Metodología de Encuestas
Volumen 14, 2012, 5-23
ISSN: 1575-7803

Evaluación de la fiabilidad de una escala para medir la calidad percibida del servicio mediante la aplicación de un modelo de respuesta graduada en el marco de la Teoría de la Respuesta al Ítem

Jose Luis Palacios Gómez
Universidad Autónoma de Madrid

Evaluación de la fiabilidad de una escala para medir la calidad percibida del servicio mediante la aplicación de un modelo de respuesta graduada en el marco de la Teoría de la Respuesta al Ítem

RESUMEN: El propósito de este trabajo es mostrar el procedimiento y los resultados de la evaluación de la fiabilidad de una escala para medir la calidad percibida del servicio mediante el modelo de respuesta graduada (MRG) de Samejima, uno de los más empleados dentro de la perspectiva psicométrica de la Teoría de Respuesta al Ítem cuando los gradientes de respuesta son ordinales o de tipo Likert. Se usa para tal fin un conjunto de datos proveniente de dos encuestas consecutivas realizadas por un servicio municipal socioeducativo para recoger la percepción de la calidad del servicio por sus usuarios. Los resultados permiten observar la capacidad que posee la escala para discriminar a los usuarios por su nivel de satisfacción con el servicio recibido y una mejor comprensión de la misma. Se pone de relieve la ventaja de esta metodología para evaluar la precisión de la medida de la calidad percibida en esta clase de servicios.

PALABRAS CLAVE: Fiabilidad, Modelo de respuesta graduada, Escala, Calidad percibida.

Reliability evaluation of a scale to measure the perceived quality of the service by means of a Graded Response Model in the framework of Item Response Theory

ABSTRACT: The aim of this work is to show the procedure and results of the reliability evaluation of a scale to measure the perceived quality of a public service, by means of the Graded Response Model (GRM) from Samejima, of common use in the psychometric point of view of IRT framework when using ordinal or Likert scales. Data come from two consecutive surveys implemented in a socio-educational agency to capture users' perceived quality of the service. Results permit to observe the scale´s ability to discriminate users by their level of satisfaction with the obtained service and get a better understanding of it. This article also demonstrates the comparative advantage of this methodology to assess the accuracy of the measures of perceived quality in this kind of services.

KEYWORDS: Reliability, Graded response model, Scale, Perceived quality.

Recibido: 16 de enero de 2012.
Primera decisión: 25 de julio de 2012.
Aceptado: 26 de septiembre de 2012.

1. Introducción

Generalmente, cuando se elaboran escalas para medir actitudes con el propósito de incluirlas en cuestionarios de encuestas de calidad percibida del servicio, el marco teórico de referencia es la Teoría Clásica de los Tests (TCT). La TCT explica la puntuación de un sujeto en la escala, es decir, su "puntuación en el test", como una función modelizada como $X = V + \varepsilon$: el nivel de ese sujeto en la actitud medida depende de su puntuación verdadera y de un término de error que se puede estimar (siempre y cuando el test se componga de dos o más ítems, que es la circunstancia habitual). Sencilla, estadísticamente robusta y muy útil y operativa, la TCT adolece, sin embargo, de problemas de muy difícil solución entre los que destacan dos fundamentales: que la medida de una variable es inseparable del instrumento empleado para medirla (el test o escala) y que las propiedades del instrumento de medida están en función de los sujetos a los que se aplica. Estas dos limitaciones tienen importantes consecuencias no solo metodológicas (definición de la variable por el instrumento que la mide) sino prácticas (necesidad de desarrollar o adaptar un instrumento en cada muestra medida), que pueden hacer aconsejable contemplar otro enfoque psicométrico alternativo que las evite. La "teoría del rasgo latente" o Teoría de la Respuesta al Ítem (TRI) surge precisamente con el propósito de solucionar estos problemas de difícil solución que aquejan a la TCT, convirtiéndola en un marco teórico potencialmente útil para mejorar la evaluación de la fiabilidad de la medida del valor otorgado a un servicio por sus clientes o usuarios. Hay que poner de relieve que, a pesar de algunas excepciones (v.g. Martín y Gil, 2006; Ramos, Sanfiel y Oreja, 2006; Santos, 1999; Varela, Rial y García-Cueto, 2003), el enfoque TRI ha sido muy escasamente empleado en diseño y validación de escalas de medida de la calidad percibida del servicio.

El enfoque que propone la TRI supone una serie de ventajas respecto al enfoque clásico. Posiblemente las tres más relevantes son (Hambleton, Swaminathan y Rogers, 1991) que proporciona parámetros invariantes (no dependiendo de las características de la muestra ni de los ítems), que se estima mejor el error de medida (errores en función del nivel de rasgo) y que se puede estudiar la bondad del ajuste entre modelo y datos.

La TRI propone la existencia de un modelo matemático-estadístico que relaciona el rasgo del sujeto (inteligencia, habilidad, etc.) con su probabilidad de "acertar" el ítem o, en modelos *politómicos* (más de dos posibles respuestas por ítem), su probabilidad de escoger una determina alternativa de respuesta. Este modelo está descrito habitualmente mediante una función que recibe el nombre de Curva Característica del Ítem (CCI) o, cuando hay múltiples alternativas de respuesta, Función de Respuesta al Ítem (FRI).

Existen numerosos modelos de aplicación de la TRI cuyas características se describen en la literatura (Baker, 2001; Bock y Moustaki, 2007; López-Pina, 1995; Muñiz, 1997). El modelo que hemos elegido para llevar a cabo la evaluación de la fiabilidad de esta escala para medir la calidad percibida en servicios socioeducativos es el *Modelo de Respuesta Graduada* (MRG) de Samejima (1969, 1997),

especialmente apropiado para nuestros fines, ya que es uno de los más utilizados y mejor estudiados cuando se trata de modelizar el comportamiento de ítems con formato ordinal o tipo Likert (Abad, Ponsoda y Revuelta, 2006; Asún y Zúñiga, 2008), que son precisamente los que se usan en una escala de medida de actitudes como la estudiada.

