Estudios Pedagógicos XLVIII, N° 3: 205-226, 2022
DOI: 10.4067/S0718-07052022000300205

INVESTIGACIONES

Versión abreviada de la escala de matrices progresivas de Raven para población con talento académico: Una aproximación desde la Teoría de Respuesta al Ítem[1]

Short version of Raven progressive matrices scale for gifted population: An Item Response Theory approach

Carlos Calderón Carvajal[a], Diego Palominos-Urquieta[a], Mauricio Briceño[a], Jorge Rojas[a], Kevin Peña[a], Diego Henríquez[a]

[a] Universidad Católica del Norte, Escuela de Psicología, Chile.
ccalderon@ucn.cl, diego.palominos@alumnos.ucn.cl, mbricenosa@gmail.com, Jorge.rojas.chaile@gmail.com, kevindaniel.pr@gmail.com, xdiegohenriquez@gmail.com

RESUMEN

La escala Matrices Progresivas General de Raven es un test utilizado en los procesos de admisión de los programas de talento académico en Chile. El presente estudio tiene como objetivo evaluar las propiedades psicométricas de la escala Raven en una amplia muestra de estudiantes utilizando la Teoría de Respuesta al Ítem (TRI). Se administró la prueba a 935 estudiantes participantes del proceso de admisión de un programa de talentos académicos. Los resultados muestran que la escala total no se ajusta a un modelo unidimensional. Los análisis por separado muestran que sólo las subescalas D y E presentan un ajuste adecuado. Los análisis de ambas subescalas muestran un adecuado ajuste al modelo TRI y no presentan funcionamiento diferencial por sexo. Se discuten las implicancias de la presencia de dos factores y se proponen como una versión reducida para procesos de admisión de programas de talento académico.

Palabras clave: Análisis factorial, Teoría de respuesta al ítem, Matrices progresivas de Raven, Inteligencia general, Funcionamiento diferencial.

ABSTRACT

Raven's General Progressive Matrices scale is a test used in the admission processes of academic talent programs in Chile. The present study aims to evaluate the psychometric properties of the Raven scale in a large sample of students using the Item Response Theory (IRT). The test was administered to 935 students participating in the admission process to an academic talent program. The results show that the full scale does not fit a one-dimensional model. Separate analyzes show that only the D and E scales show adequate fit. The analysis of both subscales shows a proper fit to the two-parameter IRT model and it doesn´t show differential functioning by gender. The implications of the presence of two factors are discussed and its use is proposed as a reduced version to the admission process of academic talent programs.

Key words: Differential function, Item Response Theory, Factor analysis, General intelligence, Raven's progressive matrices scale.

[1] Este estudio ha sido financiado por el proyecto Fondecyt 11150182 y auspiciado por el programa Desarrollando y Liderando Talentos Académicos (Delta) de la Universidad Católica del Norte.

Recibido: 23/05/2021
Aceptado: 14/03/2022

1. INTRODUCCIÓN

La escala de Matrices Progresivas de Raven (Raven, 1936; 1991; Raven, Raven & Court, 1998), es un instrumento ampliamente utilizado en gran variedad de contextos, tanto en investigación básica, como en ámbitos aplicados. Es considerada una de las mejores estimaciones del factor general de inteligencia (Abad, Colom, Rebollo & Escorial, 2004; Paul, 1986; Jensen, 1998), ya que es una medida no verbal de capacidad intelectual y posee un mínimo componente cultural (Flynn, 1984, 1987, 1998). Estas características lo han hecho un instrumento de amplio uso, útil en variados contextos (Van der ven & Ellis, 2000).

El factor general de inteligencia, es un constructo enmarcado en la Teoría Bifactorial de inteligencia (Spearman, 1927). Según esta aproximación, el concepto de inteligencia puede ser explicado a través de dos tipos de factores, un factor general, denominado factor G, y diversos factores específicos, denominados S. El factor G es definido como el fundamento esencial de las operaciones mentales. Corresponde a la capacidad de reflexionar y observar el propio trabajo mental interno, captar las relaciones sustanciales entre dos o más conceptos y percibir las ideas implícitas en una relación de elementos. Por otro lado, los factores S se relacionan con habilidades características de cada tarea y no adoptan el carácter transversal que tiene el factor G.

La evidencia previa señala que las puntuaciones de la escala de Matrices Progresivas de Raven no solo estarían vinculadas con el factor G, sino que también serían indicador de otros tipos de habilidades, como lo son la resolución de problemas, la detección de patrones, la capacidad analítica-verbal y habilidades viso-espaciales (Abad *et al*., 2004; Carpenter, Just & Shell, 1990; Gignac, 2015; Hayes, Petrov & Sederber, 2015; DeShon, Chan & Weissbein, 1995; Dillon, Pohlmann & Lohman, 1981). A pesar de ello, sigue planteándose como una prueba que evalúa una habilidad general (Gignac, 2015).

Aun cuando el estudio original (Raven, 1936) y estudios posteriores (Alderton & Larson, 1990; Arthur & Woehr, 1993) propone un modelo unidimensional, otros trabajos han mostrado evidencia que apoya la presencia de más factores (Dillon *et al*., 1981; DeShon *et al*., 1995, Fernández-Liporace, Ongarato, Saavedra, & Casullo, 2004). Por ejemplo, DeShon, *et al*. (1995) proponen dos factores, los cuales estarían relacionados con una dimensión viso-espacial y otra analítico-verbal. Estas dimensiones no serían independientes ni excluyentes, sino más bien constituirían estrategias de respuesta que pueden complementarse o bien ser utilizadas simultáneamente para resolver una misma tarea. Dillon *et al*. (1981), por su parte, también plantean un modelo compuesto por dos factores. Una dimensión de detección progresiva de patrones y una segunda dimensión de suma y resta de patrones. Waschl, Nettelbeck, Jackson & Burns (2016) proponen que los factores expuestos anteriormente tanto por Dillon *et al*. (1981) como por de DeShon *et al*. (1995) pueden ser útiles para comprender cómo un individuo aborda los ítems. Sin embargo, no lograrían distinguir entre ítems que involucran o no, una u otra habilidad.

Las Matrices Progresivas se presentan en tres formas distintas. Una de ellas es la Matriz Progresiva Coloreada (MPC) que está destinada para emplearse con niños/as entre 5 y 10 años aproximadamente, y adultos mayores (Raven *et al*., 1993). La versión avanzada (MPA) que está orientada a discriminar principalmente a los individuos con un alto nivel de rasgo (Raven *et al*., 1993). Esta escala se divide en dos series, la primera consta de 12 ítems con una dificultad acorde a la escala general y que funciona solo a modo de entrenamiento, ya que no es considerada en la puntuación total. La serie II consta de 36 ítems, que tendrían

una mayor dificultad en relación con la escala general y son los que se consideran en el puntaje de la prueba.

Por último, la escala general (MPG) está destinada a abarcar todo el rango intelectual, y es utilizada para todas las edades (Raven *et al*., 1993). Está constituida por 60 ítems, distribuidos en cinco series (A, B, C, D y E) de 12 ítems cada una. Estas subescalas están organizadas en nivel de dificultad creciente.

