Comparación de modelos IRT realizando análisis a la subprueba conceptos con dibujos del WISC IV

Sebastian Castro Alvarez[1]
Estudiante de Psicología
Universidad Nacional de Colombia

Resumen

En el área de la medición psicológica, se destaca en la actualidad la teoría de respuesta al ítem (IRT). Esta consta de diferentes modelos para el análisis de ítems dicotómicos, que pueden representar un problema para el investigador al momento de decidir qué modelo utilizar para sus datos. Así este articulo busca realizar una comparación de los parámetros estimados (a, b, c) con los modelos logísticos de uno, dos y tres parámetros y con dos métodos diferentes de estimación, por máxima verosimilitud marginal y estimación modal bayesiana, para una base de 388 datos de la subprueba conceptos con dibujos del WISC IV, aplicado a niños y niñas entre los 6 y 16 años. Para el análisis, se realizó la comprobación de los supuestos de unidimensionalidad e independencia local con los programas MODFIT y NORHAM, y se realizó la estimación de los parámetros con el programa MULTILOG, finalmente se compararon los parámetros estimados, por medio de la correlación de Pearson y pruebas de inferencia estadística cuando no hubo correlación significativa. En los resultados se encontró que los parámetros a y b se encuentran muy relacionados, independientemente del método de estimación o del modelo utilizado, con el parámetro c del modelo logístico de tres parámetros, se encontró una diferencia significativa entre los dos métodos de estimación, siendo el método de estimación modal bayesiana, el que realiza la estimación más fina. En conclusión se puede decir que la elección del modelo de IRT, va de acuerdo a los intereses mismos del investigador, y queda abierta la discusión en torno a la pertinencia y practicidad de los diferentes modelos.

Introducción

[1] e-mail: secastroal@unal.edu.co / secastroal@gmail.com

La Teoría de Respuesta al Ítem (IRT por sus siglas en ingles), surgió como respuesta a los inconvenientes y limitaciones que presentaba, la Teoría Clásica de los Test (TCT). Técnicamente las limitaciones eran, en primer lugar la ausencia de invarianza de los parámetros, más exactamente la habilidad de las personas en un constructo iba a depender totalmente de la longitud y dificultad del test, razón por la cual este valor estimado iba a variar de un test a otro, en segundo lugar se asumía que la precisión del test era la misma para cualquier nivel de rasgo, y finalmente no era posible especificar qué tan adecuado o no, era el modelo para explicar el comportamiento de las puntuaciones observadas (Abad, Olea, Ponsoda, & García, 2011).

Así la IRT ya lleva un poco más de 50 años desde que se formuló, aunque sus antecedentes se pueden remontar algunos años más atrás con los trabajos de Thurstone en 1929 sobre la relación de la edad y la probabilidad de responder los ítems del test de inteligencia de Binet, con Ferguson en 1942 y sus trabajos en psicofísica (Muñiz, 1997). Como tal, el nacimiento formal de la IRT se da con Frederic Lord en 1952, siendo el resultado de su tesis doctoral, en la cual ya hace la formulación de tres modelos para ítems dicotómicos que tienen en cuenta los diferentes parámetros como la dificultad, la discriminación y la adivinanza, sin embargo estos primeros modelos se vieron enmarcados matemáticamente en la función de la curva normal acumulada, lo cual representaba una altísima complejidad en los cálculos (Muñiz, 1997). Posteriormente fue Allan Birnbaum, quien desarrolló los modelos logísticos de uno, dos y tres parámetros, que vendrían a sustituir los modelos de ojiva normal de Lord, y en 1960 George Rasch propone un modelo logístico de un parámetro, con pequeñas variaciones, que se conoce como el modelo de Rasch (Muñiz, 1997).