El modelo de Samejima se deriva del de Thurstone para escalamiento de objetos (Edwards y Thurstone, 1952), aplicando su lógica al escalamiento de personas. De este modo se asume que la reacción subyacente del sujeto *j* al elemento *i* será el valor z_{ij} situado dentro del continuo de acuerdo z_j. Por tanto, la respuesta del sujeto dependerá de la posición relativa de ese valor z_{ij} respecto a una serie de *m*-1 umbrales τ_{jk}. Así, si el valor z_{ij} se encuentra por debajo del umbral τ_{j1} el sujeto escogerá la primera opción de respuesta. Si el valor es mayor que este, pero menor que τ_{j2} escogerá la segunda, y así sucesivamente. El valor z_{ij} del sujeto estará condicionado por su nivel en el rasgo latente y por un componente de error aleatorio: $z_{ij} = \lambda_j \theta_i + \varepsilon$; donde ε es el error de estimación y λ_j es la correlación entre z_{ij} y el rasgo latente del sujeto θ_i. Dada la relación lineal entre θ y *z*, se asume que para sujetos con igual nivel θ la distribución de z [$f(z_j|\theta)$] es la normal con media $\lambda\theta_j$ y desviación típica $\sqrt{1-\lambda_j^2}$. Este último valor es la desviación típica del error de estimación ε y es el mismo para cualquier valor de θ. Utilizando la aproximación logística, el modelo queda expresado como:

$$P^*(x_{ij} \geq k \mid \theta = \theta_i) = \frac{1}{1 + \exp\left[-Da_j(\theta_i - b_{jk-1})\right]}$$

donde D = 1 (métrica logística), $a_j = \frac{\lambda_j}{\sqrt{1-\lambda_j^2}}$ y $b_{jk} = \frac{\tau_{jk}}{\lambda j}$; y asumiendo que $P^*(x_{ij} \geq 1 \mid \theta = \theta_i) = 1$ y $P^*(x_{ij} \geq m+1 \mid \theta = \theta_i) = 0$, ya que lógicamente el sujeto tendrá que escoger alguna de las *m* alternativas presentadas. A partir de estas probabilidades acumuladas podemos establecer la probabilidad de escoger la opción *k* como la diferencia entre la probabilidad de escoger la *k* o una superior y la probabilidad de escoger la opción *k* + 1 o una superior. El parámetro *a* informa sobre la capacidad discriminante del ítem, mientras que el parámetro *b* indica el nivel de rasgo en el que la probabilidad de escoger una alternativa o las superiores es 0,5 (en cierto modo los parámetros *b* son un indicador de la atracción de las alternativas: reflejan el nivel de rasgo que tienen los sujetos que prefieren una alternativa determinada).

2. Metodología

2.1. Muestra

Se han utilizado dos muestras de 310 y 429 sujetos respectivamente, provenientes de sendas encuestas llevadas a cabo, con una diferencia de dos años (en 2009 y en 2011), en un servicio municipal de Educación de Personas Adultas de una ciudad de la corona metropolitana de Madrid (España). La descripción de estas muestras se refleja en la tabla 1.

Tabla 1.
Estadísticos descriptivos de las muestras utilizadas

		Encuesta 2009		Encuesta 2011	
Estadísticos		Sexo	Edad	Sexo	Edad
N	Válidos	294	310	405	428
	NS/NC	16	-	24	1
Media		-	43,73	-	41,21
Moda		Mujer (74,8%)	56,00	Mujer (78,65)	53,00
Desviación típica		-	15,78	-	14,42

2.2. Instrumento

La escala evaluada está compuesta de doce items, con cinco referidos a las instalaciones del servicio (limpieza, accesibilidad, conservación, seguridad y confort), cuatro referidos al personal docente (simpatía, profesionalidad, motivación y comunicación con el usuario) y tres a los trámites para usar el servicio (horarios, rapidez y comodidad), con un gradiente de puntuaciones con rango 1-10, en el que 1 es la valoración mínima del ítem y 10 la valoración máxima. Para responder al ítem se pide al usuario que lo califique con una puntuación dentro de ese rango. La escala se utiliza solamente en los servicios socioeducativos de este municipio madrileño y su calidad métrica ha sido evaluada anteriormente bajo la perspectiva TCT, ofreciendo evidencias de validez y fiabilidad notables (Palacios, 2007).

2.3. Procedimiento

La escala estaba contenida en los cuestionarios administrados a los usuarios del servicio, junto con otras preguntas de carácter sociodemográfico y relativas a la formación recibida (curso, taller, etc.) con fines de clasificación y segmentación analítica de la respuesta. Los cuestionarios fueron entregados a todas las personas asistentes al servicio, con la petición de que los cumplimentasen y los depositaran en un buzón dispuesto al efecto. En condiciones estadísticas convencionales (n.

confianza = 2σ; p = q) y en el supuesto de aleatoriedad, las encuestas arrojaron un error muestral de ±3,81% y ±3,91%, respectivamente.

2.4. Análisis

Debido a que la metodología de TRI exige muestras con tamaño n ≥ 500 sujetos (Muñiz, 1997: 52), se decidió sumar las muestras de ambas encuestas, alcanzándose por tanto una n = 729, suficiente para la aplicación de MRG. Sin embargo, la correcta estimación de parámetros en el ámbito de TRI requiere una representación suficiente de sujetos en todas las alternativas de respuesta a un ítem, normalmente una proporción cercana al 5% como tasa inferior. Como puede verse en la tabla 2, las cuatro primeras alternativas tienen tasas próximas al 0% ó 1% para prácticamente todos los ítems. En estas circunstancias, se aconseja (Abad, Ponsoda y Revuelta, 2006: 85-88) reunir o colapsar las alternativas adyacentes menos elegidas, empezando por los extremos. Para el caso que nos ocupa, esto se refiere únicamente a la parte inferior de la escala, es decir, a los items 1 a 5, pero no a la parte superior, donde las frecuencias son siempre superiores a 0,05. Por consiguiente, las cinco primeras alternativas fueron colapsadas en una sola para satisfacer una distribución adecuada para la mayoría de los ítems. Así, se han recodificado las alternativas de respuesta de tal manera que las opciones 1 a 5 se consideraron una sola (alternativa 1), mientras que las siguientes se renumeraron para seguir esta nueva ordenación : la alternativa 6 pasó a ser la 2, la 7 pasó a ser la 3, la 8 a ser la 4, la 9 a ser la 5 y la 10 a ser la 6. Quedaron así seis alternativas de respuesta para cada ítem, que, por otra parte, es una cantidad de opciones juzgada como ideal para logra el buen funcionamiento del modelo en lo que se refiere a su precisión (Hernández, Muñiz y García, 2000). Debido también a las restricciones de estimación de los modelos TRI, se eliminaron los sujetos con valores perdidos. Como resultado de ello, el tamaño de la muestra quedó fijado en 628 sujetos.