En términos generales los ítems presentan la misma estructura. A la persona se le presenta una matriz con distintos elementos, donde hay un segmento faltante. Se le pide elegir entre una serie de alternativas (ver figura 1). Se asume que, para identificar la alternativa correcta, las personas deben desplegar habilidades perceptuales de observación y razonamiento lógico. Cada una de las matrices debe seguir una secuencia tanto en horizontal como en vertical, buscando el sentido a los patrones y sus características. Su administración puede ser individual y/o colectiva, y el tiempo de aplicación es variable (entre 30 a 60 min.). Se puede aplicar en niños, adolescentes y adultos.

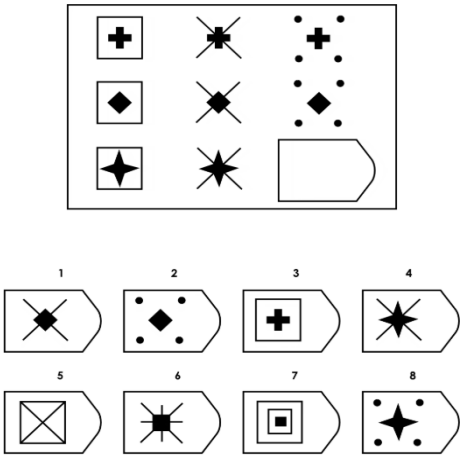

Figura 1. Ejemplo de uno de los ítems de las matrices progresivas escala general.

Nota: se aprecia que la matriz sigue horizontalmente un patrón de cuadrados, una equis y cuatro puntos, manteniendo la misma figura al centro. Verticalmente la figura del centro cambia de una cruz a un rombo y posterior mente a una estrella de cuatro puntas, sin embargo, se mantiene la parte exterior de la figura. Respetando esos patrones, la respuesta correcta es el objeto número 8 (extraído de Raven *et al*., 1993).

En Chile, la escala de inteligencia de mayor predominancia en los establecimientos educacionales es la escala de Wechsler (Mansilla, Vásquez & Estrada, 2012; Rosas, Tenorio, Pizarro, Cumsille, Bosch, Arancibia, Carmona-Halty, Pérez-Salas, Pino, Vizcarra & Zapata-Sepúlveda, 2014). Sin embargo, la escala de MPG continúa siendo ampliamente utilizada y es solicitada con frecuencia por distintas organizaciones que financian programas de talento académico (García Cepero, Proestakis, Lillo, Muñoz, López y Guzmán, 2012). En comparación a la escala de Wechsler, es un instrumento de aplicación mucho más

económico en cuanto a acceso, aplicación y evaluación, convirtiéndolo en uno de los test más utilizados en Chile (Mansilla *et al*., 2012).

La aplicación de la escala de Matrices Progresivas (MP) ha sido ampliamente utilizada en programas de Talento. Tanto es así, que abundante literatura aborda la utilidad y pertinencia del test de MP para los procesos de admisión (Lohman, Korb & Lakin, 2008; Matthews, 1988; Alvino, McDonnel & Richert, 1981). Sin perjuicio de ello, los resultados de estas investigaciones proporcionan evidencia que la escala de Matrices Progresivas en la identificación de talento no debe utilizarse como único criterio. Esto es debido a que el talento académico considera un conjunto de características más amplias y no sólo capacidades intelectuales (Matthews, 1988; Alvino, McDonnel & Richert, 1981). Estos trabajos aconsejan la utilización de otras escalas complementarias para la identificación de talento académico. Desde esta perspectiva integral, su uso se limita a evaluar una de las tres dimensiones del talento planteado por Renzulli (2016) como lo es una capacidad intelectual por encima de la media.

1.1. ESCALA RAVEN: EVIDENCIA PREVIA

La escala de Matrices progresivas (Raven, 1936; Raven *et al*., 1993), tiene algunas versiones reducidas, tanto de la versión avanzada (MPA) como de la escala general (MPG).

Arthur & Day (1994), desarrollaron una versión abreviada con una muestra de estudiantes universitarios. El propósito principal de esta versión fue reducir el tiempo de aplicación. La versión definitiva quedo conformada por 12 de los 36 ítems de la escala original. Los criterios de eliminación se basaron en la correlación ítem-test, los índices de dificultad y el aumento en el coeficiente α de Cronbach si el ítem es eliminado. Los resultados de esta versión muestran una consistencia interna (α de Cronbach) de .65, en comparación con el AMP original de .86. La correlación entre ambas escalas fue de .66. Los resultados del test-retest fueron de .75. El análisis factorial confirmatorio (AFC) mostró la existencia de un modelo de un factor (χ^2_{54}= 54.78; p>.05; GFI= .985, AGFI= .978). Estos resultados han sido replicados en estudios posteriores (Chiesi, Ciancaleoni, Galli, Morsanyi & Primi, 2012).

Por su parte, Bors & Stokes (1998), llevaron a cabo la versión reducida de la escala MPA a través de tres estudios. El propósito era generar una norma para estudiantes de primer año de universidad a partir de la cual se obtuvo una versión reducida. El primer estudio se llevó a cabo con una muestra de 506 estudiantes, a los que se les aplicó la versión de la MPA con los dos sets ítems (serie I de prueba de 12 ítems, y serie II de 36). A partir de las correlaciones ítem-test, la versión reducida quedó conformada por 12 de los 36 ítems originales. Los resultados del AFC apoyaron una estructura interna de dos factores correlacionados (χ^2_{87}= 129.72; p>.05; CFI= .906, NFI= .887). El coeficiente de α de Cronbach fue de .73 y los resultados del análisis test-retest mostraron una estabilidad aceptable (r= .82).

En un tercer estudio analizaron las correlaciones de los puntajes de la versión abreviada propuesta, la versión completa y la versión abreviada propuesta por Arthur & Day (1994). Los resultados mostraron una correlación de .88 entre la versión corta de Bors & Stokes (1998) y la escala completa, mientras que la correlación entre ambas versiones abreviadas fue de .87.

Respecto a la escala MPG y sus versiones reducidas, encontramos el trabajo de Wytek, Opgenoorth & Presslich (1984). El propósito fue reducir el tiempo de aplicación para su uso en contextos psiquiátricos. La prueba completa se aplicó a 300 pacientes. Mediante el

modelo de Rasch y análisis de la Teoría Clásica de los Test, se redujo la escala de 60 a 30 ítems. Pese a la reducción del tamaño de la escala, no se vio afectada significativamente su fiabilidad. Se estudió la relación entre la versión reducida desarrollada y una versión reducida de la escala Wechsler (WIP-IQ; Dahl, 1972) en una muestra de 390 pacientes. Se obtuvieron correlaciones de .61 entre la escala total del Raven y el puntaje de la escala WIP-IQ, y de .59 entre la nueva escala reducida y el puntaje del WIP-IQ.

Una segunda versión abreviada de la MPG es la de Bilker, Hansen, Brensinger, Richard, Gur, & Gur (2012). Estos autores desarrollaron una versión abreviada con el propósito de evaluar de manera breve la capacidad intelectual en investigaciones con pacientes psiquiátricos y con déficits cognitivos. La primera etapa consideró la selección de un grupo de ítems (Escala A) que fuera altamente predictivo de la escala original de 60 ítems. Adicionalmente se identificó un segundo subconjunto de ítems (Escala B) que no estaban contenidos en el primero (Escala A), y que fueran también altamente predictivos.

