Psicothema
2001. Vol. 13, nº 2, pp. 324-329

ISSN 0214 - 9915 CODEN PSOTEG

SOFTWARE, INSTRUMENTACIÓN Y METODOLOGÍA

Relaciones empíricas entre los estadísticos de la teoría clásica de los tests y los de la teoría de respuesta a los ítems

Mª Isabel Barbero García, Pedro Prieto Marañón*, Juan Carlos Suárez Falcón y Concepción San Luis Costas
Universidad Nacional de Educación a Distancia y * Universidad de La Laguna

En este trabajo se analiza empíricamente las relaciones entre las estimaciones de los parámetros de los ítems de la Teoría de la Respuesta al Item (TRI) y de la Teoría Clásica de los Tests (TCT). Asimismo, se analiza el grado de invarianza que presentan los parámetros de ambos marcos teóricos. Para ello se llevó a cabo un estudio de simulación Monte Carlo en el que se generaron las muestras a partir del modelo logístico de tres parámetros. A partir de estas muestras se formaron tres submuestras en función del nivel de aptitud: baja, media y alta aptitud, respectivamente. A continuación se estimaron los parámetros de los ítems en cada submuestra y en la muestra total con los modelos de uno, dos y tres parámetros de la TRI, así como con el modelo de la TCT. Los resultados mostraron, en general, correlaciones muy altas entre las estimaciones de la TRI y de la TCT en el parámetro de habilidad de los sujetos y la dificultad de los ítems, especialmente en los grupos aptitud media. Correlaciones más bajas se encontraron entre las estimaciones de la discriminación de los ítems en todos los grupos. Por último, en relación al segundo objetivo de este trabajo, los resultados no son todo lo buenos que cabría esperar salvo en lo referente al índice de dificultad.

Empirical relationhips between classical test theory and item response theory statistics. The purpose of this paper was examine empirically: (1) the relationships between the item and person statistics from Item Response Theory (IRT) and Classical Test Theory (CTT) and, (2) the invariance of parameters in both measurement frameworks. In order to achieve these objectives, a simulation study was carry out where 30 samples with 500 subjects and 50 items were generated from the 3-PL model; 10 of these samples with low level of ability, 10 with medium and 10 with high level of ability. The items were calibrated by the 1-, 2-, and 3-PL models. Likewise, the person and item statistics in CTT was calculated in each sample. The findings indicate that the relationships between the statistics from both, IRT and CTT, were very strongs, with very high correlation indices, except in the discrimination parameters. The degree of invariance of parameters across the samples wasn't very high in both measurement frameworks.

Uno de los principales problemas a los que tiene que enfrentarse la Teoría de los Tests es el relativo a los errores de medida implícitos en cualquier proceso de medición; por eso, desde los primeros trabajos de Spearman (1904, 1907, 1913), que representaron el nacimiento formal de la Teoría Clásica de los Tests (TCT), hasta nuestros días, han sido muchos los esfuerzos y las investigaciones realizadas, tanto en otros países como en el nuestro, centradas en el desarrollo de nuevos modelos y técnicas que permitieran la detección y estimación de dichos errores. Prueba del esfuerzo realizado en nuestro país en los últimos años son algunos de los trabajos publicados en el número especial de metodología de la revista Psicothema (Vol. 12, suple. Nº 2, 2000).

Algunos de los modelos que fueron surgiendo no eran más que extensiones del modelo lineal de Spearman asumido en la TCT (por ejemplo la Teoría de la Generalizabilidad); sin embargo, a finales de los años sesenta, y sobre todo en la década de los ochenta, surge con fuerza un nuevo marco teórico dentro del campo de la teoría de los tests, la Teoría de Respuesta al Ítem (TRI) que permite encuadrar los modelos psicométricos dentro de modelos probabilísticos de carácter más general (Mellenberg,1994;Van der Linden y Hambleton, 1997) y permitirá solucionar algunos de los

Correspondencia: Mª Isabel Barbero García
Facultad de Psicología
Universidad Nacional de Educación a Distancia
28040 Madrid (Spain)
E-mail: mbarbero@psi.uned.es

problemas de la medición de variables psicológicas que quedaban pendientes dentro del marco de la TCT: a) la dependencia de las puntuaciones de los sujetos del instrumento utilizado para su obtención, y b) la dependencia de las propiedades del instrumento de medida de la muestra de sujetos a los que se aplica.