Con objeto de comprobar la unidimensionalidad del constructo "calidad percibida" se llevó a cabo un análisis factorial exploratorio (AFE) y un análisis factorial confirmatorio (AFC). El AFE se ejecutó con el procedimiento de extracción de factores por el método de factorización de ejes principales y el AFC con un modelo de ecuaciones estructurales con estimación de parámetros por el método de máxima verosimilitud con el programa *AMOS.18*. Con el mismo fin de comprobar la unidimensionalidad del constructo, se halló la correlación ítem-total, condición necesaria, aunque no suficiente, para determinarla (Morales, Urosa y Blanco, 2003). Se estimaron posteriormente los parámetros *a* y *b* del MRG, con el programa *Xcalibre 4.1.4.* mediante el procedimiento de "máxima verosimilitud marginal" (Bock y Aitkin, 1981). Una vez calculados los parámetros del modelo, se obtuvieron las Funciones de Respuesta al Ítem, es decir, la representación gráfica de las probabilidades de elegir cada opción de respuesta en función del nivel del rasgo (también denominadas "curvas características operantes"). Respecto a la escala completa, se obtuvieron las funciones de la información del test y del error típico de medida.

Tabla 2.
Distribución de frecuencias relativas de las alternativas según los ítems antes de colapsar.

	Alternativas									
ÍTEMS	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**
Limpieza	0,01	0,00	0,01	0,01	0,08	0,11	0,18	0,26	0,15	0,17
Accesibilidad	0,01	0,01	0,01	0,02	0,09	0,11	0,17	0,24	0,13	0,20
Conservación	0,01	0,01	0,03	0,04	0,14	0,20	0,20	0,21	0,08	0,09
Seguridad	0,02	0,00	0,02	0,03	0,08	0,14	0,20	0,23	0,12	0,17
Confort	0,02	0,01	0,03	0,06	0,13	0,16	0,18	0,18	0,11	0,12
Simpatía	0,01	0,00	0,00	0,00	0,02	0,04	0,10	0,17	0,23	0,43
Profesionalidad	0,00	0,00	0,00	0,00	0,02	0,04	0,08	0,17	0,23	0,44
Motivación	0,01	0,00	0,00	0,01	0,01	0,05	0,10	0,18	0,22	0,42
Comunicación	0,00	0,00	0,01	0,01	0,02	0,05	0,09	0,14	0,21	0,48
Horarios trámites	0,00	0,00	0,00	0,00	0,03	0,07	0,13	0,24	0,20	0,31
Rapidez trámites	0,01	0,00	0,01	0,01	0,03	0,07	0,11	0,19	0,21	0,35
Comodidad trámites	0,00	0,00	0,01	0,01	0,03	0,07	0,12	0,19	0,20	0,37

Finalmente, se ha llevado a cabo una prueba de bondad del ajuste para comprobar si el modelo se ajusta a los datos empíricos. Se ha utilizado también para ello el programa *Xcalibre 4.1.4*, que proporciona un índice de bondad del ajuste basado en χ^2. Pero el uso de χ^2 como estadístico de contraste se ha puesto en ocasiones en tela de juicio, ya que la distribución real de los estadísticos es desconocida (Sueiro y Abad, 2009). Debido a esto último, también hemos utilizado otro procedimiento complementario para comprobar la bondad del ajuste del modelo a los datos: el análisis de residuos. Para ello se divide θ en q intervalos y se calcula para cada una de ellas el residuo estandarizado. Obviamente, cuanto mayor es el residuo (más alejado de cero en términos absolutos) peor será el ajuste del modelo. Como el modelo de TRI que utilizamos es el MRG, en el que cada ítem presenta varias opciones de respuesta, el estudio del ajuste se lleva a cabo sobre las probabilidades teóricas y observadas de cada opción de respuesta, siguiendo un procedimiento adaptado a esta circunstancia (Abad, Olea, Ponsoda y García, 2011: 431-433). Estas probabilidades se representan de forma gráfica, con detalle, además, del intervalo de confianza asociado a la probabilidad observada para cada uno de los niveles de rasgo, que en nuestro caso es del 95%. El programa *MODFIT* que utilizamos al efecto proporciona los gráficos de residuos que permiten juzgar la bondad del ajuste del modelo a los datos opción a opción de cada item. Teniendo en consideración tanto el valor de χ^2 como el del residuo para cada ítem, podemos juzgar si el modelo propuesto se ajusta aceptablemente a los datos empíricos resultantes de la aplicación de la escala evaluada.

Una cuestión que conviene aclarar aquí es la relativa a la comprobación del requisito de unidimensionalidad del rasgo medido en los modelos TRI y su relación con el requisito de la denominada "independencia local" (las puntuaciones de un ítem no dependen de las de los demás). En la práctica es muy frecuente que sólo se someta a comprobación la unidimensionalidad, pues, aunque existen procedimientos para comprobar la independencia local aparte (Abad *et al*., 2011; Lord, 1980; Stout 2002), si el test posee unidimensionalidad ello implica necesariamente independencia local de sus items, ya que "si el supuesto de unidimensionalidad exige que la respuesta del sujeto esté determinada solamente por su nivel de rasgo latente, es evidente que dicha respuesta no podrá estar influenciada por cómo haya contestado los anteriores items (independencia local) o cualesquiera otras variables" (Muñiz *et al*., 2005: 82). Basta, por tanto, con comprobar la unidimensionalidad para satisfacer los requisitos necesarios para aplicar la metodología de TRI, y a ello nos hemos atenido realizando las pruebas al efecto recomendadas en la literatura (Abad *et al*., 2006; Muñiz, Fidalgo, García-Cueto, Martínez y Moreno, 2005; Reckase, 1979).