El estudio incluyó a 180 personas con edades entre 16 y 77 años. Esta muestra consideró personas sin diagnóstico, con diagnósticos de esquizofrenia y pacientes no psicóticos con trastornos del eje I del DSM-IV. Se elaboraron dos escalas, A y B, con nueve ítems cada una. Los índices de correlación con la versión original de 60 ítems fueron de .98 y .97 respectivamente.

Debido a que la estrategia analítica para la propuesta de la versión reducida de la escala Raven considera la aplicación de técnicas en el marco de la Teoría de Respuesta al Ítem (TRI), el siguiente apartado intenta presentar las principales características de esta aproximación, así como los alcances y ventajas en comparación a la Teoría Clásica de los Test (TCT).

1.2. TEORÍA DE RESPUESTA AL ÍTEM (TRI)

La Teoría Respuesta al ítem (TRI) ha sido el modelo psicométrico con mayor desarrollo en los últimos años (Hambleton & Swaminathan, 1985; Lord, 1980; Martínez-Arias, 1995; Muñiz, 1997; 2010).

Ello debido a que ha logrado resolver importantes limitaciones de la Teoría Clásica de los test (Muñiz, 2010). Uno de los problemas de la TCT es que sus mediciones no resultan invariantes respecto del instrumento utilizado. Dicho de otro modo, no es posible comparar las puntuaciones de dos personas, las cuales han sido evaluadas por test diferentes que miden el mismo constructo. Otras de las limitaciones de la TCT es que las propiedades psicométricas de los ítems (e.g. el índice de dificultad), dependen de las características de la muestra a partir de la cual estos parámetros han sido estimados. La TRI intenta resolver algunos de estos problemas.

El modelo TRI se basa en establecer una relación entre la puntuación verdadera de los sujetos o también denominados rasgos latentes (θ), y la probabilidad de responder un determinado ítem P(θ). Esta relación es representada a través de una función matemática denominada Curva Característica del Ítem (CCI), la cual se basa en una función de regresión logística.

El modelo requiere de dos supuestos fundamentales. El primero de ellos corresponde a la independencia local. Este supuesto asume que las respuestas de las personas a los ítems sólo sean función del nivel de rasgo y no estén condicionadas al resto de los ítems del instrumento. El segundo corresponde al supuesto de unidimensionalidad. Este se refiere a que las respuestas

de los sujetos dependan sólo de un factor o rasgo latente. Aunque ambos supuestos se refieren a propiedades distintas, poseen una estrecha relación, con lo cual, el cumplimiento de la unidimensionalidad trae consigo el cumplimiento de independencia local.

En términos generales, el modelo plantea la existencia de una función matemática que vincula el nivel de habilidad de las personas y la probabilidad de acierto al ítem. Esta corresponde a una función logística definida por la CCI. Aun cuando existen dos tipos de modelos para definir CCI, el modelo de ojiva normal y el modelo logístico, es este último el que ha sido mayormente utilizado debido a su simplicidad matemática.

Dentro de los modelos logísticos, existen distintos tipos dependiendo del número de parámetros con los cuales es definida la CCI; el Modelo de un parámetro o Modelo de Rasch (1PL), modelo de dos parámetros (2PL) y modelo de tres parámetros (3PL).

En el Modelo de Rasch o Modelo de un parámetro (1PL), la CCI está definida por el parámetro *b*, que expresa el nivel de rasgo donde la probabilidad de responder correctamente al ítem es igual a .5. Cómo es posible advertir, aunque este parámetro se denomina índice de dificultad, no tiene la misma interpretación que el índice de dificultad en la Teoría Clásica. Este parámetro indica la posición en la CCI en relación al continuo de aptitud θ de los sujetos. Mientras más alto es el valor del parámetro *b*, se requerirá de un mayor nivel de aptitud para tener una probabilidad de acierto superior a .5.

En el Modelo de dos parámetros (2PL), además del parámetro *b* propuesto por el modelo de Rasch, se añade a la función un segundo parámetro denominado *a*. Este corresponde al valor de máxima pendiente de la curva en el punto del rasgo definido por el parámetro *b*. Este parámetro también es denominado parámetro de discriminación, ya que entrega información de la máxima capacidad de discriminación de los sujetos en el nivel de rasgo definido por el parámetro *b*.

Finalmente, el Modelo de tres parámetros (3PL), considera un parámetro denominado *c*. Este es conocido como parámetro de pseudo-azar, ya que indica la probabilidad de acierto de una persona con el menor nivel de rasgo. Debido a que los métodos de estimación tienen dificultades para la estimación de este parámetro, es que los modelos más utilizados corresponden al de 1PL y 2PL.

Uno de los aportes más relevantes de la TRI corresponde al Análisis del funcionamiento diferencial (DIF). Este consiste en estudiar las diferencias entre grupos de sujetos que, teniendo el mismo nivel de rasgo, presentan diferentes probabilidades de responder correctamente a un ítem debido a su pertenencia grupal. Este aspecto resulta relevante en cuanto promueve el uso de este tipo de análisis, no solo en la construcción de test invariantes entre distintas poblaciones, sino que abre la posibilidad de estudiar el funcionamiento diferencial entre grupos de interés psicológico. Gran atención ha tenido la evaluación del efecto de género, etnia o raza, en el funcionamiento de los test. Ello, para evitar conflictos éticos o sociales, o bien evaluar el origen de estas diferencias (Attorresi, Lozzia, Abal, Galibert & Aguerri, 2009).

1.3. SESGO DE GÉNERO EN LA EVALUACIÓN DE LA INTELIGENCIA

La diferencia de género en la inteligencia ha sido ampliamente discutida en la literatura. Desde comienzos del siglo XX y hasta comienzos del siglo XXI, algunos autores defendieron ampliamente la inexistencia de diferencia de género en habilidades intelectuales (Burgaleta, Head, Álvarez-Linera, Martínez, Escorial, Haier & Colom,

2012; Burt & Moore, 1912; Cattell, 1971; Eysenck & Kamin; 1981; Haier, 2007; Halpern, 2012; Herrnstein & Murray, 1994; Lubinski, 2000; Mackintosh, 1996, 2011; Savage-McGlynn, 2012).

Por contraparte, un grupo no menor de investigadores, han defendido la existencia de diferencias, planteando la superioridad de los hombres sobre las mujeres en la escala de rendimiento intelectual (Lynn, 1994; Colom & Lynn, 2004). Lynn & Irwing (2004) presentaron un meta-análisis con los resultados de las diferencias de género en la evaluación de inteligencia general en población adulta usando las MP de Raven. En este estudio plantean una ventaja de los hombres de CI de hasta 5 puntos por sobre las mujeres a partir de los 16 años. En otro meta-análisis, Irwing & Lynn (2005) indican que la diferencia de CI a favor de los hombres en las MP alcanzaba 4.6 puntos. Estos resultados fueron respaldados por una muestra escocesa que presenta una diferencia a favor de los hombres de 4.35 puntos de CI (Deary, Whiteman, Starr, Whalley & Cox, 2004) y de 4.05 puntos de CI en una muestra serbia (Cvorovic & Lynn, 2014).