En teoría, por lo tanto, los modelos de la TRI proporcionan estadísticos de los ítems independientes de la muestra de sujetos utilizada y estadísticos de los sujetos independientes del conjunto de ítems que se les ha administrado (Hambleton y Swaminathan, 1985; Hambleton, Swaminathan y Rogers, 1991; Muñiz, 1997).

La propiedad de invarianza es, pues, el punto fuerte de la TRI y la que permite abordar algunos problemas de medida difíciles de afrontar desde el marco de la TCT como son, por ejemplo, la equiparación de los tests y los problemas relacionados con los tests adaptativos informatizados (TAI´s). No obstante, tal y como señala Fan (1998), si no se verificara esta propiedad de invarianza, o si los resultados obtenidos desde el marco teórico de la TRI y desde el de la TCT no se diferenciaran sustancialmente, sería difícil justificar la utilización de algunos de los modelos de la TRI debido a su mayor complejidad.

Como ya señaló Lord en 1980, los estadísticos obtenidos desde ambos enfoques (TCT y TRI) se relacionan monotónicamente; no obstante, hay pocas investigaciones empíricas que ofrezcan resultados claros acerca de la naturaleza y cuantía de esta relación. Por otra parte, en algunas de las investigaciones empíricas, en las que se comparan los resultados obtenidos mediante los modelos de la TRI con los obtenidos desde la TCT, se concluye que los primeros superan a la TCT en el plano teórico, pero que si se utiliza la TCT para el análisis de ítems y para la estimación del nivel de aptitud de los sujetos, los resultados aportados por ambas aproximaciones son bastante similares (Fan, 1998; Gil, Suárez y Martínez-Arias, 1999; Lawson, 1991; Stage, 1998 a y b, 1999).

En muchos de los trabajos empíricos se han utilizado muestras de tamaño muy grande. Fan (1998) utilizó una base de datos de más de 193.000 sujetos; Gil, Suárez y Martínez Árias (1999) utilizaron una muestra de 8.230 niños. No obstante, dado que en la práctica, la posibilidad de utilizar muestras de este tamaño es bastante difícil, se considera necesario constatar si con tamaños muestrales más pequeños se siguen verificando los supuestos de los modelos.

Con el fin de aportar algunos datos empíricos que puedan contribuir a esclarecer el problema, se ha diseñado un estudio de simulación con dos objetivos fundamentales:

– Analizar las relaciones entre los estadísticos obtenidos desde la TCT y desde los modelos de la TRI.

– Analizar la invarianza de los estadísticos obtenidos desde la TCT y desde la TRI.

Método

Diseño

Como ya se ha comentado anteriormente se trata de un estudio de simulación en el que la obtención de los datos se llevó a cabo de la siguiente manera:

En primer lugar, se generaron los valores de θ (nivel de habilidad de los sujetos) para 30 muestras de 500 sujetos cada una. Las 30 muestras se dividían en tres niveles:

– 10 muestras con un nivel de habilidad bajo ($\theta < -1{,}50$)

– 10 muestras con un nivel de habilidad medio ($-1{,}50 < \theta < 1{,}50$)

– 10 muestras con nivel de habilidad alto ($\theta > 1{,}50$)

A continuación se generaron los parámetros a, b y c correspondientes a 50 ítems. La distribución de los parámetros fue uniforme entre los siguientes niveles:

El parámetro *b* entre –3 y +3

El parámetro *a* entre 0,5 y 2

El parámetro *c* entre 0 y 0,25

Con estos datos se generaron 30 ficheros que incluían las respuestas simuladas de los sujetos de las distintas muestras *ante el mismo test de 50 ítems.*

En cada uno de los 30 ficheros de respuestas se estimaron los parámetros de los 50 ítems. Para ello se utilizaron los modelos logísticos de 1, 2 y 3 parámetros de la TRI y el modelo de la TCT. Los parámetros estimados de los ítems fueron:

Modelo logístico de 1 P: índice de dificultad *b*

Modelo logístico de 2 P: índice de dificultad *b* e índice de discriminación *a*

Modelo logístico de 3 P: índice de dificultad *b*, índice de discriminación *a* y pseudoazar *c*

TCT: índice de dificultad *p* e índice de discriminación $r_{ítem\text{-}test}$

– Asimismo se estimó el parámetro de habilidad θ de los sujetos de cada una de las muestras.