3. Resultados

El AFE resultó ser pertinente a la luz de los estadísticos calculados al efecto: el test KMO arrojó un valor de 0,877 y la prueba de Barlett resultó estadísticamente significativa (χ^2 = 4531,511; g.l. = 66; p < 0,01). El primer factor extraído explica el 44,62% de la varianza total empírica. El gráfico de sedimentación correspondiente se muestra en el gráfico 1, mostrado claramente un factor dominante. En la matriz factorial sin rotar, todos los items saturan por encima de 0,40, como se muestra en la tabla 3, siendo este un indicador considerado apropiado para determinar si un constructo es unidimensional (García, Gil y Rodríguez, 2000; Morales, 2000). El AFC proporcionó, entre otros, los estadísticos RMSEA (raíz media cuadrática del error de aproximación) y TLI (índice de Tucker-Lewis) con valores de 0,039 y 0,973 respectivamente, para un modelo de tres factores (formados por los items relativos a las instalaciones, el personal y los trámites), y de 0,078, y 0,916, respectivamente, para el modelo unifactorial (tabla4), indicativos de un mejor ajuste para el modelo multifactorial, pero no tan débil para el modelo unifactorial que permita descartar la hipótesis de unidimensionalidad de este último, pues valores de RMSEA entre 0,05 y 0,08 y de TLI iguales o superiores a 0,90 se consideran indicativos de un ajuste aceptable modelo/datos (Hair, Anderson, Tatham y Black, 2001; Levy y Varela, 2006).

Finalmente, la fiabilidad de la escala es alta, arrojando un *alpha* de Cronbach con un valor de 0,879; además, los coeficientes de correlación ítem-total corregidos presentan casi siempre un valor superior a 0,500 (solamente la correlación ítem-total corregida relativa al ítem Limpieza desciende a 0,452). En definitiva, todo parece indicar que efectivamente la escala puede definirse como unidimensional, lo cual la hace adecuada para evaluarla con un modelo de medida basado en la TRI.

Gráfico 1.
Gráfico de sedimentación del AFE

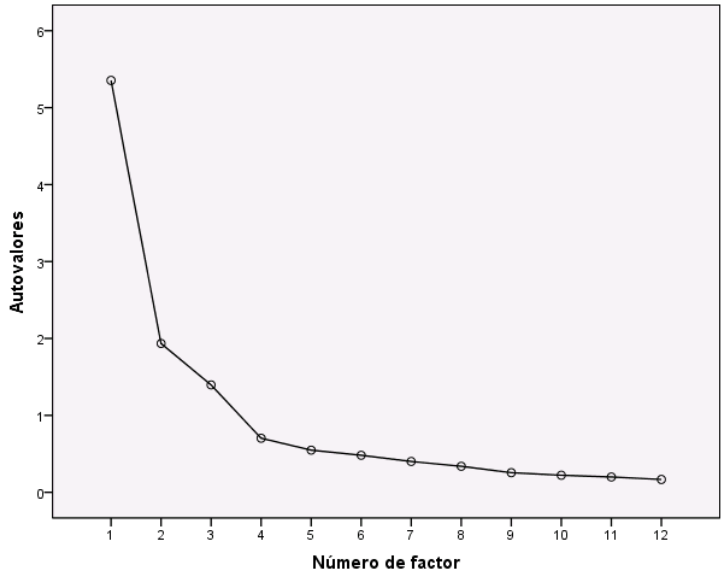


Tabla 3.
Matriz factorial del AFE sin rotar

ITEMS	Factor 1	2	3
Limpieza	0,438	0,300	0,124
Accesibilidad	0,545	0,357	0,181
Conservación	0,565	0,542	0,274
Seguridad	0,583	0,360	0,190
Confort	0,494	0,430	0,261
Simpatía	0,750	-0,380	0,157
Profesionalidad	0,805	-0,422	0,142
Motivación	0,735	-0,434	0,179
Comunicación	0,750	-0,446	0,130
Horarios	0,665	0,074	-0,399
Rapidez	0,700	0,128	-0,572
Comodidad	0,633	0,135	-0,543

Tabla 4.
Estadísticos de ajuste del AFC en dos modelos factoriales alternativos

Estadísticos de ajuste	Modelo	
	Tres factores	Un factor
CMIN/DF	1,956	2,119
AGFI	0,925	0,864
TLI	0,973	0,916
CFI	0,979	0,906
RMSEA	0,039	0,078

Tabla 5.
Parámetros estimados de los ítems

	Parámetros					
ITEMS	**b1**	**b2**	**b3**	**b4**	**b5**	**a**
Limpieza	-2,36	-1,43	-0,42	0,95	2,07	0,93
Accesibilidad	-1,87	-1,19	-0,38	0,74	1,56	1,13
Conservación	-1,52	-0,46	0,53	1,96	2,95	0,95
Seguridad	-1,90	-1,02	-0,13	0,93	1,67	1,17
Confort	-1,43	-0,48	0,42	1,54	2,61	0,90
Simpatía	-1,96	-1,55	-1,03	-0,49	0,15	3,96
Profesionalidad	-1,93	-1,49	-1,03	-0,49	0,11	4,61
Motivación	-1,89	-1,45	-0,99	-0,42	0,19	3,75
Comunicación	-1,87	-1,47	-1,01	-0,55	0,05	3,84
Horarios trámites	-2,51	-1,72	-0,98	-0,05	0,74	1,57
Rapidez trámites	-2,33	-1,61	-1,00	-0,24	0,57	1,53
Comodidad trámites	-2,47	-1,78	-1,13	-0,32	0,51	1,39

Respecto a los parámetros *b*, de localización, todos los ítems mostraron un comportamiento adecuado, con distancias suficientes entre los valores de *b* de las diferentes alternativas u opciones y situados en límites cercanos al intervalo entre -2 y 2 (tabla 5). Concretamente, se observa que para los ítems *Limpieza*, *Accesibilidad*, *Conservación*, *Seguridad* y *Confort*, los parámetros *b* se distribuyen a lo largo de todo el continuo del rasgo (entre -2 y 2) de manera bastante simétrica y bien distanciados entre sí. Esto significa que para esos ítems la elección de alternativas altas o bajas (puntuaciones altas o bajas en el gradiente de los ítems de la escala, que ofrece, una vez transformado, un rango de 1 a 6) tiene una correspondencia bastante precisa con niveles de rasgo bajos o altos. Dicho de otro modo, es necesario un nivel de rasgo alto para dar respuestas elevadas (puntuaciones altas) en estos ítems. Para

los demás ítems, sin embargo, los parámetros *b* se concentran en la parte baja y media del nivel de rasgo (entre -2 y 0,5). Por tanto, para dichos ítems, un nivel medio de rasgo es suficiente para dar una respuesta alta en el ítem.