Entre los cambios más importantes que realiza la IRT con respecto a la TCT es el cambio del modelo matemático sobre el cual se fundamenta, pasando de un modelo lineal a un modelo logístico, que realiza una representación más acertada del desempeño de las personas en los test. Con este gran cambio, se derivaron otros, como que se dejó de lado la concepción de puntuación verdadera y puntuación observada, y que solo daban información del desempeño de la persona en la totalidad de la prueba, y se pasó a hablar únicamente del nivel de rasgo expresado con “θ”, y que se estima para cada persona, en cada ítem,

transformando la escala de medida, a una escala que conceptualmente está en el intervalo de (-∞,+∞), y que la media se encuentra en 0. Esta misma escala se utiliza para dar el parámetro de dificultad del ítem que se expresa con “b”, por su lado la discriminación que se representa con “a”, teóricamente también debería estar en ese intervalo pero su rango se puede restringir fácilmente al intervalo (0,2), y finalmente el parámetro de adivinanza expresado con “c” cobra mayor importancia en la IRT pues la estimación que se realiza del mismo es mucho más fina a la que se podía realizar en la TCT y se esperan valores positivos cercanos a 0 (Lord, 1990). En este orden de ideas los diferentes modelos logísticos de uno, dos y tres parámetros, proponen una serie de ecuaciones [1] (Abad, Olea, Ponsoda, & García, 2011), que permiten en primer lugar estimar todos los parámetros incorporados en el modelo, y construir lo que se denomina la Curva Característica del Ítem.

Modelo logístico de un parámetro (ML1P) $$P_j(\theta) = \frac{1}{1+e^{-Da(\theta-b_j)}}$$

Modelo logístico de dos parámetros (ML2P) $$P_j(\theta) = \frac{1}{1+e^{-Da_j(\theta-b_j)}} \quad [1]$$

Modelo logístico de tres parámetros (ML3P) $$P_j(\theta) = c_j + (1-c_j)\frac{1}{1+e^{-Da_j(\theta-b_j)}}$$

En estas ecuaciones hay valores constantes como D=1,702, y la base de los logaritmos neperianos e=2,718, y en términos generales expresan la probabilidad, de una persona para responder correctamente un ítem dado su nivel de habilidad.

Sin embargo, la IRT pone algunas restricciones, por lo cual no será aplicable a cualquier conjunto de datos, las restricciones que supone es en primer lugar satisfacer la unidimensionalidad, que indica que un conjunto de ítems de un test, o el test en su totalidad solo evalúan un único rasgo, o habilidad, o constructo (Hambleton, Swaminathan, & Rogers, 1991). La unidimensionalidad se suele comprobar con los procedimientos denominados Análisis Factoriales (AF), que se pueden categorizar dependiendo si se basan en los criterios de mínimos cuadrados o de máxima verosimilitud, y que efectúan

correlaciones entre las variables; cuando estas son medidas dicotómicas, es recomendable que el AF se realice por medio de las correlaciones tetracóricas (Ferrando, 1996). Normalmente un grupo de datos se consideran unidimensionales cuando el primer factor explica un 50% de la varianza, sin embargo estos criterios todavía son un tema de mucha discusión, e incluso se puede llegar a considerar un instrumento unidimensional si solo se extrae un factor de segundo orden, así existan varios de primer orden (Burga León, 2012). Satisfacer este supuesto, es de las principales limitaciones que tiene la IRT respecto a aplicabilidad.

El segundo supuesto de la IRT, es el de independencia local, que consiste en que por cada par de ítems, dado un nivel de habilidad, la probabilidad de contestar correctamente el primero no depende de la probabilidad de contestar correctamente el segundo, o viceversa, es decir que en otras palabras la probabilidad de contestar los dos ítems correctamente es igual al producto de las probabilidades de contestar correctamente cada ítem (Chen & Thissen, 1997). La comprobación de este supuesto se ha realizado esencialmente por medio de las tablas de contingencia, Chen & Thissen (1997) exponen 4 índices diferentes que se pueden obtener para saber si hay o no hay independencia local, que son el X^2 de Pearson, el estadístico G^2, la diferencia de los coeficientes estandarizados ϕ, y la proporción de la diferencia estandarizada Log-Odds, aunque de estos el más utilizado es el X^2 de Pearson. En muchas ocasiones se omite la verificación del segundo supuesto, porque la independencia local se deriva de la unidimensionalidad (Abad, Olea, Ponsoda, & García, 2011), lo que supone un problema, ya que con cumplirse la unidimensionalidad no necesariamente implica que se cumpla la independencia local.