Las estimaciones se llevaron a cabo mediante el método de máxima verosimilitud marginal que es el que utiliza por defecto el programa BILOG. El proceso de simulación se llevó a cabo con el módulo correspondiente del programa GENESTE.

Procedimiento

Cada uno de los objetivos planteados requería un tipo distinto de análisis. No obstante, como paso previo, se estudió la precisión de las estimaciones de los parámetros obtenidas mediante los modelos logísticos de la TRI. Para ello, una vez estimados los parámetros de los ítems y la habilidad de los sujetos en las distintas muestras, se calculó la media de las estimaciones obtenidas en las 10 muestras de cada nivel y en el conjunto de las 30 muestras. A continuación, se llevó a cabo el estudio de la precisión calculando la correlación entre los valores medios obtenidos y los valores reales.

Una vez realizado este paso se procedió al estudio de los objetivos planteados.

Primer objetivo: analizar las relaciones entre los estadísticos obtenidos desde la TCT y desde los modelos de la TRI

Este primer objetivo se podía estudiar mediante los siguientes análisis:

– Comparación entre los valores obtenidos de θ mediante los modelos de la TRI y la puntuación empírica (X) de los sujetos en la TCT.

Para evaluar hasta qué punto difieren o no las estimaciones del nivel de aptitud o habilidad de los sujetos realizadas desde la TRI y desde la TCT, se calculó la correlación de Pearson entre ambos tipos de estimaciones. Los resultados se ofrecerán no sólo para el conjunto de las 30 muestras sino en cada uno de los niveles de habilidad diferenciados.

– Comparación entre los valores del índice de dificultad *b* obtenidos mediante los modelos de la TRI y los obtenidos de *p* desde la TCT.

Este análisis se llevó a cabo mediante la correlación de Pearson entre los valores de *b* obtenidos al utilizar los modelos de la TRI y

los valores de *p* obtenidos en la TCT. Los resultados se ofrecen no sólo para el conjunto de las 30 muestras sino para cada uno de los niveles de habilidad diferenciados.

– Comparación entre los valores del índice de discriminación *a* obtenidos mediante los modelos de la TRI y los obtenidos de $r_{ítem\text{-}test}$ desde la TCT.

Este análisis se llevó a cabo mediante la correlación de Pearson entre los valores de *a* obtenidos al utilizar los modelos de la TRI y los valores de $r_{ítem\text{-}test}$ en la TCT. Los resultados se ofrecen no sólo para el conjunto de las 30 muestras sino para cada uno de los niveles de habilidad diferenciados.

Segundo objetivo: analizar la invarianza de los estadísticos obtenidos desde la TCT y desde la TRI.

Se estudió la invarianza de los siguientes estadísticos:
– Invarianza de los índices de dificultad *b* y *p*
– Invarianza de los índices de discriminación *a* y $r_{ítem\text{-}test}$
– Invarianza de *c*

El estudio se llevó a cabo mediante la correlación de los valores medios de los estadísticos obtenidos en distintas muestras, en nuestro caso en muestras de distinto nivel de habilidad.

Resultados

Siguiendo el planteamiento utilizado por Fan (1998), se irá dando respuesta a cada una de las cuestiones planteadas comenzando, en primer lugar, por ofrecer los resultados obtenidos al estudiar la precisión de las estimaciones de los parámetros de los modelos de la TRI (ver tabla 1).

Los resultados obtenidos son similares a los encontrados en otros estudios de simulación de características semejantes en lo relativo a tamaño muestral y número de ítems. Es de destacar la falta de precisión en las estimaciones del parámetro *a* sobre todo en el modelo de 2 parámetros, esta falta de precisión es mayor en los grupos extremos de habilidad. También la precisión de las estimaciones del parámetro *c* es baja. Esta falta de precisión va a afectar a los resultados de los análisis posteriores.

Primer objetivo

– Comparación entre los valores obtenidos de θ mediante los modelos de la TRI y la puntuación empírica (X) de los sujetos en la TCT

Para analizar la relación entre la puntuación directa de los sujetos (TCT) y las estimaciones de θ obtenidas mediante los modelos logísticos de 1, 2 y 3 parámetros, se llevaron a cabo los siguientes pasos:

– En cada una de las 30 muestras simuladas se obtuvieron estimaciones de θ y las puntuaciones directas (X).
– Se calculó la correlación entre las estimaciones de θ y las puntuaciones (X) en cada muestra.
– Se calculó la media de las correlaciones obtenidas en las 10 muestras de cada nivel de habilidad.
– Se calculó la media de las correlaciones obtenidas en el conjunto de las 30 muestras.