Respecto a otras pruebas, la evidencia indica que no existe diferencias significativas de género en las puntuaciones del WPPSI (Lynn, 2017). Para el caso del WISC y sus versiones estandarizadas WISC-III Y WISC-IV la revisión de 31 muestras indica que los hombres obtuvieron una ventaja promedio de 2.85 puntos de CI sobre las mujeres. En tanto para el WAIS, una revisión de 33 estudios sugiere que existe una diferencia de 3.6 puntos en promedio que favorece a los hombres (Lynn, 2017).

Por otra parte, Colom, Juan-Espinoza, Abad y García (2000) llevaron a cabo un estudio para evaluar las diferencias de género en una amplia muestra de participantes (N = 10,475). Este estudio mostró que la diferencia del factor G entre hombres y mujeres es despreciable. Posteriormente, Colom, García, Abad y Juan-Espinosa (2002) encontraron una relación nula entre sexo y factor G en una adaptación española de la escala de WAIS-III.

Abad *et al*. (2004) llevaron a cabo una investigación con la MPA. Estos autores encontraron la presencia de DIF en un número importante de ítems que favorecía a los hombres. Adicionalmente, encontraron que, eliminando los ítems con DIF de la prueba, las diferencias entre hombres y mujeres desaparecieron.

2. EL PRESENTE ESTUDIO

Como hemos comentado anteriormente, la escala de matrices progresivas es una de las medidas más ampliamente utilizadas en la evaluación del factor G de inteligencia en el contexto chileno, tanto en la investigación básica como en el ámbito aplicado. Ello es debido a que es más económica y corresponde a una prueba no verbal, lo cual evita el efecto de terceras variables. No existe evidencia previa sobre sus propiedades psicométricas en Chile desde la TRI. Las investigaciones existentes, solo hacen referencias a estudios normativos o de estandarización (Mansilla *et al*., 2012). Adicionalmente, considerando que es el test más utilizado en Chile en los procesos de admisión de programas de talento académico, no existen estudios previos en este tipo de poblaciones. Por tanto, el propósito de este estudio es evaluar las propiedades psicométricas de la MPG desde un enfoque TRI y la elaboración de una versión reducida para su uso en población con talento académico.

Los análisis estarán centrados en tres tipos de evidencias: Análisis de Estructura interna, Análisis de fiabilidad y de ítems, y análisis de DIF entre hombres y mujeres.

Para el estudio de la estructura interna, se llevó a cabo un análisis AFC para evaluar el ajuste de los datos a la unidimensionalidad. Además, se utilizará el modelo de TRI para la evaluación del funcionamiento la prueba. Ello debido a que en Chile no existe estudios previos que utilicen esta clase de análisis, y los estudios en el contexto sudamericano son escasos (Escurra-Mayaute & Delgado-Vásquez, 2010). Dado que en el país esta escala es utilizada para procesos de admisión de programas de talento académico, es que este estudio se ha llevado a cabo con estudiantes que participan en el proceso de admisión de uno de estos programas.

Adicionalmente, el propósito de los análisis psicométricos es proponer una versión reducida seleccionando los ítems que tienen mejores propiedades psicométricas y mejor ajuste, utilizando un AFC y el modelo TRI de 2PL.

Finalmente, dado que las posibles diferencias de género en habilidades intelectuales han generado un amplio debate, y frente a la posibilidad de la existencia de DIF en el funcionamiento de la prueba, nos hemos propuesto analizar la existencia de funcionamiento diferencial desde el modelo TRI. A partir de la obtención de la versión abreviada y la selección de los ítems que tienen mejores propiedades psicométricas, compararemos el rendimiento de hombres y mujeres.

Ya que previamente no existe una evidencia sólida para poder plantear hipótesis sobre la estructura factorial de la escala, las hipótesis serán planteadas de manera exploratoria. Una primera hipótesis corresponde a la presencia de una estructura unidimensional de la prueba total, tal como lo propone el estudio original (Raven, 1936). Una segunda hipótesis corresponde a la presencia de un modelo de dos factores, vinculados a las dimensiones propuestas en estudios posteriores, como por ejemplo la visoespacial y analítico-verbal. En base a la literatura existente (Abad *et al*., 2004; Colom & Abad, 2007), creemos que los análisis de funcionamiento diferencial mostraran la presencia de DIF en algunos de los ítems de la escala.

3. MÉTODO

La propuesta corresponde a un estudio psicométrico de corte transversal y de tipo no experimental.

3.1. PARTICIPANTES

La muestra fue no probabilística e intencionada. Los participantes fueron 935 estudiantes de entre 9 y 18 años de edad (M=11.62, SD= 1.63) que cursan entre 6° de educación primaria y 1° año de educación secundaria, todos ellos participantes del proceso de admisión a un programa de talento académico de una universidad chilena. En cuanto al género, 414 eran hombres (44.3%) y 521 mujeres (55.7%).

3.2. PROCEDIMIENTO

Este estudio fue aprobado por el Comité de Ética Científica correspondiente. El procedimiento de obtención de datos se llevó a cabo en dos fines de semana durante el proceso de admisión de los estudiantes al programa de talentos académicos. La convocatoria a este proceso se realiza a través de los colegios, y la participación es voluntaria. Los

estudiantes asisten acompañados de sus padres/madres o tutores durante una mañana. En cada sesión, a los estudiantes se les describió las características del programa, se les dio entrega del asentimiento informado y se les aplicó las pruebas de admisión entre las cuales se encuentra la escala de Matrices Progresivas de Raven en su formato general (MPG). Paralelamente, en otro salón, a los padres/madres se les informó las características de programa y del estudio. Posteriormente se les entregó el consentimiento informado para su firma.

3.3. INSTRUMENTO

Como hemos mencionado anteriormente el instrumento corresponde al test de Matrices Progresivas Raven, Escala General (1991). Corresponden a un instrumento que evalúa inteligencia general de modo no verbal. Está compuesto por 60 problemas presentados en cinco subescalas de 12 ítems cada una, organizados en dificultad creciente.

3.4. PROPUESTA DE ANÁLISIS

Los análisis los hemos llevado a cabo en cuatro etapas. La primera de ellas tiene el propósito de evaluar la dimensionalidad de los datos. Como hemos comentado anteriormente, el contraste de la dimensionalidad es un requisito para el adecuado cumplimiento de los supuestos de unidimensionalidad e independencia local, condiciones necesarias para el ajuste de un modelo TRI. Para el estudio de estos supuestos hemos sometido los datos a un Análisis Factorial Confirmatorio (AFC). El ajuste de los datos a un modelo de un factor, apoyaría la existencia de un modelo unidimensional subyacente a los datos.

En segundo lugar, y una vez contrastados los supuestos de dimensionalidad e independencia local, hemos ajustado un modelo TRI para la obtención de los parámetros de los ítems. Para la evaluación del ajuste del modelo TRI, hemos obtenido tanto los índices de bondad de ajuste global, como los índices de ajuste local para los ítems individuales. Estos índices evalúan la existencia de diferencias estadísticas significativas entre las respuestas observadas y las respuestas pronosticadas a partir de los parámetros de los ítems. La ausencia de diferencias estadísticas significativa, apoyarían el ajuste de los datos al modelo TRI. Adicionalmente, para evaluar la fiabilidad, hemos obtenido los índices de fiabilidad alfa de Cronbach y las correlaciones entre las subescalas de la prueba para la evaluación de la fiabilidad global. También hemos obtenido las funciones de información del test a partir del modelo TRI, la cual nos entrega información acerca de la fiabilidad en cada uno de los sectores del continuo del nivel de rasgo.