Además de la comprobación de supuestos en la IRT salen a relucir lo métodos de estimación, todos los parámetros que se sean implicados en un modelo especifico, serán variables desconocidas, entre todos los métodos de estimación desarrollados, se pueden mencionar el de máxima verosimilitud marginal, el de máxima verosimilitud conjunta, las estimaciones bayesianas conjuntas y marginales, la estimación heurística, e incluso métodos basados en análisis factorial no lineal (Hambleton, Swaminathan, & Rogers, 1991), sin embargo no existe claridad cuando es mejor uno u otro método, y la elección queda supeditada al paquete estadístico utilizado para realizar los análisis. Aun así, los

métodos más utilizados son el de máxima verosimilitud marginal que se basa en la función de verosimilitud, y primero fija el parámetro de habilidad de las personas para luego estimar los parámetros y viceversa hasta que los valores estimados converjan (Abad, Olea, Ponsoda, & García, 2011), y las estimaciones bayesianas modal a posteriori (MAP) y esperada a posteriori (EAP) que incorporan en la función de verosimilitud la distribución de los parámetros a estimar (Abad, Olea, Real, & Ponsoda, 2002), la importancia de la selección del método de estimación esta, en que a partir de estos se calculan los errores estándar de estimación y el sesgo, índices que se relacionan directamente con el ajuste de los modelos (Abad, Olea, Real, & Ponsoda, 2002).

Adicionalmente la IRT proporciona la posibilidad de indicar la bondad de ajuste de los diferentes modelos, hay otros conceptos importantes como la curva de información del ítem, y la curva característica y de información del test (Muñiz, 1997), que siguen proporcionando información para análisis y que permiten comprender mucho mejor la prueba de lo que se podía comprender con la TCT, por ejemplo las curvas de información son lo equivalente a la medida de la confiabilidad, y además al no ser constante, indica en que niveles de habilidad es más o menos acertada la medición, por ítem y para todo el test.

Con la existencia conjunta de dos teorías y varios modelos, es razonable preguntarse cuando es mejor uno o cuando es mejor otro, y en qué tipo de condiciones del tamaño de muestra, tamaño del test, normalidad o no normalidad de los datos, etc. Por ejemplo algunos estudios han comparado los indicadores que se obtienen en las dos teorías de la medición, en Suecia con su examen de estado, SweSAT concluyen que por lo menos en los parámetros de dificultad y discriminación obtenidos bajo las dos teorías tienen correlaciones altas, aspecto por el cual concluyen que es indiferente el uso de una u otra teoría para analizar los datos recolectados (Stage, 1998) (Stage, 2003), y se da primacía a la facilidad para los cálculos que a la información que se suministra. En diferentes estudios se ha optado por realizar los análisis bajo el marco de las dos teorías como por ejemplo en Rodríguez, Casas, & Medina (2005) analizando los exámenes de estado de educación superior en Colombia, en Iraurgi, Lozano, Gónzalez-Saiz, & Trujols (2008) analizando la escala de intensidad de apoyos (SIS) para niños y adolescentes y en Guillen, Verdugo, Arias, Navas, & Vicente (2012) analizando la escala de severidad de la dependencia.