Los resultados obtenidos aparecen recogidos en la tabla 2

Se podría haber utilizado el mismo procedimiento que utilizó Fan (1998), es decir, transformar cada uno de los coeficientes de correlación a puntuaciones Z de Fisher, promediar esos valores y, finalmente, la media obtenida transformarla en un nuevo coeficiente de correlación de Pearson; quizás hubiera sido un procedimiento más correcto, pero dado que lo que nos interesaba era analizar la cuantía de la relación no creímos necesario llevar a cabo esta transformación.

Como se puede observar las correlaciones encontradas son bastante altas, coincidiendo con las obtenidas en otros trabajos (Fan, 1998; Gil, Suárez y Martínez–Arias, 1999), lo que indica que desde un punto de vista aplicado se podría utilizar la TCT para estimar el nivel de aptitud o habilidad de los sujetos ya que no se apreciarían grandes diferencias en los resultados obtenidos desde una aproximación u otra, incluso cuando se tienen en cuenta los distintos niveles de habilidad. No obstante, conviene resaltar cómo la correlación es mas alta en el nivel medio de la escala de habilidad que en los niveles alto y bajo. Esto podría confirmar lo constatado por Gil, Suárez y Martínez-Arias (1999)

Tabla 1
Precisión de las estimaciones de los parámetros

Grupos	1P		2P			3P			
	$r_{\theta\theta'}$	$r_{bb'}$	$r_{\theta\theta'}$	$r_{bb'}$	$r_{aa'}$	$r_{\theta\theta'}$	$r_{bb'}$	$r_{aa'}$	$r_{cc'}$
Alto	0,929	0,909	0,935	0,817	0,384	0,947	0,871	0,395	0,383
Medio	0,959	0,940	0,962	0,959	0,413	0,976	0,990	0,801	0,591
Bajo	0,859	0,823	0,868	0,844	0,077	0,885	0,841	0,325	0,659
Total	0,916	0,891	0,922	0,857	0,359	0,936	0,956	0,765	0,624

Tabla 2
Correlaciones entre las estimaciones de la TRI y las de la TCT

Grupos	$r_{\theta'x}$			r_{bp}			$r_{aritem\text{-}test}$	
	1P	2P	3P	1P	2P	3P	2P	3P
Alto	0,970	0,960	0,968	0,924	0,710	0,683	0,773	0,812
Medio	0,992	0,985	0,990	0,986	0,956	0,970	0,753	0,345
Bajo	0.982	0,965	0,972	0,901	0,936	0,940	0,900	-0,040
Total	0,981	0,976	0,976	0,980	0,971	0,974		

en el sentido de que en la zona central de la nube de puntos del diagrama de dispersión originado por los valores de X (TCT) y θ (TRI) se observa una relación lineal más estrecha que en los extremos, pudiendo deberse a que la curva se asemeje a una función logística y más concretamente a la Curva Característica del Test. Como señala Lord (1980), la relación entre θ y la puntuación de los sujetos en el test es una relación monotónica pero no lineal ya que θ se obtiene mediante una transformación no lineal de la puntuación directa.

– *Comparación entre los valores del índice de dificultad b obtenidos mediante los modelos de la TRI y los de p obtenidos desde la TCT*

Para analizar la relación entre el índice de dificultad *p* de la TCT y las estimaciones de *b* obtenidas mediante los modelos logísticos de 1, 2 y 3 parámetros, se llevaron a cabo los siguientes pasos:

– En cada una de las 30 muestras se obtuvieron estimaciones de *b* mediante los modelos de 1, 2 y 3 parámetros y los valores de *p*.

– Se calculó la correlación entre las estimaciones de *b* y los valores de *p* en cada muestra.

– Se calculó la media de las correlaciones obtenidas en las 10 muestras de cada nivel de habilidad.

– Se calculó la media de las correlaciones obtenidas en el conjunto de las 30 muestras.