Respecto al parámetro *a*, de discriminación, los ítems *Simpatía*, *Profesionalidad*, *Motivación* y *Comunicación* obtuvieron los valores más elevados (3,96, 4,61, 3,75 y 3,84, respectivamente), siendo por tanto ítems con gran poder discriminativo. Esto significa que un nivel de rasgo distinto produce respuestas también muy diferenciadas en estos ítems. Los ítems *Horarios*, *Rapidez* y *Comodidad* de los trámites obtuvieron valores de discriminación medios (1,57, 1,53 y 1,39, respectivamente). Los items *Limpieza*, *Accesibilidad*, *Conservación*, *Seguridad* y *Confort* de las instalaciones presentaron en cambio niveles de discriminación bajos (0,93, 1,13, 0,95, 1,17 y 0,90, respectivamente), lo cual significa que sujetos con niveles de rasgo distintos podrían dar respuestas parecidas a estos ítems o, dicho de otro modo, que un sujeto con un nivel de rasgo determinado puede dar respuestas diferentes a estos ítems.

Los gráficos 2 y 3 muestran las Funciones de Respuesta al Ítem. En cada uno de los gráficos, la probabilidad de elegir la alternativa está reflejada en el eje de ordenadas, mientras que el nivel del rasgo aparece en el eje de abscisas. En general, en todos los ítems la alternativa más baja (alternativa u opción 1) es más elegida conforme menor es el nivel de rasgo del sujeto y la probabilidad de elegir dicha alternativa decrece conforme aumenta el nivel de rasgo. Exactamente lo contrario ocurre con la alternativa más alta (alternativa u opción 6): la probabilidad de escogerla crece conforme el nivel de rasgo aumenta. Las alternativas intermedias tienen un punto óptimo en el nivel de rasgo en el que la probabilidad de escoger esa alternativa es máxima. Dicha probabilidad disminuye según nos alejamos de ese punto óptimo por un lado u otro.

En el gráfico 4 se representan gráficamente la Función de Información del Test y su inversa, el error típico de medida. En las dos curvas se observa en definitiva el mismo resultado: se obtienen resultados de medida óptimos cuando el nivel de rasgo de los sujetos está entre -2 y 0. En los niveles aún más bajos (y estadísticamente muy improbables), la precisión de la medida desciende paulatinamente. Lo mismo sucede conforme el nivel de rasgo es superior al promedio (que está representado por el valor 0).

En cuanto a la bondad del ajuste del modelo a los datos desde un punto de vista estadístico, los resultados son los mostrados en la tabla 6. Vemos que hay cuatro ítems (*Conservación* de las instalaciones y *Simpatía*, *Profesionalidad* y *Motivación* del personal) que arrojan valores de χ^2 estadísticamente significativos ($p < 0,05$) y que por tanto no permiten concluir que en estos casos el modelo ajusta suficientemente bien a los datos (rechazamos la hipótesis nula de igualdad entre modelo y datos). El conjunto del test arroja igualmente un valor de χ^2 estadísticamente significativo, indicativo por consiguiente de que el modelo utilizado no se ajusta bien a los datos provenientes de la aplicación de la escala estudiada.

Gráfico 2.
Funciones de Respuestas al Ítem de los seis primeros ítems del test

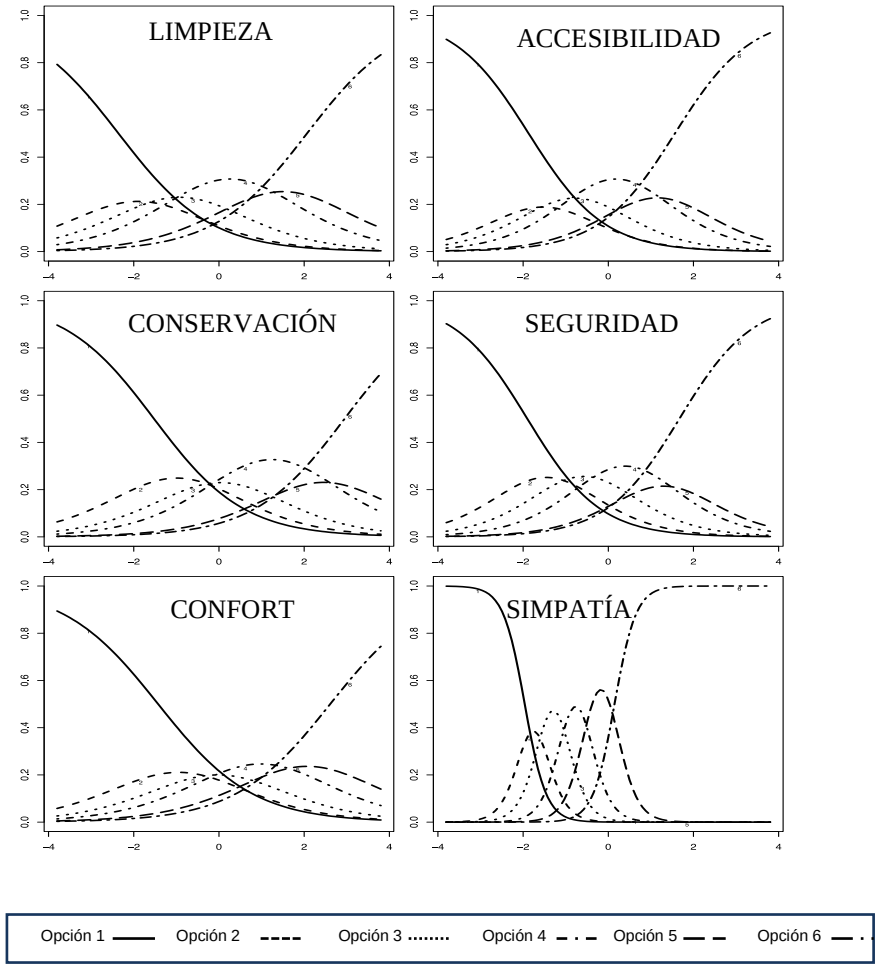
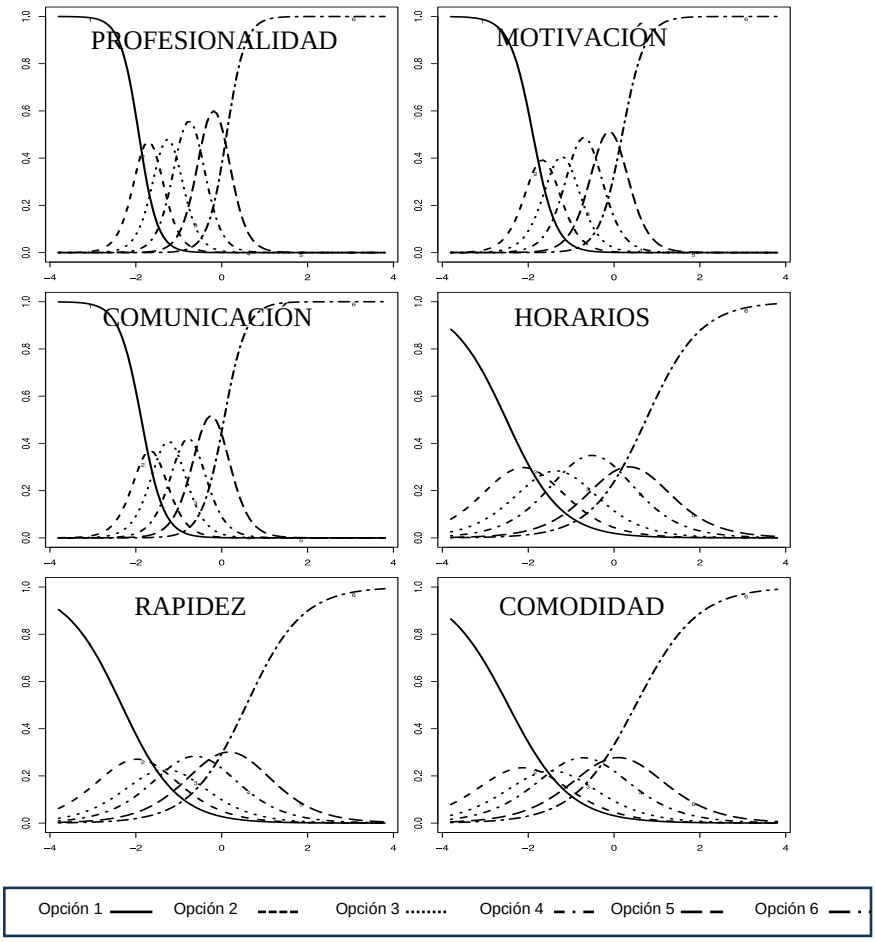