En tercer lugar, para evaluar la posible presencia de un sesgo de género, hemos sometido los datos a un análisis DIF. Para ello hemos contrastado la existencia de diferencias estadísticas significativas en los parámetros de los ítems entre hombre y mujeres, a través del estadístico de Wald. La existencia de diferencias significativas en los parámetros de los ítems, serían un indicador de la presencia de funcionamiento diferencial, y junto con ello, la presencia de sesgo producto del sexo de los participantes.

Por último, una vez descartada la presencia de DIF por sexo, hemos comparado las puntuaciones globales entre hombres y mujeres para cada una de las escalas de la prueba, con el fin de contrastar la existencia de diferencias en el rendimiento intelectual entre ambos sexos.

4. RESULTADOS

Los resultados serán presentados en cuatro secciones. En la primera de ellas se presentarán los resultados del ajuste de los modelos de medida a través del Análisis Factorial Confirmatorio (AFC). Se compara el ajuste del modelo de la escala completa y la de las subescalas por separado. Estos análisis se han llevado a cabo con el programa Mplus v7.

En segundo lugar, se presentarán los resultados de la estimación de los parámetros de los modelos TRI utilizando el paquete Mirt disponible para Rstudio. En todos los casos hemos utilizado el modelo de dos parámetros. Los resultados del AFC y el modelo TRI han sido utilizados en la elección de los ítems para la propuesta de la versión reducida de la escala.

Una vez obtenida esta nueva versión, hemos realizado un Análisis de Funcionamiento Diferencial de los Ítems (DIF) según sexo. Ello, con el propósito de determinar si las propiedades psicométricas de los ítems pueden favorecer o perjudicar a uno de los dos grupos y evaluar de efecto de DIF en las posibles diferencias que pudiesen mostrar las puntuaciones globales. Finalmente presentaremos una comparación en las puntuaciones de la versión abreviada entre hombres y mujeres, a través de un ANOVA factorial utilizando de programa SPSS v25.

4.1. ANÁLISIS FACTORIAL CONFIRMATORIO

Se realizó un análisis AFC en la escala de MPG para comprobar el supuesto de unidimensionalidad del modelo. Para su estimación hemos utilizado el método Weighted Least Square with Mean and Variance Adjusted (WLSMV), adecuado para variables categóricas, el cual está disponible en Mplus.

Para evaluar el ajuste global de los modelos se emplearon los índices de ajuste comúnmente utilizados en la práctica investigadora, Chi-cuadrado, el CFI de Bentler (1992) y el RMSEA de Steiger & Lind (1980). La tabla 1 muestra los resultados del ajuste global de la escala completa y de las subescalas por separado.

Los resultados muestran que la escala global no ajusta al modelo unidimensional [χ^2_{1710}=2586.253, CFI= .808, RMSEA=.023]. Debido a ello, hemos decidido realizar un análisis de cada una de las escalas por separado. Los resultados indican que las subescalas A y B no ajustan al modelo unidimensional (CFI< .90; RMSEA<.06). Por otro lado, la subescala C obtiene un ajuste moderado (CFI= .90-.95; RMSEA<.06). Finalmente, las subescalas D y E obtienen un ajuste excelente (CFI>.95; RMSEA<.06). La correlación entre ambas subescalas fue de .38.

Tabla 1. Índices de bondad de ajuste del AFC para cada uno de los modelos de medida

	χ^2	d.f	valor-p	CFI	RMASEA (C190)
Escala global	2586.253	1710	.000	.808	.023 (.022-.025)
Escala A	91.412	54	.001	.878	.027 (.017-.037)
Escala B	191.364	54	.000	.879	.052 (.044-.060)
Escala C	90.359	54	.001	.946	.027 (.017-.036)
Escala D	121.975	54	.000	.966	.037 (.028-.045)
Escala E	108.255	54	.000	.974	.033 (.024-.042)

Debido a que las subescalas A, B y C no se ajustan al modelo unidimensional, el resto de los análisis sólo los hemos realizados sobre las subescalas D y E, ya que sólo estas cumplen con los supuestos unidimensionalidad necesario para el análisis TRI.

4.2. TEORÍA DE RESPUESTA AL ÍTEM

Una vez ajustado la unidimensionalidad de las escalas D y E, se realizó el análisis TRI de ambas escalas. Hemos especificado el modelo de dos parámetros, debido a que es el más utilizado en la práctica investigadora, ya que estudios previos muestran la presencia de sesgo en las estimaciones del parámetro c (Muñiz, 1997). Para la estimación de los parámetros hemos utilizado un método de estimación de Máxima Verosimilitud Conjunta. Los parámetros estimados, sus errores de estimación, y el índice de ajuste local $S\text{-}\chi^2$ (Orlando & Thissen, 2000, 2003), son presentados en las tablas 2 y 3.

Tabla 2. Parámetros estimados, errores de estimación e índice de ajuste de los ítems de la escala D

	a	s.e.	b	s.e.	$S\text{-}X^2$	p	RMSEA
Ítem 37	1.413	.248	-3.174	.381	14.994	.059	.031
Ítem 38	2.060	.265	-1.848	.137	2.235	.946	.000
Ítem 39	1.959	.199	-1.869	.142	2.669	.914	.000
Ítem 40	1.646	.251	-1.711	.140	9.744	.204	.020
Ítem 41	2.238	.199	-1.837	.130	5.417	.609	.000
Ítem 42	2.166	.290	-1.518	.109	4.259	.750	.000
Ítem 43	1.459	.170	-1.149	.107	14.940	.060	.030
Ítem 44	1.386	.162	-1.328	.122	8.747	.364	.010
Ítem 45	1.245	.147	-1.106	.114	11.520	.174	.022
Ítem 46	2.051	.243	-1.207	.091	11.551	.073	.031
Ítem 47	.581	.108	1.282	.243	10.695	.098	.029
Ítem 48	.477	.141	4.090	1.132	6.549	.256	.018

Los resultados muestran que los ítems de la subescala D (tabla 2) presentan ítems con niveles de dificultad que van desde un nivel muy bajo (*b*=-3.174) en el ítem 37, a muy difícil (*b*=4.090) en el ítem 48. Sin embargo, la mayoría de ellos se ubican entre el -2 y el -1 del nivel de rasgo. Con respecto al parámetro *a*, los ítems presentan altísimos niveles de discriminación, en su mayoría por sobre 1, incluso algunos alcanzando valores sobre 2. Considerando los valores estimados de ambos parámetros es posible determinar que la subescala D posee una gran capacidad de discriminación, la cual se concentra entre 1 y 2 desviaciones típicas por debajo de la media del nivel de rasgo.

La tabla también muestra el estadístico S-χ^2, el cual contrasta la hipótesis de ajuste de las CCI de los ítems a los datos. Este estadístico compara las proporciones de respuesta correctas para cada nivel de rasgo y las proporciones pronosticadas por la CCI de cada ítem. Adicionalmente se presenta el estadístico RMSEA, el cual se calcula a partir de S-χ^2, y entrega información acerca de la magnitud de desajuste de los ítems. Valores por debajo de .05 son considerados valores aceptables.