Los resultados aparecen recogidos en la tabla 2

Teniendo en cuenta que el índice de dificultad de la TCT no es más que la proporción de aciertos, en realidad representa un índice de facilidad. Por eso, en la tabla 2 se ha cambiado el signo de las correlaciones encontradas. Se ha utilizado también la correlación directa entre los valores de *b* y los de *p* ya que lo que nos interesaba era simplemente analizar la relación entre ambos estadísticos. Se podía haber utilizado el valor de p normalizado asumiendo que la distribución subyacente del rasgo medido por el ítem es una distribución normal. Los resultados obtenidos en otros trabajos que han utilizado esta transformación indican que las correlaciones obtenidas entre *b* y *p normalizado* son algo más altas (Fan, 1998) lo que mejoraría los resultados obtenidos.

Todas las correlaciones obtenidas son estadísticamente significativas ($p \leq 0{,}01$); sin embargo, se observa que también ahora es en el nivel medio en el que las correlaciones son más altas.

Conviene destacar que las correlaciones obtenidas utilizando los valores de *b* estimados mediante el modelo de Rasch (1P) son, en general, más altas que las obtenidas al utilizar los valores de b estimados mediante el modelo de 2 y 3 parámetros. Estos resultados coinciden con los de otros trabajos (Fan, 1998; Gil, Suárez y Martínez-Arias, 1999).

– *Comparación entre los valores del índice de discriminación a obtenidos mediante los modelos de la TRI y los obtenidos de $r_{ítem\text{-}test}$ desde la TCT*

Para analizar la relación entre el índice de discriminación $r_{ítem\text{-}test}$ de la TCT y las estimaciones de a obtenidas mediante los modelos logísticos de 2 y 3 parámetros, se llevaron a cabo los siguientes pasos:

– En cada una de las 30 muestras se obtuvieron estimaciones de *a* mediante los modelos de 2 y 3 parámetros y los valores de $r_{ítem\text{-}test}$

– Se calculó la correlación entre las estimaciones de *a* y los valores de $r_{ítem\text{-}test}$ en cada muestra y para cada modelo

– Se calculó la media entre las correlaciones obtenidas en las 10 muestras de cada nivel de habilidad.

– Se calculó la media entre las estimaciones obtenidas en el conjunto de las 30 muestras.

Los resultados aparecen recogidos en la tabla 2

Los valores del índice de discriminación de la TCT se han obtenido mediante la correlación biserial puntual entre las puntuaciones de los sujetos en el ítem y las obtenidas en el test total después de haber eliminado de éstas la contribución del ítem.

Dado que en el modelo de 1 parámetro se asume que el índice de discriminación es constante, no se han analizado las relaciones entre las estimaciones de la TCT y de la TRI. En la tabla se ofrecen sólo los resultados obtenidos con los modelos de 2 y 3 parámetros.

A pesar de que las relaciones encontradas son más bajas que al comparar el índice de dificultad y el nivel de habilidad, excepto la correlación obtenida en el modelo de 3 parámetros y en el nivel de aptitud bajo, todas las correlaciones son estadísticamente significativas ($p \leq 0{,}01$).

En el modelo de dos parámetros las correlaciones encontradas son aceptables, siendo más altas en las muestras de nivel bajo; por el contrario, en el modelo de tres parámetros, en las muestras de nivel de aptitud bajo, las correlaciones encontradas son bajas e incluso próximas a cero como demuestra la media obtenida.

La justificación de los valores obtenidos, además de poder deberse a la falta de precisión de las estimaciones obtenidas del parámetro *a*, puede estar en que la relación entre los dos tipos de estadísticos sea una relación no lineal y, por lo tanto, sea necesario analizarla mediante otros procedimientos.

Segundo objetivo

La propiedad de invarianza de las estimaciones es la piedra angular de la TRI y lo que la da fuerza frente a la TCT. Realmente, el problema de la dependencia de las estimaciones que, al menos a nivel teórico, caracteriza a la TCT justificaría el gran desarrollo que ha tenido la TRI. Ahora bien, como ya se ha dicho en algunas ocasiones, las diferencias entre los resultados ofrecidos por estos

Tabla 3
Invarianza de las estimaciones de los parámetros de la TRI y de la TCT

Grupos	TRI (*b*)			TCT	TRI (*a*)		TCT	TRI (c)
	1P	2P	3P	(p)	2P	3P	$r_{item\text{-}test}$	
$r_{alto/medio}$	0,921	0,786	0,818	0,853	-0,247	0,276	-0,450	0,598
$r_{alto/bajo}$	0,713	0,701	0,728	0,579	-0,398	0,250	-0,535	0,383
$r_{bajo/medio}$	0,912	0,875	0,838	0,840	0,440	0,465	0,145	0,865

dos marcos teóricos tienen que ser sustanciales para que compense la utilización de los modelos de la TRI, con la complejidad y formalización matemática que requieren, en lugar del modelo más sencillo de la TCT.