Se han intentado hacer otros tipos de comparaciones, DeMars (2008) analizó desde los modelos clásicos dicotómicos y politómicos de la TCT, y con el modelo de tres parámetros y el modelo de respuesta nominal de la IRT, y comparó los errores estándar estimados según los modelos, argumentando, que estos si están en la misma escala. Así mismo las conclusiones las da en términos de la confiabilidad encontrada para el instrumento utilizado, por lo cual aseguran que los modelos politómicos tenían mejores estimadores de confiabilidad que los dicotómicos, en adición mencionan que los modelos clásicos tienen ventajas sobre las exigencias de los cálculos, y son más fáciles de explicar a los usuarios, aspecto que ya se corroborado con otros estudios. Otra investigación similar fue realizada por Kinsey (2003), que comparó los modelos de RASCH y el de IRT, manipulando la longitud de los test (10, 20 o 30 ítems) y 11 proporciones del tipo de ítems desde 100% dicotómicos hasta el 0% y el resto politómicos. En este caso no encontró diferencias significativas en la habilidad estimada, concluyendo que los modelos son igualmente efectivos, sin embargo frente a mayores proporciones de ítems politómicos y mayor longitud del test, los resultados son ligeramente mejores con los modelos IRT frente a los modelos de RASCH. También Muñiz, Rogers & Swaminathan (1989 citados en Muñiz, 1997), querian comparar como funcionaba el modelo de Rasch, poniendo a priori diferentes condiciones para el parametro de discriminación y adivinanza, teniendo así 12 situaciones diferentes a partir de simulación, y adicionalmente comparar si la estimación de los parametros mejoraba con los modelos de dos o tres parametros. Con este estudio se pudieron dar cuenta de que en las condiciones que propusieron las estimaciones que realiza el modelo de Rasch y el ajuste del modelo se ve poco o nada afectado cuando el indice de adivinanza es diferente de cero, y cuando el índice de discriminación no es constante para todos los ítems, además la estimación de los parametros con los otros modelos tampoco mejoraban en términos significativos, una vez más abogan por realizar lo procedimientos más simples.

Así la presente investigación, busca comparar como se dá la estimación de los parametros y el ajuste, para los modelos logisticos de uno, dos y tres parametros, utilizando dos metodos de estimación el de máxima verosimilitud marginal, y estimación modal bayesiana, para unos datos correspondientes a la de prueba figuras incompletas de la escala Wechsler de Inteligencia para Niños (WISC IV), de tal forma que se pueda dar soporte para

la toma de desición con respecto al modelo y al método de estimación cuando se esta trabajando con datos empiricos.

Metodología

Para realizar este estudio comparativo de los modelos IRT, se hizo uso de las bases de datos disponibles en el Servicio de Atención Psicológica de la Universidad Nacional de Colombia, se tomó la base de datos de la Escala Wechsler de Inteligencia para Niños (WISC IV). Estos datos fueron recopilados en contexto clínico entre el 8 de febrero de 2008 y el 21 de noviembre de 2012. La base contaba con un total de 433 casos, pero se redujo a 388, teniendo en cuenta casos repetidos, con información incompleta, o que se encontraran por fuera del rango adecuado de edad.

La muestra estuvo compuesta por niños y niñas con edad en el rango de los 6 a los 16 años (X=10.4, SD=2.94), al momento de la aplicación. El 71% de la población fue de género masculino y el 29% del género femenino; así mismo el 10,8% pertenecía a estrato 1, el 59,5% a estrato 2, el 24,2% a estrato 3, el 4,9 a estrato 4 y el 0,5% al estrato 5.

Teniendo la base de datos, se procedió con la elección de la subprueba que se seleccionaría para el análisis, de estas, en primer lugar se descartaron las que no contaban con ítems exclusivamente dicotómicos, y en segundo lugar se descartaron las opcionales, teniendo así para un posible análisis la subprueba conceptos con dibujos que evalúa la capacidad para generar categorías, hace parte de índice de razonamiento perceptual, y es un buen predictor de la capacidad de aprendizaje (Wechsler, 2005), y la subprueba de matrices que evalúa el análisis perceptual, la capacidad de encontrar patrones y relaciones lógicas, también perteneciente al índice de razonamiento perceptual (Wechsler, 2005). La decisión final, fue analizar conceptos con dibujos, que se compone de 28 ítems en orden de dificultad, porque en esta subprueba se contaban con más datos que en la subprueba de matrices, además, al estar compuesta por menos ítems, podría ser más viable satisfacer el supuesto de unidimensionalidad.