Respecto al ajuste de los ítems al modelo 2PL, el estadístico de ajuste S-χ^2 muestra que para todos los ítems se mantiene la hipótesis de ajustes (p >.05). Adicionalmente, todos los valores de RMSEA se mantienen por debajo de .05. Finalmente, para evaluar el ajuste global del modelo, hemos obtenido el índice M_2 (Maydeu-Olivares & Joe, 2005, 2006) el cual tiene una distribución χ^2, y los índices de ajuste tradicionales RMSEA, TLI y CFI, obtenidos a partir de este estadístico. Los índices de ajuste global del modelo muestran un ajuste adecuado de la subescala D (M_2=109.391; p<.05; RMSEA=.033; TLI=.979; CFI=.983).

Con respecto a la escala E (tabla 3), las estimaciones muestran problemas en los ítems 55, el cual presenta discriminación negativa (*a*=-.153), y los ítems 56, 58 y 59, los que presentan parámetros de dificultad excesivamente altos (34.870; 120.866). Adicionalmente, estos ítems presentan errores de estimación altísimos asociados a las estimaciones de *b*. Frente a estos resultados, hemos decidido su eliminación.

Tabla 3. Parámetros estimados, errores de estimación e índice de ajuste de los ítems de la escala E

	a	s.e.	b	s.e.	S- χ^2	p	RMSEA
Ítem 49	1.131	.144	-1.172	.109	11.079	.026	.044
Ítem 50	1.130	.117	-.424	.080	4.926	.425	.000
Ítem 51	1.397	.138	-.579	.074	1.855	.869	.000
Ítem 52	2.234	.218	.274	.053	5.663	.340	.012
Ítem 53	4.855	.950	-.002	.059	2.610	.456	.000
Ítem 54	1.767	.166	.286	.059	4.989	.417	.000
Ítem 55	-.153	.223	-23.104	33.633	2.890	.822	.000
Ítem 56	.096	.211	34.870	76.428	5.167	.523	.000
Ítem 57	.532	.111	3.223	.632	4.872	.560	.000
Ítem 58	.374	.154	7.134	2.823	5.750	.452	.000
Ítem 59	.019	.132	120.866	825.266	12.982	.043	.035
Ítem 60	.575	.167	4.935	1.317	5.047	.410	.003

La tabla 4 presenta los parámetros estimados de la subescala E reducida, en la cual han sido eliminados los ítems 55, 56, 58 y 59.

Tabla 4. Parámetros estimados, errores de estimación e índice de ajuste de los ítems de la escala E reducida

	a	s.e.	b	s.e.	S-X^2	p	RMSEA
Ítem 49	1.322	.144	-1.178	.108	8.622	.071	.035
Ítem 50	1.130	.117	-.424	.080	7.958	.159	.025
Ítem 51	1.396	.138	-.579	.074	3.910	.418	.000
Ítem 52	2.238	.218	.274	.053	13.930	.016	.044
Ítem 53	4.832	.933	-.001	.043	2.643	.450	.000
Ítem 54	1.773	.167	.285	.059	11.117	.050	.036
Ítem 57	.531	.111	3.227	.632	7.114	.310	.014
Ítem 58	.372	.154	7.168	2.850	3.101	.796	.000
Ítem 60	.575	.167	4.934	1.316	7.433	.283	.016

Para la subescala E reducida los resultados muestran que los niveles de dificultad se mantienen entre niveles fáciles (*b* =-1.178) en el ítem 49, hasta muy difícil (*b*=4.934) para el ítem 60. Estos valores muestran que la subescala E presenta mayores niveles de dificultad que la subescala D, la cual se mantiene en torno a la media del nivel de rasgo. Las estimaciones para el parámetro *a*, muestran que los ítems poseen una discriminación que alcanza desde niveles moderados (*a* =.53) hasta muy altos (*a* =4.86).

En cuanto al ajuste de los ítems, el estadístico de *S-*χ^2 muestra que en la mayoría de los casos se mantiene la hipótesis de ajuste, a excepción del ítem 52. En cuanto a los valores de RMSEA, todos los ítems presentan valores por debajo de .05.

Finalmente, el estadístico M_2 obtuvo un valor de 1011.39 con una probabilidad asociada de .001. Sin embargo, el estadístico RMSEA derivado de M_2 obtiene un valor de .05 y CFI (.963) y TLI (.951) por sobre .95, lo cual nos indica un ajuste adecuado del modelo a los datos.

La tabla 5 muestra los coeficientes alfa de Cronbach y las correlaciones entre las distintas versiones de la prueba. Como es posible observar, todas las versiones poseen coeficientes superiores a .70, incluso las subescalas D y E reducida, las cuales poseen un número muy reducido de ítems. Adicionalmente, todas las correlaciones entre la versión completa y el resto de las versiones poseen correlaciones altas. Cabe destacar que la versión reducida total (considerando los ítems de las subescalas D y E) obtiene una correlación de.885 con la versión completa, la cual es similar a valores reportados en estudios previos (Arthur & Day, 1994; Bors & Stokes, 1998).

Tabla 5. Correlaciones y coeficiente alfa de Cronbach entre las distintas versiones

	1	2	3	4
Versión completa (1)	(.834)			
Subescala D (2)	.760	(.731)		
Subescala E reducida (3)	.711	.382	(.707)	
Versión reducida total (4)	.885	.837	.825	(.782)

Nota: todas las correlaciones son significativa (p<.05); alfa de Cronbach en la diagonal principal entre paréntesis.

La figura 2 muestra la función de información de las subescalas D y E reducida. La función de información es un indicador de la precisión del instrumento para cada rango del continuo del nivel de rasgo, o lo que es lo mismo, la fiabilidad del instrumento para todo el nivel de rasgo de los sujetos (Muñiz, 2010). Esta función es construida a partir de las magnitudes de la discriminación (parámetro *a*) que aporta cada ítem a la prueba total.

Con respecto a la subescala D (panel superior de la figura 2), la mayor capacidad de discriminación se encuentra entre 2.5 y .5 desviaciones típicas por debajo de la media. Esto quiere decir que la mayor fiabilidad de la subescala D se encuentra desde el nivel mínimo hasta el tercer decil del nivel de habilidad aproximadamente. Por otro lado, la subescala E reducida (panel inferior de la figura 2) posee mayor nivel de discriminación entre los niveles -1 y +1 del nivel de rasgo, o lo que es lo mismo, posee la máxima fiabilidad entre los percentiles 15 y 85 aproximadamente.

