Escritos de Psicología, Vol. 4, nº 3, pp. 36-43
Septiembre-Diciembre 2011

ISSN 1989-3809 DOI: 10.5231/psy.writ.2011.1209

Modelización de una Prueba de Analogías Figurales con la Teoría de Respuesta al Ítem

Modelling Figural Analogies Test with the Item Response Theory

G. Diego Blum, María Silvia Galibert, Facundo J. P. Abal, Gabriela S. Lozzia y Horacio F. Attorresi

Instituto de Investigaciones de la Facultad de Psicología de la Universidad de Buenos Aires, Argentina

Disponible online 31 de diciembre de 2011

Se detallan las propiedades psicométricas de una Prueba de Analogías Figurales desde el marco de la Teoría de Respuesta al Ítem. Se elaboraron 36 matrices de figuras de 2x2 utilizando reglas de emplazamiento, distorsión y número. Se contó con una muestra de 499 estudiantes de Psicología de la Universidad de Buenos Aires, 79% de los cuales fueron mujeres. Se utilizó el Modelo Logístico de 3 Parámetros logrando un ajuste global altamente satisfactorio al 5% (p = .47). Sólo 3 ítems del total no ajustaron al modelo. Existe una buena potencia discriminatoria general (a: M = 1.02; DT = .33), un nivel de dificultad medio (b: M = -.03; DT = .63) y un nivel de acierto por azar ligeramente inferior a lo esperable con 6 alternativas de respuesta (c: M = .14; DT = .05). Se discuten las condiciones para modelizar la Prueba y posibles desventajas del presente estudio.

Palabras clave: Analogías; Matrices; Teoría de Respuesta al Ítem.

The psychometric properties of a Figural Analogies Test are described within the framework of Item Response Theory. Thirty-six 2x2 matrix figures were constructed by using location, distortion and number rules. The sample included 499 psychology students from the University of Buenos Aires, 79% of whom were women. The 3-Parameter Logistic Model was used obtaining a highly satisfactory global fit at 5% (p = .47). Only 3 items did not fit the model. It had good overall discriminatory power (a: M = 1.02, SD = .33), a medium level of difficulty (b: M = -.03, SD = .63) and the c level was slightly lower than expected with six possible answers (c: M = .14, SD = .05). The conditions for modelling the test and possible disadvantages of the present study are discussed.

Key Words: Analogies; Matrices; Item Response Theory.

Correspondencia: Lic. G. Diego Blum. Anchorena 1.169 3º B (1425), Capital Federal, Argentina. E-mail: blumworx@gmail.com. E-mail de los otros autores: María Silvia Galibert: galibert@psi.uba.ar., Facundo J. P. Abal: fabal@psi.uba.ar, Gabriela S. Lozzia: glozzia@psi.uba.ar, Horacio F. Attorresi: horacioattorresi@fibertel.com.ar, hattorre@psi.uba.ar.
Fuente de Financiación de la Investigación: Trabajo realizado en el marco del subsidio UBACyT P043 y Nro. 20020100100346

Las técnicas psicométricas son útiles para operacionalizar de manera cuantificable constructos psicológicos. Dentro de la Psicometría, las *Teorías de los Tests* sustentan las bases sobre las cuales se construyen comúnmente dichas pruebas. En la actualidad las dos teorías psicométricas más utilizadas son la Teoría Clásica de Tests (TCT) y la Teoría de Respuesta al Ítem (TRI) (Martínez-Arias, 1995).

La mayoría de los tests validados responden a los requerimientos de la TCT. Spearman (1904, 1907, 1913) fue el primero en formularla utilizando el Modelo Lineal de Puntuaciones, según el cual la puntuación observada del individuo resulta de la suma de su puntuación verdadera más un error de medida. Si bien la TCT representó desde principios de Siglo XX un enfoque simple y práctico para estudiar la calidad de la medición, se basa en supuestos débiles dado que toma indicadores demasiado generales (Muñiz, 1994). Sus principales inconvenientes, apuntados ya en 1928 por Thurstone, son que las propiedades estudiadas varían tanto en función de la prueba utilizada para medir el constructo como de la muestra de individuos recolectada.

En este sentido, el enfoque de la TRI permitió superar estas limitaciones y complementar, aunque no reemplazar, a la TCT. Surgida a partir de los trabajos de Rasch (1960) y Birnbaum (1968), la TRI se compone de una serie de modelos que parten del supuesto de que la respuesta de un individuo a un ítem puede predecirse y explicarse a partir de una variable inobservable, a saber, el rasgo latente (Lazarsfeld, 1950; véase también en Embretson, 1983). El objetivo fundamental de la TRI es la construcción de instrumentos de medición con propiedades invariantes entre poblaciones. La TRI permite expresar las propiedades del test en función de la aditividad de las propiedades de los ítems que lo componen.

Los modelos de la TRI se desarrollaron tradicionalmente en el ámbito de las pruebas de rendimiento máximo como son los test de inteligencia, aptitudes y rendimiento educativo (Martínez-Arias, Hernández-Lloreda y Hernández-Lloreda, 2006). Es en estas pruebas donde mayor aplicación ha tenido la TRI hasta nuestros días. La respuesta al ítem en dichos tests suele ser dicotómica (respuesta correcta o incorrecta, sin importar la cantidad de alternativas presentes). La TRI asume que la probabilidad de acertar el ítem se establece en función del nivel en la escala de aptitud. Dicha relación funcional se denomina Curva Característica del Ítem (CCI). Cada CCI queda determinada por sus propios parámetros y es independiente de las CCI de los demás ítems y de la distribución del rasgo latente en la población de individuos que sirvieron para estimarlos. Por consiguiente, la medición de un constructo se independiza tanto del conjunto de ítems administrado (test) como de las muestras estudiadas, superando las dificultades de la TCT.