Con la base de datos depurada y lista para los análisis, y poniendo en consideración que los ítems son dicotómicos, y que los datos no se ajustaban a la distribución normal, para comprobar la unidimensionalidad se realizó un procedimiento de análisis factorial no lineal con el programa Norham, para la estimación de los parámetros de los tres modelos logísticos, considerando los dos métodos de estimación se utilizó el programa estadístico Multilog 7.03[2], y la comprobación de la independencia local y la bondad de ajuste de los modelos se comprobó con el estadístico X^2 con el programa ModFIT. Finalmente se correlacionaron los parámetros estimados por los diferentes modelos con el programa estadístico SPSS 20[2].

Resultados

En el análisis factorial no lineal confirmatorio se pueden observar los siguientes pesos para cada ítem en la tabla 1, este modelo tuvo buen ajuste a los datos ya que se obtuvo un RMSR=0,007 y un índice de tanaka γ=0,985, el primero debe ser menor a 0,06 pues se busca que sea cercano a 0 porque está relacionado con los residuales, y el segundo debe ser muy cercano a 1, con esto se satisface el supuesto de unidimensionalidad para poder correr los análisis por los distintos modelos IRT.

Tabla 1: ***Pesos factoriales***

Ítem	Factor
1	0.778
2	0.799
3	0.922
4	0.767
5	0.907
6	0.830
7	0.769
8	0.829
9	0.863
10	0.665
11	0.773
12	0.674
13	0.711
14	0.747
15	0.869

[2] Con licencia de la Universidad Nacional de Colombia

16	0.717
17	0.710
18	0.666
19	0.689
20	0.736
21	0.705
22	0.712
23	0.692
24	0.479
25	0.621
26	0.662
27	0.658
28	0.631

Con respecto a la independencia local, uno de los inconvenientes que se presentan es que el programa ModFIT permite verificarlo después de corridos los análisis por IRT y cuando ya se tienen estimados los parámetros. Así el índice obtenido fue el X^2 de Pearson, que por tablas de contingencia comprueba la hipótesis nula de que hay independencia local de los datos, así con las parejas de ítems que se corroboraron cuando se realizó con los modelos de dos y tres parámetros, independientemente del método de estimación, también se satisface este supuesto, cosa que no sucede con el modelo de un parámetro donde dos pares de ítem muestran tener dependencia.

Tras hacer el análisis con los modelos de IRT se hizo la estimación de los diferentes parámetros de discriminación, dificultad y adivinanza según correspondía. Los estadísticos descriptivos con mínimo, máximo, media y desviación estándar para cada parámetro dados los diferentes modelos y métodos de estimación se pueden observar en la tabla 2. Por su lado, en general se puede decir que todos tuvieron un buen ajuste de los datos a los modelos, sin embargo se observaron ítems que no ajustaban, en el ML1P con estimación bayesiana, no ajustan los ítems 25, 26, 27, en el ML1P el ítem 27, en el ML2P independientemente del método de estimación no ajusta el ítem 15, y en el ML3P con estimación bayesiana no ajustan los ítem 26, 27 y 28. Esto da como resultado que el modelo que mejor se ajusta a los datos viene a ser el ML3P por la estimación de máxima verosimilitud marginal, en donde todos los ítems se ajustaron.