Analizar la invarianza de las estimaciones obtenidas es el segundo objetivo de nuestro trabajo, y lo iremos haciendo con cada uno de los estadísticos obtenidos.

– *Invarianza de los índices de dificultad b y p*

De verificarse esta propiedad, los valores de *b* y de *p* serían independientes de la muestra de sujetos que respondiera al ítem. Para ver hasta qué punto se cumple este supuesto se llevaron a cabo los siguientes pasos:

– Se calcularon todas las intercorrelaciones entre los valores obtenidos de *b* en las distintas muestras. Lo mismo se hizo con los valores de *p*.

– Tanto para los valores de *p* como para los de *b*, se calculó la media de las correlaciones obtenidas entre las muestras de nivel alto y las de nivel medio.

– Tanto para los valores de *p* como para los de *b*, se calculó la media de las correlaciones obtenidas entre las muestras de nivel alto y las de nivel bajo.

– Tanto para los valores de p como para los de *b*, se calculó la media de las correlaciones obtenidas entre las muestras de nivel medio y las de nivel bajo.

Los resultados aparecen recogidos en la tabla 3

Se puede observar que las correlaciones más altas son las obtenidas al utilizar el modelo logístico de 1 parámetro. Sin embargo, cuando se analiza la invarianza de *b* en los modelos logísticos de 2 y 3 parámetros los resultados obtenidos no difieren sustancialmente de los obtenidos en el marco de la TCT. Teniendo en cuenta que, tal y como se muestra en la tabla 1, es mayor la precisión de las estimaciones en el modelo logístico de 1 parámetro y dado que la propiedad de invarianza se verifica cuando el ajuste del modelo a los datos es un ajuste perfecto (Hambleton, Swaminathan y Rogers, 1991), los resultados pueden deberse al mejor ajuste del modelo de Rasch.

Por otra parte, se observa la misma tendencia en los resultados obtenidos en el marco teórico de la TCT y en el de la TRI: los valores más bajos se encuentran al correlacionar los dos grupos extremos de habilidad (nivel alto/nivel bajo). Si esto ocurriera en el caso de que el ajuste a los datos fuera un ajuste perfecto se estaría violando uno de los pilares de la TRI, la independencia de los parámetros de los ítems del nivel de habilidad de los sujetos que responden. Por otra parte, obtener un ajuste perfecto en algunas situaciones es bastante inusual, lo que implicaría que las ventajas de utilizar el marco teórico de la TRI en estas situaciones quedarían en entredicho y quizás fuera igualmente práctico utilizar el marco de la TCT, con la ventaja de su mayor sencillez.

– *Invarianza de los índices de discriminación a y* $r_{ítem\text{-}test}$

De verificarse esta propiedad, los valores de *a* y de $r_{ítem\text{-}test}$ serían independientes de la muestra de sujetos que respondiera al ítem. Para ver hasta qué punto se cumple este supuesto se llevaron a cabo los siguientes pasos:

– Se calcularon todas las intercorrelaciones entre los valores obtenidos de *a* en las distintas muestras. Lo mismo se hizo con los valores de $r_{ítem\text{-}test}$

– Tanto para los valores de *a* como para los de $r_{ítem\text{-}test}$, se calculó la media de las correlaciones obtenidas entre las muestras de nivel alto y las de nivel medio.

– Tanto para los valores de *a* como para los de $r_{ítem\text{-}test}$, se calculó la media de las correlaciones obtenidas entre las muestras de nivel alto y las de nivel bajo.

– Tanto para los valores de *a* como para los de $r_{ítem\text{-}test}$, se calculó la media de las correlaciones obtenidas entre las muestras de nivel medio y las de nivel bajo.

Los resultados se recogen en la tabla 3.

Como puede observarse, los resultados son bastante desalentadores, difieren sustancialmente de los encontrados por Fan (1998) ya que, en ningún momento, podemos considerar que se verifique la propiedad de invarianza. En el modelo de 2 parámetros se han encontrado correlaciones negativas entre las estimaciones obtenidas en las muestras de nivel alto/medio y alto/bajo. Algo mejor es el resultado obtenido entre las estimaciones en las muestras de nivel medio/bajo. Aunque con valores diferentes podemos decir que en los resultados obtenidos con los estadísticos de la TCT se sigue la misma tendencia.