Gráfico 3.
Funciones de Respuestas al Ítem de los seis primeros ítems del test



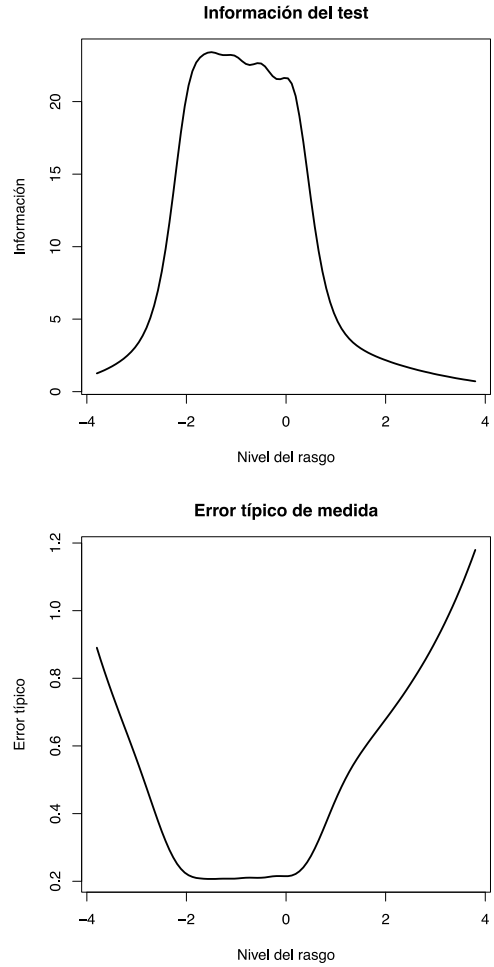
Por su parte, el análisis gráfico de residuos entre las curvas que representan las probabilidades observadas y teóricas de elegir las distintas alternativas u opciones de respuesta (1, 2, …, 6) de los ítems que componen la escala muestran unos resultados no del todo concordantes con los resultados del estudio de la bondad del ajuste basado en χ^2. En el caso del ítem *Conservación de las instalaciones* observamos que efectivamente se producen importantes desajustes entre valores teóricos y empíricos en las opciones 1, 4 y 6, y en esta última opción sobre todo en los niveles de rasgo más altos (la observación de estos gráficos permite determinar, por tanto, no solo qué opciones del ítem muestran un mejor o peor ajuste, sino también establecer en qué niveles de rasgo el modelo predice mejor o peor los resultados de la aplicación empírica del test). De manera que comprobamos que en el caso de este ítem

efectivamente hay un pobre ajuste entre probabilidades empíricas y las probabilidades teóricas o predichas por el modelo: vemos, pues, que existe coincidencia entre los análisis estadístico y gráfico de la bondad del ajuste para este ítem. Sin embargo, en los gráficos correspondientes a las distintas opciones de los ítems *Simpatía*, *Profesionalidad* y *Motivación* del personal, observamos que no hay residuos muy notables entre las probabilidades observadas y las teóricas. Reproducimos aquí solamente, por razones de espacio, los gráficos de contraste entre ambas probabilidades del ítem *Profesionalidad* del personal (gráfico 5), donde vemos que la comprobación del ajuste por el método de los gráficos de residuos contradice en cierta medida los resultados obtenidos con la prueba de χ^2.

