	–	.886	–	.837

no que tiene importantes implicaciones prácticas, ya que 0.05 o menor es el límite aceptado de los residuales estandarizados en la literatura para considerar un modelo como adecuado. Lo mismo ocurre con la diferencia en los CFI, 0.11, del modelo de escala continua estimado por máxima verosimilitud frente al estimado por métodos de distribución libre. Esta diferencia, supone que mientras el modelo se puede evaluar como adecuado al estimarlo por máxima verosimilitud, sin embargo aparece como mejorable (no alcanza 0.9) en la estimación por métodos de distribución libre. En resumen, nuevamente, y esta vez para un modelo de ajuste aceptable y validado anteriormente, el cambio de escala de medida no supone un efecto relevante, mientras que el cambio de estimación, por contra, empeora claramente el ajuste del modelo.

En algunas de las soluciones estandarizadas -como en las estimaciones para formato continuo por ambos métodos- uno de los factores de método (F5) se ha definido negativamente durante la estimación. De esta forma, el proceso iterativo de estimación ha encontrado una mejor solución sin acarrear problemas de cara a la interpretación, ya que la totalidad de las saturaciones en ese factor tienen el mismo signo.

Resulta de interés analizar los posibles efectos del método de estimación y del formato de respuesta separadamente para las saturaciones de los factores de contenido (rasgos) y los de método. Para resumir la información de todas las saturaciones factoriales ofrecidas en la tabla 4, puede calcularse la media de las diferencias en valor absoluto entre pares de saturaciones de interés. En cuanto a las diferencias debidas al formato, la media de diferencias entre las saturaciones estimadas por máxima verosimilitud para la escala continua y la escala Likert en factores de rasgo es de 0.063, mientras para factores de método es 0.051. Estas mismas diferencias entre continua y Likert, pero en estimación por métodos de distribución libre es de 0.057 en saturaciones en los factores de contenido y de 0.0538 en los factores de método. Como puede verse las diferencias en la cuantía de las saturaciones debidas al cambio de escala a través de métodos de estimación es mínima, en torno al 0.05, lo que implica en términos -medios- de porcentaje de varianza explicada un 0.25%. Por su parte, si atendemos a las diferencias medias debidas a los métodos de estimación nos encontramos con resultados muy similares. En los datos de escala continua, la diferencia media en las saturaciones entre la estimación por máxima verosimilitud y por distribución libre es de 0.049 en factores de rasgo y 0.075 en factores de mé-

Tabla 4
Saturaciones según métodos de estimación y formatos de respuesta para el modelo multirrasgo-multimétodo

	CONTINUO				LIKERT			
	ML		AGLS		ML		AGLS	
Item	Rasgo	Método	Rasgo	Método	Rasgo	Método	Rasgo	Método
1	.463 F2	-.595 F5	.380 F2	-.723 F5	.290 F2	.602 F5	274 F2	.748 F5
2	.312 F1	.228 F4	.316 F1	.200 F4	-.357 F1	.272 F4	-.396 F1	.226 F4
3	.587 F3	-.609 F5	.534 F3	-.696 F5	.706 F3	.508 F5	.656 F3	.624 F5
4	.209 F2	.548 F4	.103 F2	.628 F4	.112 F2	.615 F4	.143 F2	.666 F4
5	.577 F2	.289 F4	.489 F2	.391 F4	.534 F2	.260 F4	.592 F2	.308 F4
6	.237 F3	-.318 F5	.142 F3	-.421 F5	.046 F3	.396 F5	.102 F3	.445 F5
7	.329 F3	.316 F4	.315 F3	.378 F4	.363 F3	.393 F4	.391 F3	.508 F4
8	.187 F1	.353 F4	.216 F1	.430 F4	-.219 F1	.321 F4	-.257 F1	.358 F4
9	.132 F2	-.501 F5	.128 F2	-.593 F5	.092 F2	.450 F5	.095 F2	.529 F5
10	.018 F1	.890 F4	.009 F1	.869 F4	-.015 F1	.834 F4	.039 F1	.882 F4
11	.165 F3	-.953 F5	-.157 F3	-.963 F5	-.038 F3	.835 F5	.026 F3	.867 F5
12	.518 F3	-.668 F5	.533 F3	-.727 F5	.490 F3	.646 F5	.486 F3	.715 F5
13	.736 F1	.247 F4	.773 F1	.203 F4	-.780 F1	.211 F4	-.768 F1	.201 F4
14	.579 F2	-.450 F5	.644 F2	-.548 F5	.516 F2	.451 F5	.608 F2	.481 F5
15	.131 F1	.590 F4	.055 F1	.711 F4	-.140 F1	.631 F4	-.061 F1	.785 F4
16	.369 F3	.381 F4	.315 F3	.493 F4	.374 F3	.471 F4	.372 F3	.485 F4
17	.761 F1	.200 F4	.832 F1	.197 F4	-.772 F1	.201 F4	-.785 F1	.218 F4
18	.595 F2	.382 F4	.542 F2	.470 F4	.499 F2	.513 F4	.440 F2	.599 F4
19	.436 F2	.466 F4	.334 F2	.607 F4	.326 F2	.498 F4	.304 F2	.553 F4
20	.688 F1	.282 F4	.720 F1	.333 F4	-.690 F1	.261 F4	-.708 F1	.267 F4