En el modelo de 3 parámetros los resultados son algo más alentadores ya que, al menos, las correlaciones encontradas entre las estimaciones obtenidas en las muestras de distinto nivel de aptitud son positivas aunque bastante bajas también y siguiendo la misma tendencia que en los demás, los valores más bajos se obtienen al correlacionar los resultados obtenidos en los grupos extremos.

– *Invarianza del parámetro c de pseudoadivinación*

Si, como señala Lord (1980), el parámetro *c* no viene afectado por el origen de la escala ni por la unidad de medida, su estimación debería ser idéntica en las distintas muestras en el caso en que se verificara la propiedad de invarianza y, por lo tanto, la correlación entre las distintas estimaciones sería una correlación perfecta. Para ver hasta qué punto se cumple este supuesto se llevaron a cabo los siguientes pasos.

– Se calcularon todas las intercorrelaciones entre los valores obtenidos de *c* en las distintas muestras.

– Se calculó la media de las correlaciones obtenidas entre las muestras de nivel alto y las de nivel medio.

– Se calculó la media de las correlaciones obtenidas entre las muestras de nivel alto y las de nivel bajo.

– Se calculó la media de las correlaciones obtenidas entre las muestras de nivel medio y las de nivel bajo.

Los resultados se recogen en la tabla 3.

Aunque también ponen en duda el principio de invarianza, los resultados son algo más alentadores que los encontrados en el apartado anterior con el índice de discriminación. Se sigue dando la misma tendencia y el valor más bajo se ha encontrado el analizar la invarianza en grupos extremos.

Discusión y conclusiones

Una de las ventajas que ofrecen los trabajos empíricos es la de ofrecer la posibilidad de contrastar los resultados obtenidos por distintos investigadores y la realización de nuevos trabajos bajo distintas condiciones con el fin de estudiar las posibles variables que puedan estar incidiendo en el caso de que los datos no sean coincidentes.

Esta posibilidad, junto a la disparidad de resultados obtenidos en distintos trabajos (Lawson, 1981; Miller y Linn, 1988; Cook. Eignor y Taft , 1988) al analizar las relaciones entre los estadísticos (ítems/sujetos) de la TRI y de la TCT, es la que dirigió los objetivos fundamentales de este trabajo que podemos resumir así:

– Analizar las relaciones entre los estadísticos obtenidos desde el marco de la TCT y desde el de la TRI, así como la invarianza de los mismos.

Ahora bien, teniendo en cuenta la dificultad que conlleva, en muchas situaciones reales, la obtención de muestras de gran tamaño, al hacer el diseño de nuestro trabajo optamos por utilizar una muestra de tamaño mediano (N=500) aún a sabiendas de las incidencias que esta variable podía tener en los resultados obtenidos ya que el ajuste de los modelos de la TRI mejora, sustancialmente, al aumentar el tamaño muestral. La falta de precisión de las estimaciones de los parámetros *a* y *c* en los modelos de 2 y 3 parámetros puede ser una prueba de ello; aunque, es cierto, que la estimación del parámetro *c* suele presentar problemas de inestabilidad con independencia del tamaño de la muestra. Las estimaciones del parámetro *b* muestran una alta precisión en los tres modelos, pero es en el modelo de tres parámetros en el que se alcanzan valores más altos.

La precisión de las estimaciones del nivel de aptitud de los sujetos es más alta en el modelo de 3 parámetros, sobre todo para los grupos de nivel alto y medio, seguido por modelo de 2 parámetros y finalmente por el de 1 parámetro; en estos dos modelos se observa la misma tendencia que en el de 3 parámetros ya que la precisión es más alta en los niveles de aptitud alto y medio.

Respecto a los resultados encontrados al analizar las relaciones entre los estadísticos obtenidos mediante los modelos de la TRI y los de la TCT hemos de resaltar lo siguiente:

La relación entre las estimaciones del nivel de aptitud de los sujetos y las puntuaciones directas que han obtenido en el test es muy alta. En los distintos niveles, y en los tres modelos logísticos, está próxima a la unidad, siendo en el modelo de 1 parámetro donde los valores son más altos. La tendencia a lo largo de los tres modelos es la misma, la relación entre θ y X es más estrecha en el nivel medio que en los niveles alto y bajo lo que, como ya se ha comentado, puede ser un reflejo de la falta de relación lineal entre θ y X.