La TRI se ha aplicado de manera extensa a la medición de habilidades unidimensionales. Cuando se trabaja con datos dicotómicos, se utilizan comúnmente los Modelos Logísticos de uno, dos y tres Parámetros. El primero de estos modelos, también conocido como Modelo de Rasch (1960), fue ampliado por Fischer (1973) y colaboradores para dar origen al Modelo Logístico Lineal de Rasgo Latente (LLTM). El LLTM descompone la dificultad del ítem en una suma de efectos debidos a diversas fuentes de dificultad. En esta línea además se sitúan los desarrollos de Embretson (1984, 1991) con el Modelo Multicomponente de Rasgo Latente (MCTM). Tanto LLTM como MCTM suelen emplearse para determinar aquellos procesos cognitivos que explican mejor las respuestas a un ítem.

Existen muchas pruebas de razonamiento e inteligencia en la actualidad, sin embargo la medición del Razonamiento Analógico (RA) ocupa un lugar destacado ya que ha sido descrito de manera frecuente como un componente clave de la capacidad intelectual (Cattell, 1971; Spearman, 1904; Sternberg, 1987). Entre los diversos tipos de estímulos encontrados en pruebas de analogías, los tests de figuras abstractas ocupan uno de los lugares más destacados. Es muy común elaborar reactivos de este estilo con ayuda del modelo matricial, como sucede por ejemplo en el Test de Matrices Progresivas de Raven (Raven, Court y Raven, 1993), en el Test de Factor G de Cattell y Cattell (1997), en el Test de Inteligencia No-Verbal versión 2 (Test of Non-Verbal Intelligence 2, TONI 2) de Brown, Sherbenou y Johnsen (2000) y en los trabajos de Wolf Nelson y Gillespie (1991), aunque existen también otros modelos posibles. Todas las pruebas mencionadas han sido modelizadas con la TCT mientras que en algunas se estudió además el ajuste de los modelos de la TRI a los datos. Por ejemplo, Raven, Raven y Court (1991) emplearon el examen visual de las CCI para proporcionar información sobre la naturaleza de las aptitudes evaluadas y sobre el posible perfeccionamiento del test.

El RA es un razonamiento no-deductivo basado en la generación de inferencias sobre objetos poco conocidos partiendo de su comparación con objetos similares y mejor comprendidos. La función clave es la *extensión* de características que parten del análogo-fuente y se dirigen hacia el análogo-meta para añadir información sobre este último (Cubillo y González-Labra, 1998; Rivera, 2000; Sternberg, 1977). En aquellas analogías que definen relaciones de proporción entre elementos (analogías A:B::C:D), los dominios A:B y C:D comparten un número de roles que hacen posible extrapolar determinadas relaciones desde el análogo-fuente A:B hacia el análogo-meta C:D (Blum, Abal, Lozzia, Picón-Janeiro y Attorresi, 2011).

La educción de relaciones y de correlatos entre relaciones juegan un papel crucial en las analogías de estilo A:B::C:D (Spearman, 1923). La presencia implícita de una relación o grupo de relaciones que es correlativa a otra relación o grupo puede entenderse como una *regla* de resolución. Numerosos autores han destacado el rol de reglas puntuales en la resolución de matrices y/o ítems de analogías (e.g. Brown et al., 2000; Freund, Hofer y Holling, 2008; Whitely y Schneider, 1981). Blum et al. (2011) propusieron sugerencias para la construcción de ítems que evalúan el RA utilizando matrices figurales de 2x2. Una de dichas sugerencias es el empleo de reglas de emplazamiento

espacial (rotación, traslación, reflejo), reglas de distorsión (tamaño, forma) y reglas de número (adición, sustracción).

Desde la perspectiva de Blum et al. (2011), el ítem debería permitir que la misma regla o grupo de reglas pueda abordarse tanto a través de la comparación vertical entre relaciones horizontales (A:B con C:D) como por medio de la comparación horizontal entre relaciones verticales (A:C con B:D). Otro punto destacado fue la necesidad de evitar sesgos basados en respuestas que privilegien formas de resolución diferentes del RA. Se propuso la elaboración de distractores que se parezcan entre ellos y a la respuesta correcta, así como intentar graduar la dificultad considerando que ésta debería aumentar conforme crece el número de reglas en un ítem.

El objetivo de esta investigación es presentar los resultados de la modelización de una Prueba de Analogías Figurales desde la TRI. Si bien el foco está puesto sobre este último enfoque psicométrico, se estudian además los índices tradicionales de fiabilidad y unidimensionalidad desde la TCT para complementar el análisis.

Método

Participantes

Participaron 499 cursantes del primer año del Ciclo General de la carrera de Licenciatura en Psicología de la Universidad de Buenos Aires (UBA). El 21% del total de individuos fueron varones mientras que el 79% fueron mujeres. La edad varió entre 18 y 56 años, con una media de 21.98, una mediana de 20 y una desviación típica de 5.40. Se les informó a los individuos sobre el carácter voluntario de su colaboración y que su calificación en la asignatura que se encontraban cursando no se vería comprometida. Se han seguido las normas éticas pertinentes al tipo de procedimiento y población.