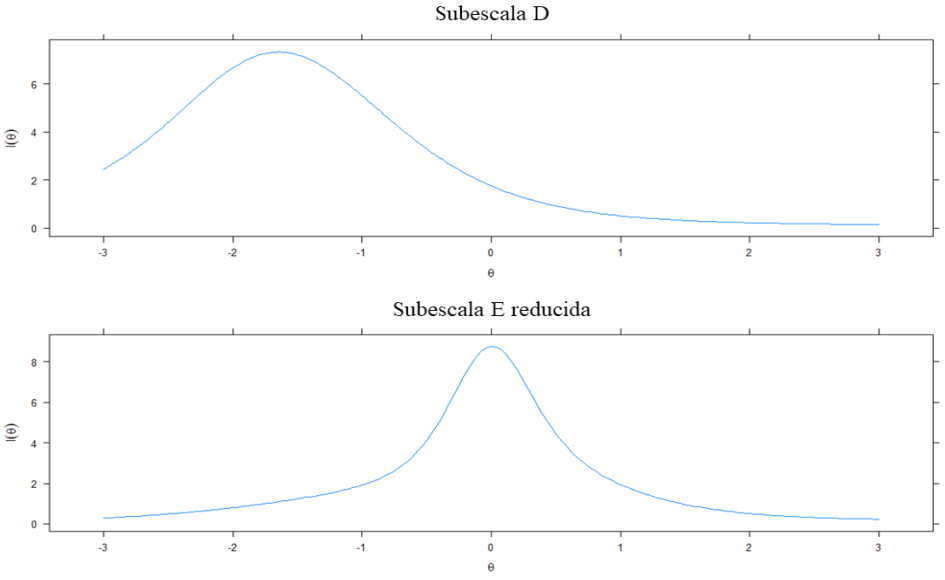

Figura 2. Función de información de la Escala D y Escala E reducida.

4.3. ANÁLISIS DE FUNCIONAMIENTO DIFERENCIAL (DIF)

Para contrastar una de nuestras hipótesis sobre el posible sesgo de género presente en las escalas MP, hemos llevado a cabo un análisis de funcionamiento diferencial en las escalas D y E reducida, utilizando como variable de agrupación el sexo de los estudiantes. Para contrastar esta hipótesis, hemos utilizado la prueba de Wald, la cual contrasta la hipótesis nula de igualdad de parámetros, *a* y *b*, de forma separada y conjunta. Si los datos son compatibles con la hipótesis nula, se puede asumir que tanto hombres como mujeres presentan los mismos parámetros *a* y *b*, lo cual significa que la probabilidad de acierto sólo depende del nivel de rasgo, y no de la pertenencia grupal. Estos hallazgos indicarían ausencia de funcionamiento diferencial. La tabla 6 muestra los parámetros estimados para hombres y mujeres, y los estadísticos Wald para cada ítem, tanto para el contraste de los parámetros por separado, como de manera conjunta. Los resultados muestran que únicamente en el ítem 38 de la subescala D, y los ítems 50 y 51 de la subescala E reducida presentan diferencias estadísticas significativas en el valor de los parámetros *b*. Por otro lado, sólo el ítem 57 de la subescala E reducida presenta diferencias significativas en el parámetro *a*. Dado que corresponde a una proporción muy baja de ítems, creemos que estos resultados no impactan en el funcionamiento diferencial del test.

Tabla 6. Parámetros estimados por sexo y prueba de Wald

	Estimaciones				Prueba de Wald					
	Hombres		Mujeres		Total		a		b	
Ítem	*a*	*b*	*a*	*b*	*χ2*	*p*	*χ2*	*p*	*χ2*	*p*
Ítem 37	1.070	-4.080	1.450	-2.920	2.200	.333	.700	.420	1.600	.213
Ítem 38	1.300	-2.210	2.300	-1.810	5.900	.053	4.700	.030	1.200	.278
Ítem 39	1.370	-2.080	1.960	-1.940	4.200	.125	2.100	.152	2.100	.149
Ítem 40	1.020	-2.180	1.440	-1.880	2.900	.235	2.000	.162	.900	.333
Ítem 41	1.840	-2.030	2.130	-1.790	.800	.667	.400	.550	.500	.501
Ítem 42	1.680	-1.590	1.490	-1.880	1.600	.452	.300	.612	1.300	.249
Ítem 43	1.030	-1.470	1.180	-1.240	.800	.686	.400	.551	.400	.528
Ítem 44	1.140	-1.500	1.000	-1.610	.300	.853	.300	.589	.000	.871
Ítem 45	1.100	-1.140	1.060	-1.230	.200	.923	.000	.866	.100	.716
Ítem 46	2.100	-1.190	1.990	-1.170	.200	.884	.100	.798	.200	.671
Ítem 47	.570	1.490	0.710	1.010	1.700	.429	.600	.451	1.100	.290
Ítem 48	.710	3.000	.440	4.300	1.800	.416	1.300	.263	.500	.480
Ítem 49	1.350	-1.210	1.420	-1.010	1.600	.452	.000	.824	1.500	.215
Ítem 50	1.840	-.350	.980	-.370	9.100	.011	8.500	.004	.500	.471
Ítem 51	1.800	-.540	1.200	-.540	4.800	.091	4.000	.046	.800	.371

Ítem 52	1.840	.290	1.670	.350	.400	.814	.300	.602	.100	.710
Ítem 53	1.900	.090	2.000	-.030	1.400	.488	.100	.801	1.400	.242
Ítem 54	1.180	.360	1.450	.360	1.200	.545	1.100	.300	.100	.714
Ítem 57	.160	12.270	.520	2.980	8.200	.017	2.800	.095	5.400	.021
Ítem 60	.590	4.540	.520	5.710	1.400	.488	.100	.822	1.400	.240

4.4. COMPARACIÓN DE LAS PUNTUACIONES ENTRE HOMBRES Y MUJERES

Una vez obtenida evidencia acerca de la ausencia de funcionamiento diferencial del test, hemos realizado una comparación de medias. El propósito ha sido evaluar la posible existencia de diferencias significativas entre hombres y mujeres. Para ello hemos llevado a cabo un análisis de ANOVA factorial con el propósito de incorporar las medias de las subescalas D y E reducida como un factor intra-sujeto, y así poder evaluar la posible interacción entre las subescalas y el sexo. Nuestro modelo corresponde a un modelo factorial sencillo, de 2 x 2, con una variable inter-sujeto (sexo) y una variable intra-sujeto (subescalas D y E). Los resultados muestran que no existe diferencias entre hombres y mujeres ($F_{(933,1)}$=.527; p>.05), ni efecto de doble interacción genero*subescala ($F_{(933,1)}$=.383; p>.05). Sin embargo, si encontramos diferencias significativas entre las subescalas ($F_{(933,1)}$=1467.852; p<.05). Como lo muestra la gráfica de la figura 3, la muestra global posee un rendimiento superior en la subescala D en comparación a la subescala E reducida, no existiendo diferencias entre hombre y mujeres. Las mayores puntuaciones de la subescala D obtenida por los estudiantes, son coherentes con la mayor magnitud de los parámetros de dificultad obtenidos para la subescala E. Estos resultados muestran que no existen diferencias en el rendimiento entre hombres y mujeres para las escalas analizadas, una vez descartada la existencia de funcionamiento diferencial del test por sexo.

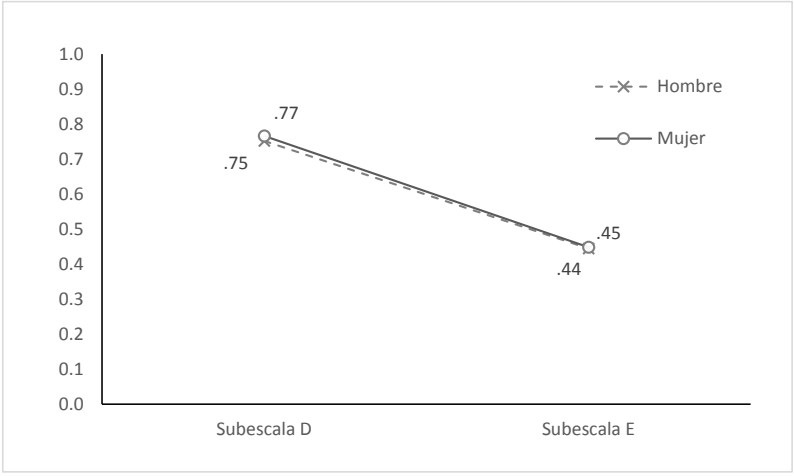

Figura 3. Gráfico de medias para la doble interacción Género*Subescalas.