Tabla 6.
Bondad del ajuste del modelo MRG a los datos

ITEMS	χ^2	g.l.	p
Limpieza	37,706	36	0,391
Accesibilidad	31,027	35	0,660
Conservación	76,281	32	0,000
Seguridad	33,327	35	0,549
Confort	43,292	35	0,159
Simpatía	39,713	19	0,004
Profesionalidad	48,642	19	0,000
Motivación	44,507	20	0,001
Comunicación	32,008	21	0,058
Horarios trámites	32,032	27	0,231
Rapidez trámites	30,791	29	0,375
Comodidad trámites	28,214	30	0,559
Total escala	477,540	338	0,000

Gráfico 4.
Funciones de información y de error típico de medida del test completo



4. Discusión y conclusiones

Las características psicométricas del modelo MRG le hacen especialmente adecuado para utilizarse en escalas de actitudes en las cuales los items tengan varias opciones de respuesta. Una escala como la estudiada aquí, diseñada para medir la calidad percibida del servicio con un formato sencillo para sus usuarios, presenta doce items con diez posibles opciones de respuesta cada uno (una calificación evaluativa numérica escasamente polisémica) y se presta bien a ser evaluada con esta metodología. El hecho de haber colapsado las cinco primeras opciones en una

sola no resta potencia analítica al modelo, debido a que la propiedad de aditividad del modelo absorbe desde el punto de vista probabilístico la transformación realizada.

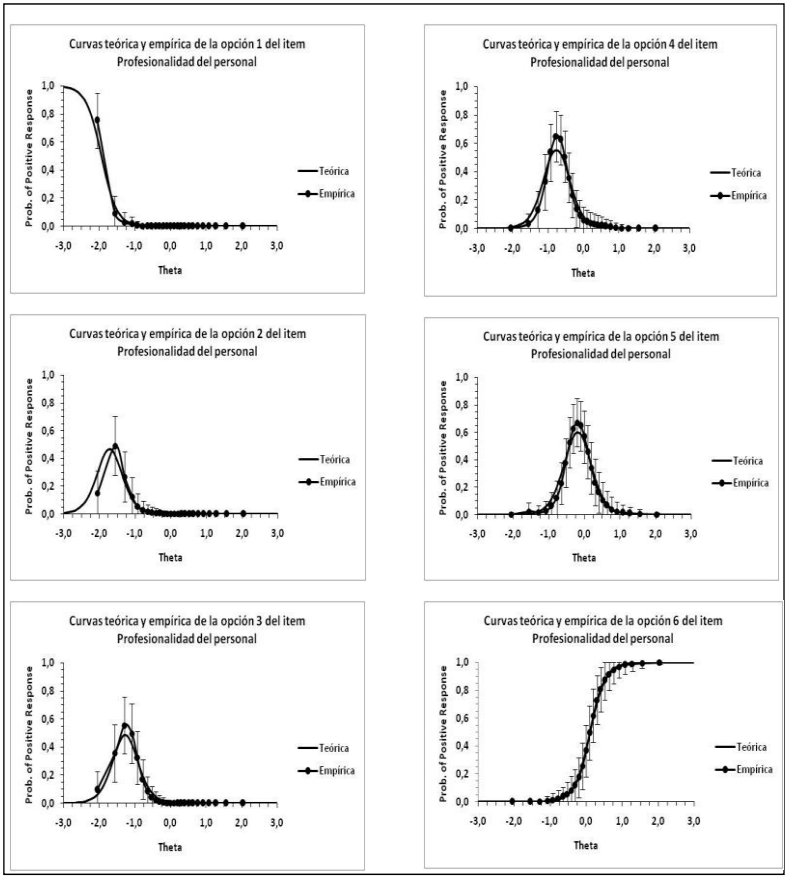
El comportamiento de los parámetros de discriminación y localización, *a* y *b*, resulta muy clarificador de las propiedades métricas de la escala, mostrando el primero qué ítems discriminan mejor en función del nivel de rasgo del usuario (calidad percibida) y mostrando el segundo la menor o mayor correspondencia entre su nivel de rasgo y la puntuación escalar y de cada ítem. Esto supone una indudable ventaja con relación a la medida que proporciona un enfoque clásico TCT, que en definitiva solo aporta una puntuación global de la escala y puntuaciones parciales para cada ítem, sin vincular con precisión estos datos con el nivel de rasgo del sujeto. En nuestro caso, el uso de la metodología del MRG pone de manifiesto que la escala evaluada ofrece una estimación muy adecuada del nivel de rasgo de los sujetos con bajo nivel de rasgo, mientras que la medida se vuelve poco precisa conforme el nivel de rasgo de los sujetos crece. La Función de Información del Test proporciona una visión muy clarificadora del poder discriminatorio de la escala para clasificar a los sujetos a los que se aplica en virtud de su nivel de rasgo, pero además los parámetros del modelo dan noticia muy concisa de qué items de la escala tienen mayor capacidad discriminatoria y para qué niveles de rasgo del sujeto los distintos items de la misma resultan más discriminantes. Los resultados mostrados apuntan a que la escala detectará de manera adecuada cuándo un sujeto se encuentra insatisfecho con el servicio (es decir, cuando su calidad percibida es baja), pero no tanto cuándo se encuentra satisfecho con el mismo. Es decir, la escala es más adecuada para detectar baja calidad percibida que alta calidad percibida. La escala no parece muy adecuada para discriminar entre dos sujetos con un nivel de rasgo medio-alto y alto cuál lo tiene superior, pero detectará con precisión a los sujetos que tengan niveles de rasgo por debajo de la media. Esta información no se puede recabar, o no con tanta precisión, con la óptica metodológica de TCT.

La comprobación de la bondad del ajuste del modelo arrojaba, como hemos visto, algún resultado contradictorio entre el método estadístico y el método gráfico. Los contrastes basados en la prueba de χ^2 indicaban mal ajuste de los items 3, 6, 7 y 8 pero los basados en los residuos gráficos limitaban el mal ajuste al ítem 3 (*Conservación de las instalaciones*). Teniendo en cuenta los problemas señalados en la literatura para esta última prueba, como son una gran sensibilidad de χ^2 a los tamaños muestrales, sobre todo, pero también incertidumbre sobre la distribución de contraste (Sueiro y Abad, 2009), no parece injustificado dar mayor prevalencia al método gráfico para valorar la bondad del ajuste del modelo. Por consiguiente, puede concluirse que la mayor parte de los ítems de la escala evaluada presentan una bondad del ajuste satisfactoria y que, por consiguiente, la escala en su conjunto es capaz de predecir la puntuación empírica del usuario del servicio a la calidad percibida del mismo con una razonable precisión.