Tabla 2: ***Estadísticos descriptivos de los parámetros estimados***[3]

	N	Mínimo	Máximo	Media	SD
b ML1P EMB	28	-2,08	3,21	-,0004	1,37924
b ML1P MML	28	-2,10	3,27	,0002	1,39528
b ML2P EMB	28	-2,25	3,20	-,0001	1,39326
b ML2P MML	28	-2,24	3,26	,0005	1,38279
b ML3P EMB	28	-2,15	3,93	,0805	1,44424
b ML3P MML	28	-2,27	3,22	-,0362	1,37830
a ML2P EMB	28	,73	1,61	1,1452	,19548
a ML2P MML	28	,72	1,70	1,1746	,21966
a ML3P EMB	28	,74	1,62	1,1741	,21063
a ML3P MML	28	,73	1,70	1,1781	,22011
c ML3P EMB	28	,01	,20	,0654	,04517
c ML3P MML	28	,00	,00	,0000	,00019

Con los parámetros estimados se corrió la prueba de normalidad de Shapiro-Wilk, encontrando que los parámetros de dificultad y discriminación para los diferentes modelos, se ajustan a una distribución normal. Con los parámetros de adivinanza no se puede afirmar esto ya que la significancia obtenida en la prueba de normalidad para c ML3P EMB fue p=0,011 y para c ML3P MML p=0,000. Siendo así se procedió a comparar si existía relación por medio de la correlación de Pearson para los parámetros b y c, y por la correlación de Spearman para el parámetro c. Como se puede ver en la tabla 3 y en la tabla 4, todas las correlaciones entre los parámetros a y b son altas y significativas. Por su lado la correlación entre los parámetros c dadas las diferentes estimaciones fue de r_s=0,322 con p=0,095, razón por la cual se hizo la prueba Wilcoxon para corroborar si existía o no una diferencia significativa que dio un valor p=0,000.

Tabla 3: ***Correlación parámetros de dificultad***

		b ML1P EMB	b ML1P MML	b ML2P EMB	b ML2P MML	b ML3P EMB
b ML1P MML	**r de Pearson**	1,000				
	Sig. (2 colas)	,000				
b ML2P EMB	**r de Pearson**	,997	,997			
	Sig. (2 colas)	,000	,000			

[3] ML1P: Modelo logístico de un parámetro, ML2P: Modelo logístico de dos parámetros, ML3P:Modelo logístico de tres parámetros, MML: Máxima verosimilitud marginal, EMB: Estimación Bayesiana

b ML2P MML	**r de Pearson**	,996	,996	1,000		
	Sig. (2 colas)	,000	,000	,000		
b ML3P EMB	**r de Pearson**	,993	,994	,996	,997	
	Sig. (2 colas)	,000	,000	,000	,000	
b ML3P MML	**r de Pearson**	,996	,996	1,000	1,000	,997
	Sig. (2 colas)	,000	,000	,000	,000	,000

Tabla 4: ***Correlaciones parámetros de discriminación***

		a ML2P EMB	**a ML2P MML**	**a ML3P EMB**
a ML2P MML	**r de Pearson**	,998		
	Sig. (2 colas)	,000		
a ML3P EMB	**r de Pearson**	,959	,969	
	Sig. (2 colas)	,000	,000	
a ML3P MML	**r de Pearson**	,998	1,000	,969
	Sig. (2 colas)	,000	,000	,000

Discusión

Con los resultados obtenidos, en principio se observa que no hay gran variabilidad en los parámetros estimados para los diferentes modelos logísticos y métodos de estimación, aplicados a los datos recolectados, por esto la selección del modelo y método de estimación queda absolutamente mediada por los índices de bondad de ajuste que se obtengan, en este caso la decisión es fácil ya que solo el modelo logístico de 3 parámetros, utilizando el método de máxima verosimilitud marginal, fue el que se ajustó a todos los ítems del test, sin embargo si se presentan otras situaciones donde más de dos modelos diferentes se ajustan a los datos, ¿Qué decisión se debe tomar?. Si este fuera el caso, el investigador tendría la libertad de elegir con que información se queda, e ingresa en el dilema si optar por la simplicidad o la complejidad, como ya se ha visto en otros estudios, hay tendencias por optar por la simplicidad (Muñiz, 1997; Stage, 1998; Stage, 2003), aspecto que se ve en la preferencia por el modelo de RASCH. Sin embargo optar por este camino sacrifica información importante respecto a la discriminación y adivinanza del ítem, aspectos que pueden favorecer la permanencia o no permanencia del mismo en un test en proceso de construcción.