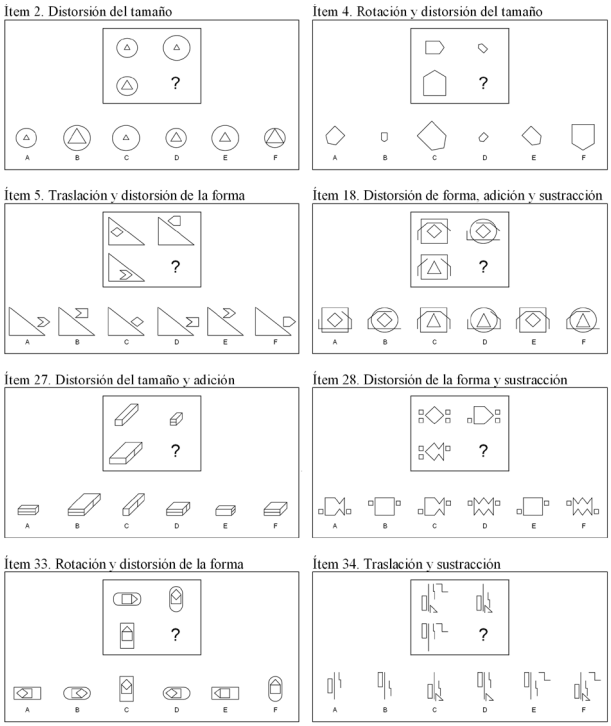
Material y procedimiento

Se utilizaron las sugerencias mencionadas en la introducción (Blum et al., 2011), con el fin de elaborar una Prueba de Analogías Figurales. Se recurrió al modelo matricial de 2x2 para el diseño de reactivos. La figura inferior derecha de cada matriz fue reemplazada con un signo de pregunta, marcando de este modo el problema principal a resolver. Se ofrecieron seis opciones de figuras para completar dicha falta, solo una de las cuales es correcta en función del empleo del RA. Las reglas utilizadas para los reactivos del protocolo fueron rotación, traslación, distorsión del tamaño, distorsión de la forma, adición y sustracción.

Numerosos autores (e.g. Embretson y Reise, 2000; Freund, Hofer y Holling, 2008; Mulholland, Pellegrino y Glaser, 1980) han destacado que el aumento de la cantidad de reglas produce un incremento de la dificultad del ítem, lo cual guarda una relación directa con la memoria de trabajo. Se construyeron ítems con una, dos y tres reglas, bajo el supuesto de que los ítems con una regla deberían ser fáciles de resolver, los que poseen dos reglas tendrían una dificultad mayor o media, y los de tres reglas serían los más difíciles. Con el objetivo de privilegiar la presencia de una dificultad media y aumentar así la varianza del test (García-Cueto y Fidalgo, 1995), la mayoría de los ítems de la prueba (21 de ellos) se desarrollaron con dos reglas mientras que 9 poseyeron una sola regla y 6 tuvieron tres reglas, conformando un total de 36 reactivos. La Figura 1 muestra ocho de los reactivos del protocolo. Se desarrolló además una consigna inicial con tres ejemplos de ítems resueltos y por resolver. Cada uno de los mismos contó con una de las tres reglas siguientes: reflejo, sustracción o distorsión de la forma.

Se controló el efecto de la fatiga sobre las respuestas (Pereda, 1987) por medio de la elaboración de seis pruebas de distinto orden de los reactivos. Para alterar dicho orden se rotaron grupos de seis ítems. Cada grupo poseyó ítems con una, dos y tres reglas. Se repartió a cada individuo una de dichas pruebas y se los invitó a contestarla durante el transcurso del espacio de clases. Se les pidió que contesten en forma absolutamente individual.

Figura 1
Ejemplos de ítems de la Prueba.



Análisis de los Datos

La muestra fue depurada según los criterios siguientes. Dado que la puntuación total no debería variar en función del tiempo empleado para concretar la tarea, se eliminaron las respuestas de 12 personas quienes invirtieron poco tiempo (menos de 22 minutos) en la Prueba y a su vez obtuvieron una puntuación total baja ($M = 6.25$; $DT = 3.28$). Esto último sugiere que ocurrió un sacrificio de la precisión de respuesta en función de la economización del tiempo propio. La muestra original que incluía

a estos 12 protocolos experimentó una correlación media-baja entre el tiempo total y la puntuación total (r = .27; p < .001), mientras que sin los mismos la correlación bajó considerablemente (r = .17; p < .01). También se descartaron 9 protocolos que poseían más de 10 ítems sin responder y otros 3 por no contestar los últimos 5 ítems, sugiriendo que los individuos no terminaron la tarea a tiempo. En total 24 registros fueron eliminados, conformando una muestra depurada de 475 individuos.

Para estudiar la calidad psicométrica de la Prueba desde la TCT se obtuvieron el coeficiente de consistencia interna α de Cronbach, las correlaciones item-test corregidas y los índices de dificultad (ID). Con estos fines se utilizó el Statistical Package for Social Sciences (SPSS), versión 15. También se estudió la unidimensionalidad por medio de un análisis factorial de componentes principales. En este caso se trabajó con la matriz de correlaciones tetracóricas dada la naturaleza dicotómica de las respuestas (García-Cueto y Fidalgo, 1995). El análisis factorial se realizó tanto con MicroFact 1.1 (Waller, 1995) como con el uso conjunto de SPSS y TetCorr 2.1 (Enzmann, 2005). Se adoptaron como criterios para corroborar la unidimensionalidad un porcentaje de varianza total explicada por el primer autovalor igual o mayor a 40% (Carmines y Zeller, 1979) y una razón del primer autovalor al segundo igual o mayor a 5 (Martínez -Arias, 1995).