todo. Por su parte, cuando se toman los datos en escala tipo Likert, las diferencias absolutas en las saturaciones estimadas por ambos métodos son de 0.032 en el cao de factores de rasgo y de 0.06 en el caso de factores de método. Por lo tanto, tampoco puede decirse que exista un efecto relevante del método de estimación sobre las saturaciones ni ningún patrón homogéneo de diferencias, salvo que éstas son mínimas.

Para facilitar una apreciación global de las diferencias entre las saturaciones obtenidas para los formatos de respuesta y los métodos de estimación considerados, se ofrecen coeficientes de reproductibilidad. Estos son la razón entre las saturaciones medias para diversas condiciones. Permiten saber en qué proporción son coincidentes las saturaciones de las dos condiciones a comparar. Utilizaremos las saturaciones medias en los factores para evaluar efectos de formato de respuesta y método de estimación. La relación entre las saturaciones en factores de método estimadas para formato Likert y continuo por máxima verosimilitud es 1.011 y por métodos arbitrarios 0.988. Los coeficientes hallados de la razón entre saturaciones por máxima verosimilitud y métodos arbitrarios son 0.895 para formato Likert y 0.875 para formato continuo. En cuanto a los coeficientes hallados para las saturaciones promedio en los factores de rasgo, para formato Likert frente a continuo por máxima verosimilitud es 0.916 y por métodos arbitrarios 0.995. Los coeficientes hallados de la razón entre saturaciones por máxima verosimilitud y distribución libre son: 0.981 para formato Likert; y 1.065 para formato continuo. Tomando globalmente las saturaciones de rasgo y método, los coeficientes de reproductibilidad son: 0.967 en la comparación de formatos en máxima verosimilitud; 0.991 entre formatos para métodos arbitrarios; 0.931 comparando métodos de estimación en formato Likert; y 0.954 en el continuo. Estos coeficientes son muy cercanos a 1, especialmente en la comparación de formatos en métodos de distribución libre. Es la comparación de métodos de estimación en el formato Likert la que ofrece una discrepancia mayor entre los casos comparados, en torno al 7% (coeficiente de reproductibilidad de 0.931) de diferencia entre máxima verosimilitud y métodos arbitrarios.

Discusión

De los tres analizados, el modelo de mejor ajuste a la escala SESS analizada es el multirrasgo-multimétodo. Además de presentarnos una configuración de los ítems adscritos a cada uno de los cinco factores que conforman el modelo, éste nos permite responder a cuestiones sobre el comportamiento de métodos de estimación y formatos de respuesta. Sería esperable teóricamente que una parametrización más alta, situación característica de los modelos multirrasgo-multimétodo analizados mediante factorial confirmatorio, conllevara diferencias en las estimaciones según métodos y formatos, frente a las ofrecidas para modelos más parsimoniosos. En modelos con más parámetros a estimar como los multirrasgo-multimétodo podría generarse una 'situación límite', con los característicos problemas de estimación asociados a su complejidad (Marsh, 1989), entre ellos la aparición de varianzas de error negativas, casos Heywood. Aunque en estas condiciones podría tener una mayor repercusión el incumplimiento o violación de los supuestos, no ha sido así en este caso. No se producen mayores problemas de estimación para escala Likert que para continua. Tampoco se aprecian diferencias, tan siquiera mínimas, en la cuantía de las saturaciones en los factores de método en función del formato de respuesta o del método de estimación. Este mismo patrón ocurre para las saturaciones en factores de rasgo. No ocurre en