Además observando las altas correlaciones presentadas entre los parámetros de dificultad y discriminación, y siendo que no se presentan diferencias considerables, a pesar del desajuste que se dio en algunos modelos, entra a discusión si es posible tomar como parámetros de los ítems los parámetros estimados a pesar del desajuste, y si estos realmente representan las propiedades de interés del ítem, o si por el contrario existen falencias en los indicadores de bondad de ajuste que se han tenido en cuenta, y estos realmente no son los mejores y es necesario considerar un mayor número y cruzar la información obtenida para realizar una toma de decisión adecuada, cabe aclarar que estos índices no son satisfactorios totalmente (Muñiz, 1997). Con respecto al parámetro de adivinanza, que se observaron discrepancias significativas dados los métodos de estimación, probablemente este justifica la preferencia por los modelos de uno o dos parámetros, pues en la literatura es bien mencionado la alta dificultad para el cálculo y variabilidad del parámetro c (Abad, Olea, Ponsoda, & García, 2011; Lord, 1990), y observando los parámetros c estimados, se podría pensar que el método de estimación de Bayes cuenta con mucha más precisión pues los índices cuentan con un mayor rango que es lo que se esperaría del comportamiento de los ítems, mientras que con el método de máxima verosimilitud este parámetro es subvalorado y no brinda información de real utilidad, sin embargo para poder asegurar esto sería necesario realizar investigación con simulación.

En relación a la comprobación de supuestos, esto es una de las principales dificultades que se presentan para la aplicación de los modelos IRT, ya sea por la falta de acuerdo en el tipo de procedimientos que se deben seguir o por los índices que se han propuesto para corroborar tales supuestos, por parte del supuesto de la independencia local es un punto bastante débil que la comprobación de esta esté mediada por el modelo utilizado y los parámetros estimados, pues lo ideal es que esto no varíe de modelo a modelo como se pudo ver con los análisis realizados, que en algunos casos si se cumplía y en otros no, el asunto es que si esto puede variar de modelo a modelo no se podría considerar un supuesto, en un sentido estricto; y siendo así implica realizar mayor número de análisis hasta encontrar el modelo ideal, en este sentido ¿hasta qué punto se podría considerar la independencia local más bien como otro índice de bondad de ajuste?.

Para finalizar, teniendo en cuenta que la medición en psicología y en ciencias sociales en general, es tan variable en cuanto a los constructos medidos y las formas para medirlos, es importante clarificar que métodos son los más adecuados y convenientes para el análisis de los ítems, ya que esto brinda información relevante acerca de la pertinencia y validez de las mediciones, siendo así que si fuera posible determinar cuándo es mejor un modelo u otro, y cuando es mejor un método de estimación u otro, dependiendo de la cantidad de ítems del test, y de los tamaños de la muestra o de las características del test por la puntuación en los ítems y si este pretende abarcar un amplio rango de habilidad o más bien un rango restringido, esto economizaría considerablemente la cantidad de análisis que se le hacen a los datos y facilitaría la toma de decisión que realiza el investigador con respecto a los ítem y a los test. Aunque las ventajas que trae la IRT son evidentes respecto a las limitaciones que existían en la TCT, es evidente que las dos teorías son complementarias como lo evidencian diferentes estudios que realizan sus análisis bajo ambos puntos de vista, y como lo expresan Manzi & San Martín (2003), aun así la IRT no está fuera de crítica teniendo en cuenta que a pesar de haber revolucionado el acercamiento al análisis de los test psicológicos, fue un cambio en el ámbito metodológico, por lo cual los aspecto epistemológicos de como se conoce el fenómeno, queda a merced de las críticas que ha tenido constantemente la medición y la evaluación en las ciencias sociales.