Se calculó la t de Student para detectar diferencias entre grupos y el tamaño del efecto de dichas diferencias considerando la fórmula que se muestra a continuación (Coe y Merino, 2003). Cohen (1988) tomó en cuenta tamaños del efecto pequeños, moderados y grandes según valores cercanos a .20, .50 y .80 respectivamente.

$$TE = \frac{\bar{x}_{G1} - \bar{x}_{G2}}{DS_{común}} \quad \text{Donde:} \quad DS_{común} = \sqrt{\frac{(N_{G1}-1)DS_{G1}^{\ 2} + (N_{G2}-1)DS_{G2}^{\ 2}}{N_{G1}+N_{G2}-2}}$$

En la modelización psicométrica con la TRI se aplicó el Modelo Logístico de 3 Parámetros (ML3P) sobre los 36 ítems mediante el programa BILOG-MG (Zimowski, Muraki, Mislevy y Bock, 1996). La formulación del ML3P es la siguiente:

$$P_i(\theta) = c_i + \frac{1 - c_i}{1 + \exp[-1.7a_i(\theta - b_i)]} \qquad \theta \in \mathbf{R}$$

donde:

θ es el rasgo latente que se desea medir con el ítem i.

$P_i(\theta)$ es la probabilidad de respuesta correcta al ítem i para un nivel dado de θ.

b_i es el índice de dificultad del ítem i.

a_i es el índice de discriminación del ítem i.

c_i es el índice de acierto por azar del ítem i.

La estimación de los parámetros para cada ítem se realizó utilizando el algoritmo de máxima verosimilitud marginal. BILOG también ofreció indicadores para evaluar el ajuste global del modelo a los datos. Además se examinó la bondad de ajuste ítem a ítem mediante un estadístico basado en la distribución de χ^2 que contrasta las frecuencias observadas con las esperadas por el modelo analizando las discrepancias en distintos niveles del rasgo latente.

Asimismo, se calculó la Función de Información del Test junto con el Error Estándar de estimación, las cuales constituyen herramientas de análisis útiles para estudiar la precisión de la medida para cada nivel del rasgo latente.

Resultados

Se confirmó la alta consistencia interna del instrumento (α = .91). Las correlaciones ítem-test corregidas puntuaron en su totalidad por encima de .30. Los ID variaron entre .31 y .91. Se comprobó además la unidimensionalidad por medio de los criterios citados. Sin efectuar rotación, el primer autovalor puntuó 14.17, que corresponde a un 39.37% de la varianza total (aproximado a 40%). El segundo autovalor puntuó 1.69, con lo cual $\lambda_1 / \lambda_2 = 8.38$ (mayor que 5).

Así como los Modelos de 1 y 2 Parámetros, el ML3P tiene dos supuestos fundamentales que condicionan su aplicación, a saber, que el rendimiento de los individuos es explicado por un factor dominante (supuesto de unidimensionalidad) y que no existe relación entre las respuestas de examinados a diferentes ítems cuando se mantiene constante la aptitud (supuesto de independencia local) (Martínez-Arias, 1995). Respecto de este último supuesto, Lord y Novick (1968) demostraron que se deduce la existencia de independencia local si se cumple la unidimensionalidad del espacio latente. En este sentido, los resultados antes descritos sugieren que ambos supuestos se confirman satisfactoriamente.

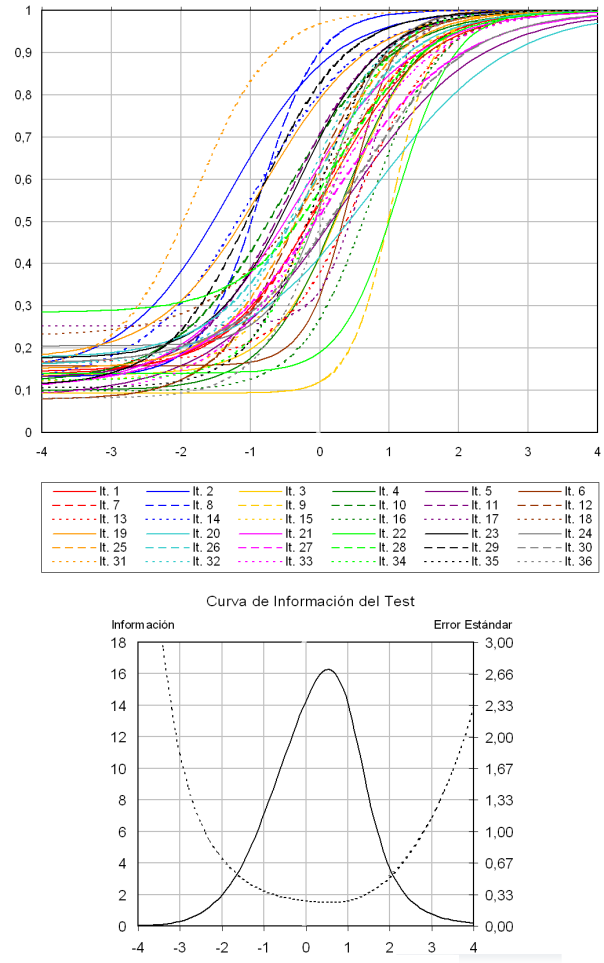
En la modelización con el ML3P se alcanzó un criterio de convergencia de .0001 (Largest Change = .00007) y no se rechazó el ajuste global al 5% (χ^2 = 276.8; p = .47). La Tabla 1 muestra tanto las estimaciones de los parámetros como la prueba χ^2 de cada ítem. En cuanto al ajuste individual, puede observarse un rechazo al 5% en los ítems 5, 11 y 18. Las medias y desviaciones estándar promedio de las estimaciones de los parámetros dan cuenta de una buena potencia discriminatoria general (a: M = 1.02; DT = .33), un nivel de dificultad medio (b: M = -.03; DT = .63) y un nivel de acierto por azar ligeramente inferior a lo esperable con seis alternativas de respuesta (c: M = .14; DT = .05). La Figura 2 muestra las CCI de los ítems y la Función de Información del Test, obtenidas con los parámetros de este modelo logístico. La Prueba da su mayor información en torno a un θ = 0.53. BILOG proporcionó además un índice de confiabilidad calculado como la razón entre la varianza de las puntuaciones estimadas θ y la varianza de las puntuaciones observadas, el cual resultó de .92, muy similar al obtenido con el α de Cronbach de la TCT.