estos datos el efecto de atenuación hallado por Ferrando (1995). En algunos casos las saturaciones factoriales en la escala Likert eran mayores que en la continua, y en otros el comportamiento era el contrario. Tampoco se detecta un efecto de multidimensionalidad espúrea (Bernstein y Teng, 1989), ya que por ambos tipos de formato la evaluación del ajuste global de los diferentes modelos es coincidente. El mayor efecto de los estudiados es el ocasionado por cambios en el método de estimación. Si bien las diferencias en saturaciones no son excesivas, un cambio de método para ambos tipos de formato de respuesta puede generar evaluaciones distintas del ajuste datos-modelo.

Ante estos resultados, y en la línea de Ferrando (1995), para soluciones confirmatorias tan similares, no se justifica el uso de formatos más complejos y costosos, como la respuesta continua. Además, conviene resaltar que el número de casos válidos, completos, es ligeramente superior en el caso de formato Likert. Si este efecto se corroborara podría considerarse otra debilidad del formato continuo. No compensa en este ámbito desarrollar medidas tan costosas, puesto que no supone mejoras en la fiabilidad de los ítems.

Teniendo en cuenta estos aspectos y los tratados en los trabajos revisados, se hace necesaria más investigación que ofrezca directrices y facilite la labor de medición en el campo aplicado sin pérdida de rigor. Esta investigación, a nuestro entender, debería integrar datos simulados y empíricos. Aunque los estudios de simulación ayudan a mimetizar gran parte de las características psicométricas que describen a los ítems, no parece ser un acercamiento suficiente ya que en la práctica real los sujetos parecen funcionar de acuerdo a esquemas más simplificadores. Dependiendo de diferentes grados de atención e interés en la tarea, entre otros aspectos, los sujetos pueden tender a limitar su espacio de respuesta, focalizando su atención en grandes anclajes *invisibles.* Parece que, si la escala de respuesta fuese realmente continua, al discretizarla deberían producirse efectos de atenuación en las saturaciones, y esto no ocurre, los sujetos no han aprovechado la continuidad del formato. Se abre el interrogante de cómo operarán los sujetos, supeditados a los procesos psicológicos que rigen su forma de responder, ¿dónde situar el umbral en que la introducción de más anclajes no supone una mejora psicométrica, independientemente del rigor que exige el cumplimiento de los supuestos?

Agradecimiento

Los autores agradecen las sugerencias de la profesora M. D. Sancerni y de dos revisores anónimos, que mejoraron la versión final del trabajo.

Referencias

Bagozzi, R. P. y Heatherton, T. F. (1994). A general approach to representing multifaceted personality constructs: Application to state self-esteem. *Structural Equation Modeling, 1* (1), 35-67.

Bentler, P.M. (1989). *EQS estructural equation program manual*. Los Angeles, BMDP Statistical Software.

Bernstein, I. H. y Eveland, D. (1982). State vs. trait anxiety: A case study in confirmatory factor analysis. *Personality and Individual Differences, 3*, 361-372.

Bernstein, I. H. y Teng, G. (1989). Factoring ítems and factoring scales are different: spurious evidence for multidimensionality due to

item categorization. *Psychological Bulletin, 105*, (3), 467-477.

Bock, R. D. (1972). Estimating item parameters and latent ability when responses are scored in two or more nominal categories. *Psychometrika, 37*, 1, 29-51.

Budesco, D. V. y Nevo, B. (1985). Optimal number of options: an investigation of the assumption of proportionality. *Journal of Educational Measurement, 22, 3*, 183-196.

Comrey, A. L. y Montag, I. (1982). Comparison of factor analytic results with two-choice and seven choice personality item formats. *Applied Psychological Measurement, 6*, 285-289.

Ferrando, P. J. (1995). Equivalencia entre los formatos Likert y continuo en ítems de personalidad: Un estudio empírico. *Psicológica, 16*, (3), 417-428.

Garner, W.R. (1960). Rating scales, discriminability, and information transmission. *Psychological Review, 67*, 343-352.