Asimismo, se observa una alta relación entre el índice de dificultad de la TCT y las estimaciones del parámetro *b* de la TRI. También en este caso, las estimaciones más próximas a la unidad de obtienen al utilizar el modelo de 1 parámetro, excepto en el grupo de nivel bajo en el que la correlación más alta se observa en el modelo de 3 parámetros.

Respecto a la relación entre las estimaciones del índice de discriminación de la TCT y de la TRI conviene destacar que los valores más altos se obtienen en el modelo de 2 parámetros y en el grupo de nivel de aptitud bajo.

En cuanto al problema de la invarianza los resultados son bastante desalentadores excepto en lo que se refiere a los índices de dificultad *b* y *p.*

Cook, Eignor y Taft (1988) observaron una falta de invarianza en los índices de dificultad *b* y *p,* y Miller y Linn (1988) utilizando muestras grandes encontraron que no se verificaba la propiedad de invarianza en sus resultados.

Esta controversia en los resultados obtenidos en distintas investigaciones, nos lleva a creer en la necesidad de continuar investigando en este sentido; por eso, en un trabajo que estamos realizando se están probando distintos procedimientos para el análisis de la invarianza de los parámetros con el fin de aportar algún nuevo dato que permita obtener mejores resultados.

Referencias

Cook,L.L., Eignor, D.R. y Taft, H.L. (1988). A comparative study of the effects of recency of instruction on the stability of IRT and conventional item parameter estimates. *Journal of Educational Measurement,* 25, 31-45.

Fan, X. (1998). Item response theory and classical test theory: an empirical comparison of their item/person statistics. *Educational and Psychological Measurement,* 58, (3), 357-381.

Gil, G., Suárez, J.C. y Martínez, R. (1999). Aplicación de un procedimiento iterativo para la selección de modelos de la teoría de respuesta al ítem en una prueba de rendimiento lector. *Revista de educación,* 319. 253-272.

Hambleton, R.K. y Swaminathan, H. (1985). *Item response theory: principles and applications.* Boston, Kluwer.

Hambleton, R.K. y Swaminathan, H. Y Rogers, H.J. (1991). *Principles and application of item response theory.* Beverly Hills, Sage.

Lawson, S. (1991). One parameter latent trait measurement: Do the results justify the effort?, en B. Thompson (Ed.). *Advances in educational research: substantive findings, methodological developments.* Vol. 1, 159-168. Greenwich, JAI.

Lord, F.M. (1980). *Applications of item response theory to practical testing problems,* Hillsdale, LEA.

Mellenberg, G.J. (1994). Generalized Linear Item Response Theory, *Psychological Bulletin,* 115, (2), 300-307.

Miller, M.D. y Linn, R. L. (1988) Invariance of item characteristic functions with variations in instructional coverage. *Journal of educational measurement,* 25, 205-219.

Muñiz, J. (1997). *Introducción a la teoría de respuesta a los ítems,* Madrid, Pirámide.

Psicothema (2000). Número especial de Metodología. Vol. 12, Supl. Nº 2.

Spearman, C. (1904). The proof and measurement of association between two things, *American Journal of Psychology,* 15, 72-101.

Spearman, C. (1907). Demonstration of formulae for true measurement of correlation, *American Journal of Psychology,* 18, 151-169.

Spearman, C. (1913). Correlations of sums and differences, *British Journal of Psychology,* 5, 417-426.

Stage, C. (1998a). A comparison between item analysis based on Item Response Theory and Classical Test theory. A study of the SweSAT Subtest WORD. *Educational Measurement,* 29. (Umeà University, Department of Educational Measurement). On-line.

Stage, C, (1998b). A comparison between item analysis based on Item Response Theory and Classical Test Theory. A study of the SweSAT Subtest ERC. *Educational Measurement,* 30. (Umeà University, Department of Educational Measurement). On-line.

Stage, C, (1999). A comparison between item analysis based on Item Response Theory and Classical Test Theory. A study of the SweSAT Subtest READ. *Educational Measurement,* 33. (Umeà University, Department of Educational Measurement). On-line.

Van der Linden, W. J. y Hambleton, R.K. (1997). *Handbook of modern item response theory.* New york, Springer-Verlag.

Aceptado el 27 de diciembre de 2000