La correlación entre los ID y los b brinda un valor de -.97, corroborando una asociación elevada y negativa. Esto último es un dato esperable, ya que una puntuación alta del primer índice da cuenta de la facilidad del reactivo en la TCT mientras que una alta de b expresa lo contrario en la TRI. Los valores mínimos y máximos del parámetro b son -1.77 y 1.13 mientras que los de a son .60 y 1.95.

Tabla 1
Parámetros de los ítems y pruebas de χ^2 para la bondad de ajuste.

Ítem	a	b	c	χ^2 (*p*)	*DF*
1	.81	.08	.14	8.5 (.49)	9
2	.79	-1.29	.14	5.8 (.44)	6
3	1.08	.40	.14	3.0 (.96)	9
4	1.05	.33	.10	8.9 (.45)	9
5	.61	.34	.08	24.5 (.004)	9
6	1.81	.46	.16	7.4 (.28)	6
7	.88	.06	.14	3.2 (.95)	9
8	1.40	-.86	.13	2.2 (.70)	4
9	1.95	1.05	.09	6.5 (.60)	8
10	.75	-.53	.12	11.5 (.18)	8
11	.91	-.44	.13	16.8 (.01)	6
12	.98	-.25	.08	7.5 (.49)	8
13	1.03	.62	.17	6.8 (.66)	9
14	.75	-.98	.12	8.5 (.29)	7
15	.93	.08	.12	7.4 (.50)	8
16	1.16	.73	.10	5.8 (.76)	9
17	1.78	.67	.25	7.4 (.49)	8
18	.74	.06	.23	20.5 (.01)	8
19	.79	-.81	.17	4.2 (.76)	7
20	.60	.78	.16	2.6 (.98)	9
21	.71	-.31	.10	8.7 (.47)	9
22	1.44	1.13	.14	4.5 (.87)	9
23	.97	-.33	.18	1.4 (.98)	7
24	1.39	.14	.21	3.7 (.72)	6
25	.94	-.12	.15	7.0 (.54)	8
26	.84	-.10	.18	6.0 (.65)	8
27	.64	.15	.11	13.3 (.15)	9
28	.87	.25	.28	7.8 (.46)	8
29	.93	-.90	.11	10.2 (.11)	6
30	.71	.45	.16	5.9 (.75)	9
31	1.07	-1.77	.15	2.7 (.44)	3
32	1.01	-.21	.16	7.3 (.40)	7
33	.77	.06	.09	16.5 (.06)	9
34	1.19	.03	.12	2.8 (.91)	7
35	1.21	-.06	.11	2.8 (.90)	7
36	1.19	.11	.08	7.3 (.51)	8

Figura 2
CCI de los ítems y Función de Información del Test.



Los promedios de los *b* para ítems con una, dos y tres reglas son -.48, .11 y .16 respectivamente. La diferencia entre los primeros dos promedios resultó significativa al 5% (t = -2.345; p = .026), con un tamaño del efecto de .93. Según Cohen (1988), este último corresponde a un efecto grande. Por otro lado, la diferencia entre el segundo y el tercer promedio no fue significativa al mismo nivel (t = -.204; p = .84).

Discusión

Los indicadores globales que brinda la TCT son adecuados para evaluar la calidad de la medición mediante tests, pero los avances actuales de la Psicometría ofrecen la posibilidad de realizar un análisis más exhaustivo y enriquecedor de los ítems. Cobran cada vez mayor importancia las evidencias de validez interna-estructural y las medidas de precisión provenientes de fuentes que toman a los ítems como unidad de análisis (Elosua, 2003). En este sentido, la TRI ofrece un análisis exhaustivo en términos del estudio y modelización de escalas psicométricas.

Los resultados descritos sugieren que todos los ítems poseen propiedades psicométricas altamente satisfactorias desde el marco de la TCT mientras que 33 del total de 36 ítems se ajustan al ML3P desde el marco de la TRI. Esto corrobora la efectividad de utilizar las sugerencias descritas por Blum et al. (2011) para la construcción de ítems de analogía figural con características psicométricas apropiadas.

La razón del desajuste de los reactivos 5, 11 y 18 se debe a que la CCI empírica no es siempre creciente en cierto intervalo de habilidad. Esto último pudo verificarse en los plots obtenidos con BILOG. Por consiguiente, los resultados del análisis desde la TRI sugieren tomar alguna decisión sobre estos ítems, ya sea eliminándolos o modificándolos. Debido a que son pocos

ítems los que desajustan, la posibilidad de su eliminación no comprometería de manera importante la validez de contenido de la Prueba. Sin embargo, dichos reactivos presentan buenos índices desde la TCT. Como se expresó líneas arriba, esto se debe a que la TCT proporciona índices globales mientras que la CCI provee información con respecto a cada nivel de habilidad.

A continuación se discutirán algunas características inherentes a los parámetros estimados. En primer término, el rango de valores que adopta *a* es aceptable así como su promedio. Del total de reactivos, quince poseen un $a > 1$, diez poseen un *a* entre .80 y 1, y once poseen un $a < .80$. Dos de los tres ítems que no ajustan al modelo (los reactivos 5 y 18) poseen niveles de *a* dentro de este último rango indicando que su capacidad discriminatoria es inferior, lo cual brindaría una razón alternativa para eliminar dichos ítems, mientras que el ítem 11 con un $a = .91$ podría simplemente modificarse.