Garner, W. R. (1962). *Uncertainity and structure as psychological concepts.* New York, Wiley.

Gómez, J., Artés, M. y Deumal, E. (1989). Efecto del número de rangos de la escala de medida sobre la calibración de ítems y sujetos mediante credit. *II Conferencia Española de Biometría,* Segovia, Septiembre.

Gorsuch, R. L. (1983). *Factor analysis* (second edition). Hillsdale, New Jersey, Erlbaum.

Gregoire, T. G. y Driver, B. L. (1987). Analysis of ordinal data to detect population differences. *Psychological Bulletin, 101,* (1), 159-165.

Grier, J. B. (1975). The number of alternatives for optimum test reliability. *Journal of Educational Measurement, 12*, 109-112.

Heatherton, T.F. y Polivy, J. (1991). Development and validation of a scale for measuring state self esteem. *Journal of Personality and Social Psychology, 60*, 895-910.

Jöreskog, K.G. y Sörbom, D. (1988). *LISREL 7: A guide to the program and applications.* Chicago, SPSS.

King, L. A., King, D. W., y Klockars, A. J. (1983). Dichotomous and multipoint scales using bipolar adjectives. *Applied Psychological Measurement, 7,* 2, 173-180.

Lord, F. M. (1980). *Applications of Item Response Theory to practical testing problems.* New Jersey, Lawrence Earlbaum Associates.

Marsh, M.W. (1989). Confirmatory factor analysis of multitrait multimethod data: Many problems and a few solutions. *Applied Psychological Measurement, 13*, 335-361.

Morales, P. (1988). *Medición de actitudes en psicología y educación. Construcción de escalas y problemas metodológicos.* San Sebastián, Ttarttalo.

Oliver, A. y Tomás, J.M. (1994). Indices de ajuste absolutos e incrementales: comportamiento en Análisis Factorial Confirmatorio con muestras pequeñas. *Psicológica, 15*, (3).

Oswald, W. I. y Velicer, W. F. (1980). Item format and the structure of the Eysenck personality inventory: a replication. *Journal of Personality Assessment, 44*, (3), 283-288.

Rasmussen, J. L. (1989). Analysis of Likert-scale data: A reinterpretation of Gregoire and Driver. *Psychological Bulletin, 105* (1), 167-170.

Romero, E., Luengo, M. A. y Otero-López, J. M. (1994). La medición de la autoestima: una revisión. *Psicologemas, 8*, (15), 41-60.

Ruch, G. M. y Charles, J. W. (1928). A comparison of five types of objective tests in elementary psychology. *Journal of Applied Psychology, 12,* 398-404.

Ruch, G. M. y Stoddard, G. D. (1925). Comparative reliabilities of five types of objective examinations. *Journal of Educational Psychology, 16,* 89-103.

Ruch, G. M. y Stoddard, G. D. (1927). *Tests and measurement in high school instruction.* Chicago: World Book.

Russell, C. J. y Bobko, P. (1992). Moderated regression analysis and Likert scales: Too coarse for comfort. *Journal of Applied Psychology, 77,* (3), 336-342.

Sancerni, M. D., Meliá, J. L. y González-Romá, V. (1990). Formato de respuesta, fiablidad y validez en la medición del conflicto de rol. *Psicológica, 11,* (2), 167-175.

Tomás, J. M., Oliver, A. y Pastor, A. (1996). Modelos confirmatorios y efectos de método en la medida de la autoestima. *Boletín de Psicología,* 51, 33-44.

Toops, H. A. (1921). Trade tests in education. *Teachers College Contributions to Education* (nº 115). New York, Columbia University.

Tversky, A. (1964). On the optimal number of alternatives of a a choice point. *Journal of Mathematical Psychology, 1,* 386-391.

Vale, C. D. y Weiss, D. J. (1977). *A comparison of information functions of multiple choice and free parameters vocabulary ítems.* Research Report 77-2. Minneapolis: Psychometric Methods program, Departament of Psychology, University of Minnesotta.

Williams, B. J. y Ebel, R. L. (1957). The effect of varying the number of alternatives per item on multiple-choice vocabulary test ítems. *The Fourteenth Yearbook.* National Council on Measurements Used in Education.

Aceptado el 12 de mayo de 1997