En definitiva, puede afirmarse que la utilización del *Modelo de Respuesta Graduada* para la evaluación de la fiabilidad de escalas de calidad percibida del servicio ofrece considerables ventajas psicométricas en relación con las técnicas propias de la TCT para medir la fiabilidad como los test de consistencia interna, que

resultan sensibles a la longitud de la escala (Morales, 2000: Morales *et al*., 2003), o los test de estabilidad de las medidas, que precisan de observaciones con los mismos sujetos en dos tiempos distintos (Martínez, Hernández y Hernández, 2006) muy difícilmente conseguibles en condiciones no experimentales. El uso de este modelo, por tanto, parece una opción a tener muy en cuenta para comprobar la fiabilidad de escalas como la aquí manejada y de las escalas de actitudes en general.

Gráfico 5.
Curvas teóricas y empíricas de las opciones del item Profesionalidad



Referencias

Abad, F.J.; Ponsoda, V. y Revuelta, J. (2006). *Modelos politómicos de respuesta al ítem*. Madrid: La Muralla.

Abad. F.J.; Olea, J.; Ponsoda, V. y García, C. (2011). *Medición en ciencias sociales y de la salud*. Madrid: Síntesis.

Asún, R. y Zúñiga, C. (2008). Ventajas de los modelos politómicos de Teoría de Respuesta al Ítem en la medición de actitudes sociales. Un estudio de caso. *Psykhe*, 17 (2) 103-115.

Baker. F. B. (2001). *The basics of item response theory*. Maryland: ERIC Clearinghouse on Assessment and Evaluation.

Bock, R. D.y Moustaki, I. (2007). Item Response Theory in a general framework en Rao, C. R. y Sindahari, S. (eds.). *Handbook of Statistics*, North Holland: Elsevier, 469-514.

Bock, R.D. y Aitkin, M. (1981). Marginal maximum likehood estimation of item parameters: application of an EM algorithm. *Psychometrika*, 46, 443-459.

Edwards, A. L. y Thurstone, L. L. (1952). An internal consistency check for scale values determined by the method of successive intervals. *Psychometrika*, 17, 169-180.

García Jiménez, E.; Gil Flores y Rodríguez Gómez, G. (2000). *Análisis factorial*. Madrid: La Muralla-Hespérides.

Hair, J.F.; Anderson, R.E.; Tatham, R.L. y Black, W.C. (2001). *Análisis multivariante*. Madrid: Prentice Hall.

Hambleton, R.K.; Swaminathan, H.y Rogers, H.J. (1991). *Principles and applications of item response theory*. Beverly Hills (Cal.): Sage.

Hernández, A., Muñiz, J. y García-Cueto, E. (2000). Comportamiento del modelo de respuesta graduada en función del número de categorías de la escala. *Psicothema*, 12 (Suplemento 2), 288-291.

Lévy, J.P. y Varela, J. (2006). *Modelización con estructuras de covarianzas en ciencias sociales*, s.l.: Gesbiblo.

López Pina, J.P. (1995). *Teoría de la respuesta al ítem: fundamentos*. Barcelona: Promociones y Publicaciones Universitarias.

Lord, F.M. (1980). *Applications of item response theory to practical testing problems*. Hillsdale (NJ): Lawrence Erlbaum Associates.

Martín, D. y Gil, E. (2006). La calidad percibida por el turista en un destino de litoral. Medida y análisis mediante el modelo de Rasch. En Oreja, R. y Febles, J. (coords.). *Modelos de Rasch en Administración de Empresas*. Santa Cruz de Tenerife: Fundación FYDE-Caja Canarias 122-133.

Martínez Arias, M.R.; Hernández Lloreda, M.V. y Hernández Lloreda, M.J. (2006). *Psicometría*. Madrid: Alianza.

Morales, P. (2000). *Medición de actitudes en psicología y educación*. Madrid: Universidad Pontificia de Comillas.

Morales, P.; Urosa, B. y Blanco, A.(2003): *Construcción de escalas de actitudes tipo Likert*. Madrid: La Muralla-Hespérides.

Muñiz J. (1997). *Introducción a la Teoría de Respuesta a los Items*. Madrid: Pirámide.

Muñiz, J.; Fidalgo, A.M.; García-Cueto, E.; Martínez, R. y Moreno, R. (2005). *Análisis de los ítems*. Madrid: La Muralla.

Palacios, J.L. (2007). Diseño y validación empírica de una escala para medir la calidad percibida del servicio en instituciones socioeducativas. *Dirección y Organización*, 33,74-83

Ramos, A.M.; Sanfiel, M.A. y Oreja, J.R. (2006). Medida de la calidad percibida del servicio turístico por medio del Modelo de Rasch: el caso del norte de Tenerife. En Oreja, R. y Febles, J. (coords.). *Modelos de Rasch en Administración de Empresas*. Santa Cruz de Tenerife: Fundación FYDE-Caja Canarias, 167-180.

Reckase M.D. (1979). Unifactor latent trait models applied to multi-factor tests: resultant implications. *Journal of Educational Statistics*, 4, 207-230.

Samejima, F. (1969). Estimation of ability using a response pattern of graded scored. *Psychometrika*, 17 (monograph).

Samejima, F. (1997). Graded Response Model. En Van Der Linden, W.J. y Hambleton, R.K. *A handbook of modern item response theory*. New York: Springer, 85-100.

Santos, J.L. (1999). *La satisfacción del turista en el destino Marbella. Medida y análisis mediante el modelo Rasch*. Málaga: Centro de Ediciones de la Diputación de Málaga.

Stout, W. (2002). Psychometrics. From practice to theory and back. *Psychometrika*, 67, 485-518.

Sueiro, M.J. y Abad, J.F. (2009). Bondad de ajuste en ítems politómicos: tasas de error tipo I y potencia de tres índices de ajuste. *Psicothema*, 21 (4) 639-645.

Thissen, D. y Steinberg, L. (1986). A taxonomy of item response models. *Psychometrika*, 51, 567-577.

Varela, J.; Rial, A. y García-Cueto, E. (2003). Presentación de una escala de satisfacción con los servicios sanitarios de atención primaria. *Psicothema*, 15 (4) 656-661.