Analizando la curva de Función de Información del Test, puede apreciarse que la Prueba proporciona su máxima información cuando la habilidad de los individuos tiende a ser media. En términos teóricos, esta curva se aproxima bastante a una curva esperable, es decir, una curva simétrica respecto de $\theta = 0$ y una oscilación de la habilidad entre -3 y +3. Sin embargo, el rango de valores en que fluctúa *b* sugiere que los reactivos no poseen una dificultad muy reducida ni muy elevada. Si el objetivo futuro es confeccionar un banco de ítems que evalúe el RA en toda su complejidad, se deberían construir reactivos muy fáciles y muy difíciles que se añadan a los vigentes. Investigaciones similares como la de Embretson y Reise (2000) confirman que dicha construcción es posible. La autora presentó los resultados de la modelización de 30 ítems de Razonamiento Abstracto con el ML3P, cuyos *b* fluctúan entre -2.81 y 3.46.

En relación con una de las sugerencias citadas en Blum et al. (2011) y según los hallazgos de otros autores que trabajaron con ítems de figuras (Embretson y Reise, 2000; Freund et al., 2008; Mulholland et al., 1980), se esperaba graduar la dificultad conforme aumenta el número de reglas de resolución. Los resultados revelan que existe una diferencia importante entre el promedio de los *b* del grupo de ítems con una regla y el promedio de los *b* del grupo con dos reglas, mientras que la diferencia es despreciable cuando dicha comparación se efectúa entre el grupo con dos reglas y el grupo con tres reglas. Además, la elaboración de ítems con tres reglas ha resultado una tarea sumamente compleja. Lo cual corrobora el hecho de que construir este último estilo de reactivos con la intención de graduar la dificultad es una tarea infructuosa considerando las condiciones de diseño establecidas. Por consiguiente, se evaluará la pertinencia de modificar aquellos ítems con tres reglas para que adopten sólo dos.

Una cuestión llamativa es que el parámetro de acierto por azar de los ítems 17, 18, 24 y 28 puntúa por encima de .20 cuando lo esperable para un reactivo con seis alternativas es alrededor de .17. Esto tal vez se debe a que alguna alternativa de respuesta de dichos reactivos resultó excepcionalmente despreciable, lo cual redujo la decisión a menos de 6 opciones, elevando así la puntuación de *c*. Además, estos ítems poseen la particularidad de que sus opciones de contestación son algo más diferentes entre sí que las de otros ítems, lo cual promueve el descarte de aquellas opciones muy distintas de la correcta. Se revisarán las alternativas de respuesta de dichos reactivos para evaluar la posibilidad de hacer más *atractivos* a aquellos distractores contestados con menor frecuencia. Sin embargo, tal vez convenga eliminar directamente el ítem 18 o a lo sumo establecer una fuerte modificación del mismo, debido a que posee las otras desventajas mencionadas: no ajusta al ML3P, su *a* es reducido y fue ideado con tres reglas.

En conclusión, los resultados corroboran la adecuación de las decisiones efectuadas para la construcción de reactivos de analogía figural. Sin embargo, en el presente estudio existen limitaciones potenciales relacionadas con la muestra de individuos recolectada. Dado que sólo el 20% de la misma fue de sexo masculino, sería necesario un aumento del tamaño de dicho grupo para otorgar mayor representatividad al estudio. El tamaño de la muestra podría considerarse reducido según determinados autores que, para un ML3P, recomiendan un $n = 1000$ (e.g. Hanson y Beguin, 2002; Yen, 1987). La intervención de la velocidad en las respuestas constituye otra limitación potencial, ya que si bien se consiguió una correlación baja mediante la reducción de la muestra, el *r* de Pearson continúa siendo significativo al 1%. Dado que la ausencia de velocidad como determinante de las respuestas es también un supuesto importante de la TRI (Martínez-Arias, Hernández-Lloreda y Hernández-Lloreda, 1996), se realizarán estudios a futuro que permitan controlar mejor dicha variable.

Las Figuras 3 y 4 muestran una posible Forma Revisada de la Prueba, que contiene 15 ítems modificados y 21 ítems sin modificar respecto de la versión utilizada para este trabajo. Los criterios de revisión fueron los siguientes:

1. Cantidad de reglas: se modificaron los ítems 6, 12, 18, 24, 30 y 36 para que adopten dos reglas en lugar de tres.

2. Redistribución de ítems: los ítems antes nombrados fueron redistribuidos entre sí para que no compartan reglas con los ítems contiguos. Con este fin también se redistribuyeron entre sí los ítems 13, 19 y 31. La Forma Revisada contiene la numeración final de los reactivos. El propósito de dicha acción es disminuir el riesgo de que una de las reglas aprendidas en un ítem sirva para resolver el ítem siguiente.

3. Modificación de ítems que no ajustan al ML3P: se modificó la estructura de dos de los tres reactivos que no ajustaron al modelo logístico. El ítem 5 no fue modificado ya que se desea investigar si en una próxima toma dicho reactivo genera ajuste. Por la misma razón no se eliminó ningún ítem de la Prueba.

4. Modificación y/o redistribución de distractores: los distractores de los ítems 1, 3, 6, 9, 11, 12, 13, 17, 18, 19, 20, 24, 28, 30 y 36 que aparecen en las Figuras 3 y 4 han sufrido algún tipo de intervención según poseyeran distribuciones asimétri-

Figura 3
Primeros 18 ítems de la Forma Revisada de la Prueba de Analogías Figurales.