Para futuros estudios, sería recomendable realizar los análisis a datos simulados donde se haga control de diferentes variables, que ya se han mencionado, además de evaluar las implicaciones que trae no satisfacer los supuestos de los modelos IRT, de tal forma que se pueda responder a la pregunta de cuándo es mejor uno u otro modelo, y de los métodos de estimación. En conclusión el tema es suficientemente amplio, además cuando se toman en cuenta otras aplicaciones como los estudios de Funcionamiento Diferencial de los Ítems y Equiparación. Una de las principales limitaciones fue el tamaño de la muestra que hay una exigencia desde la IRT por grandes muestras y que esta probablemente estuviera sesgada al ser aplicada población de contexto clínico.

Referencias Bibliográficas

Abad, F. J., Olea, J., Ponsoda, V., & García, C. (2011). *Medición en Ciencias Sociales y de la Salud.* Madrid: Sintesis.

Abad, F. J., Olea, J., Real, E., & Ponsoda, V. (2002). Estimación de habilidad y presicion en test adaptativos informatizados y test optimos: un caso práctico. *Revista Electrónica de Metodología Aplicada, 7*(1), 1-20.

Burga León, A. (2012). La unidimensionalidad de un instrumento de medición: perspectiva factorial. *Revista de Psicología, 24*(1), 53-80.

Chen, W.-H., & Thissen, D. (1997). Local Dependence Indexes for Item Pairs Using Item Response Theory. *Journal of Educational and Behabioral Statistics, 22*(3), 265-289.

DeMars, C. E. (2008). *Scoring Multiple Choice Items: A Comparison of IRT and Classical Polytomous and Dichotomous Methods.* In annual meeting of the National Council of Measurement in Education, James Madison University, New York.

Ferrando, P. J. (1996). Evaluación de la unidimensionalidad de los ítems mediante análisis factorial. *Psicothema, 8*(2), 397-410.

Guillen, V., Verdugo, M. A., Arias, B., Navas, P., & Vicente, E. (2012). *Desarrollo de la escala SIS para niños y adolescentes. Resultados y conclusiones preliminares.* VIII Jornadas científicas de investigación sobre personas con discapacidad: simposios, comunicaciones y posters. Cambio organizacional y apoyo a las graves afectaciones: dos prioridades.

Hambleton, R. K., Swaminathan, H., & Rogers, H. J. (1991). *Fundamentals of Item Response Theory.* California: Sage Publications Inc.

Iraurgi, I., Lozano, O., Gónzalez-Saiz, F., & Trujols, J. (2008). Valoración psicométrica de la Escala de Sevaridad de la Dependencia a partir de dos modelos de análisis: la Teoría Clásica de los Test y la Teoría de Respuesta al Ítem. *Boletin de Psicología*(93), 41-57.

Kinsey, T. L. (2003). *A Comparison of IRT and Rasch Procedures in a Mixed-Item Format Test.* Doctoral dissertation , University of North Texas.

Lord, F. M. (1990). *Applications of Item Response Theory to Practical Testing Problems.* New Jersey: Lawrence Erlbaum Associates Publishers.

Manzi, J., & San Martín, E. (2003). La necesaria complementariedad entre Teoría Clásica de la Medición (TCM) y Teoría de Respuesta al Item (IRT): Aspectos conceptuales y aplicaciones. *Estudios Públicos, 90*, 145-183.

Muñiz, J. (1997). *Introducción a la Teoría de Respuesta al Itém.* Madrid: Ediciones Piramide S.A.

Rodríguez, O. R., Casas, P. P., & Medina, Y. (2005). Analisis psicométrico de los examenes de evaluación de la calidad de la educación superior (ECAES) en Colombia. *Avances en MEdición, 3*, 153-172.

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

Stage, C. (2003). Classical Test Theory or Item Response Theory: the swedish experience. *UMEA Universitej, 42*.

Wechsler, D. (2005). *WISC IV: Escala Wechsler de Inteligancia para Niños: Manual técnico.* Manual Moderno.