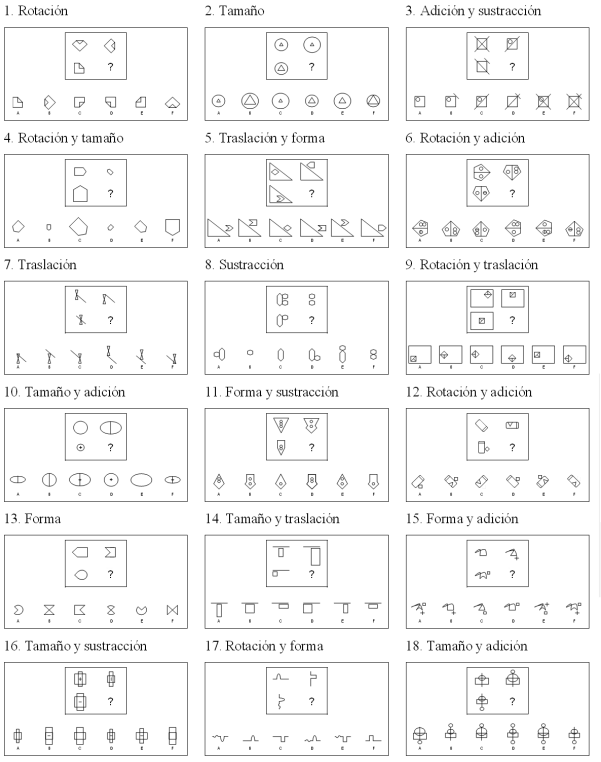
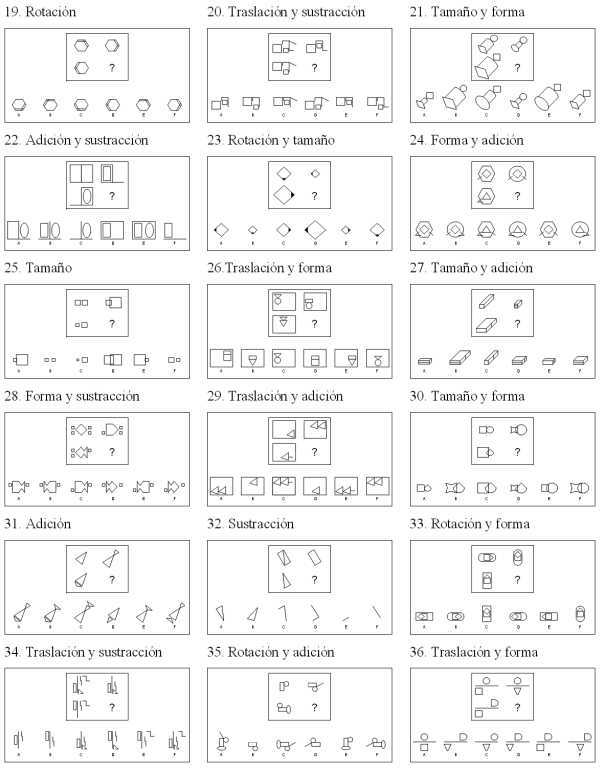


Figura 4
Últimos 18 ítems de la Forma Revisada de la Prueba de Analogías Figurales.



cas en la frecuencia de respuesta, un parámetro *c* inusualmente elevado y/o permitieran generar agrupaciones entre alternativas más parecidas entre sí.

Si se mantiene esta revisión, dichos reactivos serán administrados a una nueva muestra para averiguar sus características psicométricas y además se estudiará el Funcionamiento Diferencial de los Ítems. De este modo se espera realizar una contribución importante a la mejoría de la calidad psicométrica de la Prueba, dejándola lista para su uso en los distintos ámbitos de aplicación de la Evaluación Psicológica.

Referencias

1. Birnbaum, A. (1968). Some latent trait models and their use in inferring an examinee's ability. En F. M. Lord & M. R. Novick (Eds.). *Statistical Theories of Mental Test Scores* (pp. 397-479). Reading, MA: Addison Wesley.
2. Blum, G.D.; Abal, F.J.P.; Lozzia, G.S.; Picón Janeiro, J.C. & Attorresi, H.F. (2011). Analogías de figuras: Teoría y construcción de ítemes. *Interdisciplinaria. Revista de psicología y ciencias afines, 28 (1)*, 131-144.
3. Brown, L., Sherbenou, R.J. & Johnsen, S.K. (2000). *TONI 2. Test de Inteligencia No Verbal. Apreciación de la habilidad cognitiva sin influencia del lenguaje. Manual*. Madrid: TEA.
4. Carmines, E.G. & Zeller, R.A. (1979). *Reiability and validity assessment*. Londres: Sage.
5. Cattell, R.B. (1971). *Abilities: Their structure, growth and action*. Boston: Houghton Mifflin.
6. Cattell, R.B. & Cattell, A.K.S. (1997). *Factor "g" 2 y 3. Manual*. Adaptación española: A. Cordero, M.V. De la Cruz, M. González & N. Seisdedos. Madrid: TEA.
7. Cohen, J. (1988). *Statistical power Analysis for the behavioral sciences* (2nd Ed.). Hillsdale, N.J., Erlbaum.
8. Coe, R. & Merino, C. (2003) Magnitud del efecto: Una guía para investigadores y usuarios. *Revista de Psicología - PUCP, 21,* 147-177.
9. Cubillo, J.C. & González Labra, M.J. (1998). El razonamiento analógico como solución de problemas. En M.J. González Labra (Ed.), *Introducción a la psicología del pensamiento* (pp. 409-451). Madrid: Trotta.
10. Elosua, P. (2003). Sobre la validez de los tests. *Psicothema, 15,* 315-321.
11. Embretson, S.E. (1983). Construct validity: Construct representation versus nomothetic span. *Psychological bulletin, 93,* 179-197. http://dx.doi.org/10.1037/0033-2909.93.1.179
12. Embretson, S.E. (1984). A general latent trait model for response processes. *Psychometrika, 49,* 175-186. http://dx.doi.org/10.1007/BF02294171
13. Embretson, S.E. (1991). A multidimensional item response model for learning processes. *Psychometrika, 56,* 495-515. http://dx.doi.org/10.1007/BF02294487
14. Embretson, S.E. & Reise, S.P. (2000). *Item response theory*

for psychologists. Mahwah: Lawrence Erlbaum Associates Inc.

15. Enzmann, D. (2005). *Dirk Enzmann – Statistical Software (Some Useful Things)*. Extraído el 1 de setiembre de 2010 de http://www2.jura.uni-hamburg.de/instkrim/kriminologie/Mitarbeiter/Enzmann/Software/Enzmann_Software.html.
16. Fischer, G. (1973). Linear logistic test model as an instrument in educational research. *Acta Psychologica, 37*, 359-374. http://dx.doi.org/10.1016/0001-6918(73)90003-6
17. Freund, P.A., Hofer, S. & Holling, H. (2008). Explaining and controlling for the psychometric properties of computer-generated figural matrix items. *Applied Psychological Measurement, 32,* 195-210. http://dx.doi.org/10.1177/0146621607306972
18. García-Cueto, E. & Fidalgo, A.M. (2005). Análisis de los ítems. En J. Muñiz, A.M. Fidalgo, E. García-Cueto, R. Martínez y R. Moreno (Eds.), *Análisis de los ítems* (pp. 53-130). Madrid: La Muralla.
19. Hanson, B.A. & Beguin, A.A. (2002). Obtaining a common scale for item response theory item parameters using separate versus concurrent estimation in the common-item equating design. *Applied Psychological Measurement, 26,* 3-24. http://dx.doi.org/10.1177/0146621602026001001
20. Lazarsfeld, P.F. (1950). The logical and mathematical foundation of latent structure analysis. En S. A. Stoufer et al. (Eds.). *Measurement and Prediction*. Princeton: Princeton University Press.
21. Lord, F.M. & Novick, M.R. (1968). *Statistical theories of mental tests scores*. Reading, MA: Addison-Wesley.
22. Martínez Arias, R. (1995). *Psicometría: Teoría de los tests psicológicos y educativos*. Madrid: Síntesis.
23. Martínez Arias, R., Hernández Lloreda, M.V. & Hernández Lloreda, M.J. (2006). *Psicometría*. Madrid: Alianza.
24. Mulholland, T.M., Pellegrino, J.W. & Glaser, G. (1980). Components of geometric analogy solution. *Cognitive Psychology, 12*, 252-84. http://dx.doi.org/10.1016/0010-0285(80)90011-0
25. Muñiz, J. (1994). *Teoría clásica de test*. Madrid: Pirámide.
26. Pereda Marín, S. (1987). *Psicología Experimental*. Madrid: Pirámide.
27. Rasch, G. (1960). *Probabilistic models for some intelligence and attainment tests*. Copenhague: The Danish Institute for Educational Research.
28. Raven, J.C., Court, J.H. & Raven, J. (1993). *Test de matrices progresivas. Escalas coloreada, general y avanzada. Manual*. Buenos Aires: Paidós.
29. Raven, J., Raven, J.C. & Court, J.H. (1991). *Manual for Raven's Progressive Matrices and Vocabulary Scales. Sections 1, 2, 3 and 4*. Oxford: Oxford Psychologists Press.
30. Rivera, S. (2000). Las ciencias formales en la era posmoderna. En E. Díaz (Ed.). *La posciencia: el conocimiento científico en las postrimetrías de la modernidad* (pp. 83-113). Buenos Aires: Biblos.
31. Spearman, C.E. (1904). General intelligence objectively determined and measured. *American Journal of Psychology, 15,* 201-293. http://dx.doi.org/10.2307/1412107
32. Spearman, C.E. (1907). Demonstration of formulae for true measurement of correlation. *American Journal of Psychology, 18*, 161-169. http://dx.doi.org/10.2307/1412408
33. Spearman, C.E. (1913). Correlations of sums and differences. *British Journal of Psychology, 5*, 417-126.
34. Spearman, C.E. (1923). *The Nature of intelligence and the principles of cognition*. Londres: MacMillan.
35. Sternberg, R.J. (1977). *Intelligence, information processing and analogical reasoning: the componential analysis of human abilities*. Hillsdale, NJ: Lawrence Erlbaum Associates.
36. Sternberg, R.J. (1987). *Inteligencia humana II: Cognición, personalidad e inteligencia*. Barcelona: Paidós.
37. Thurstone, L.L. (1928). Attitudes can be measured. *American Journal of Sociology, 33*, 529-554. http://dx.doi.org/10.1086/214483
38. Waller, N.G. (1995). *MicroFact 1.1. A microcomputer factor analysis program for ordered polytomous data and mainframe size problems*. St. Paul Minessota: Assessment System Corporation.
39. Whitely, S.E. & Schneider, L.M. (1981). Information structure for geometric analogies: A test theory approach. *Applied Psychological Measurement, 5,* 383-397. http://dx.doi.org/10.1177/014662168100500312
40. Wolf Nelson, N. & Gillespie, L.L. (1991). *Analogies for thinking and talking. Words, pictures and figures*. Tucson: Communication Skill Builders.
41. Yen, W.M. (1987). A comparison of the efficiency and accuracy of BILOG and LOGIST. *Psychometrika, 52,* 275-291. http://dx.doi.org/10.1007/BF02294241
42. Zimowski, M., Muraki, E., Mislevy, R. y Bock, R. (1996). *BILOG-MGTM: Multiple-group IRT analysis and test maintenance for binary items* [Computer program]. Chicago, IL: Scientific Software International.

Fecha de recepción: 13 de marzo de 2011
Fecha de recepción de la primera versión modificada: 29 de agosto de 2011
Fecha de aceptación: 12 de septiembre de 2011