5. DISCUSIÓN Y CONCLUSIONES

El objetivo principal de nuestra investigación fue realizar un análisis psicométrico de la escala Matrices Progresivas General de Raven, en una muestra de estudiantes que se encontraban participando en el proceso de admisión de un programa de talento académico promovido por una universidad chilena. El modelo utilizado fue la Teoría de Respuesta al Ítem (TRI), ya que los estudios que consideran el análisis de la escala Raven con esta clase de modelo son escasos y no existe evidencia sobre análisis psicométricos con población con talento académico. Este trabajo cobra especial relevancia debido a que este instrumento es uno de los utilizado en programas de talento académico en Chile. Adicionalmente, estos análisis han permitido la conformación de una propuesta de versión reducida del instrumento para su utilización en procesos de admisión de esta clase de programas.

Los resultados obtenidos permitieron la obtención de hallazgos interesantes. En primero lugar, los análisis de dimensionalidad mostraron que la escala total no ajusta al modelo de un factor. Estos resultados ya han sido reportados en la literatura previa, la que presenta evidencia que apoya la existencia de dos factores (Abad *et al*., 2004; DeShon *et al*., 1995; Dillon *et al*., 1981). Al contrastar la unidimensionalidad de las subescalas por separado, sólo las subescalas D y E obtuvieron un ajuste adecuado. Estos resultados son posibles de explicar desde las particulares características de la muestra. Dado que los participantes corresponden a estudiantes seleccionados por sus profesores y/o a estudiantes autoconvocados, es posible asumir que los participantes corresponden a personas con altas capacidades e intereses intelectuales y/o académicos. Ello ha generado un alto rendimiento en las subescalas por parte de los estudiantes, especialmente en los ítems de las subescalas A, B y C, las cuales presentan una dificultad baja. Las altas tasas de respuestas correctas en las subescalas con bajo nivel de dificultad (las cuales han estado en torno a 98%, 96% y 90% para las escalas A, B y C, respectivamente), genera baja capacidad de discriminación de estos ítems para diferenciar adecuadamente a estos estudiantes. Adicionalmente, las altas tasas de correctas, genera baja dispersión de las respuestas, generando dificultades para la adecuada estimación de las covariaciones de los ítems y para la adecuada recuperación de la dimensionalidad de estas subescalas.

En cuanto a la dimensionalidad de las subescalas D y E, estudios previos ya han advertido la presencia de dos factores (Abad *et al*., 2004; DeShon *et al*., 1995; Dillon *et al*., 1981). Por ejemplo, DeShon *et al*. (1995) demostraron que la presentación de los ítems estaba organizada según dimensiones y/o estrategias de respuesta. En concreto, mostraron que los ítems de las subescalas A, B y D correspondían a una dimensión analítico-verbal, mientras que los ítems de las subescalas C y E correspondían a una dimensión viso-espacial. A pesar de ello, creemos que estos resultados no son concluyentes debido al creciente nivel de dificultad en la presentación de los ítems a lo largo del instrumento. Es posible que los resultados que sugieren la presencia de estas dos dimensiones correspondan a pseudo factores producto de la dificultad de los ítems y no debido a dimensiones sustantivas. Este problema ha sido planteado hace bastante tiempo en la literatura metodológica (e.g. Bernstein, 1988; Ferrando & Lorenza-Seva, 1992). Más estudios son necesarios para diferenciar el efecto de pseudo factores de la real presencia de factores sustantivos subyacentes.

Respecto a los análisis desde el modelo TRI, los resultados son prometedores. En términos generales, tanto la subescala D como E poseen excelentes indicadores de discriminación y dificultad. En el caso de la subescala E, cuatro ítems han sido detectado con

problemas en la estimación de sus parámetros, los cuales han sido eliminado. Finalmente, la versión reducida propuesta ha quedado conformada por 20 ítem, 12 correspondientes a la subescala D y 8 ítems de la subescala E. Ambas subescalas han obtenido excelentes indicadores de ajuste tanto a nivel de ítems como a nivel global. Finalmente, las funciones de información muestran que la versión obtenida alcanza excelente fiabilidad, especialmente en los rangos bajos y medios, abarcando gran parte del nivel de rasgo. En cuanto a la relación entre la versión original y la versión reducida, el coeficiente de correlación de Pearson obtuvo un valor por sobre .80, lo cual es coherente con los resultados obtenidos por otras propuestas de versiones reducidas elaboradas en otros contextos (Arthur & Day, 1994; Bors & Stokes, 1998). Estos resultados convierten a la versión propuesta en una excelente opción para procesos de selección educativa que considere la evaluación de alto rendimiento intelectual, y en especial en procesos de admisión de programas de talento académico.

Dado el arduo debate respecto a las posibles diferencias en rendimiento intelectual entre hombre y mujeres, y al posible efecto de funcionamiento diferencias de los test de inteligencia, hemos llevado a cabo un análisis DIF para la versión propuesta. Los resultados muestran que un número reducido de ítems presenta funcionamiento diferencial, lo que no impactan en el funcionamiento diferencial del test. Estos resultados son apoyados por la ausencia de diferencias entre hombre y mujeres obtenida por el análisis de comparación de medias. En definitiva, la versión obtenida no se ve afectada por la existencia de funcionamiento diferencial del test debido al sexo, y las puntuaciones por subescala no presentan diferencias significativas entre hombre y mujeres.

Varias son las limitaciones de nuestro estudio, de las cuales debemos destacar dos que nos parecen las más relevantes. En primer lugar, las particularidades de la muestra de participantes dificultan la generalización de los resultados a otros contextos o usos del test. Dada la amplia utilización de la MPG de Raven en diferentes contextos y para diferentes usos, es necesario replicar este estudio en esos contextos y en otras muestras de participantes. Particularmente necesario es llevar a cabo estudios con población general, que permitan evaluar las propiedades psicométricas en una amplia muestra de estudiantes, lo cual permitirá poder acumular evidencias de validez del test y de su versión reducida para ampliar su utilidad a diferentes contextos y poblaciones. En segundo lugar, los resultados de este estudio no son concluyente respecto a dimensionalidad del test. Dado que las dimensiones recuperadas coinciden con la dificultad creciente en la presentación de los ítems, no es posible descartar la presencia de pseudo factores de dificultad que puedan estar afectando los resultados. Estudios posteriores deben enfocarse en la evaluación de evidencias de validez externa, relacionando las puntuaciones de cada dimensión con instrumentos que evalúen dimensiones viso-espaciales y analítico-verbales en específico. La obtención de resultados diferenciales en las correlaciones entre las distintas medidas puede contribuir a determinar la presencia de dimensiones sustantivas o efectos instrumentales de dificultad.

Dado que este estudio es limitado, al estar centrado en la evaluación de las propiedades psicométricas de la MPG en población con talento académico, son necesario más estudio orientado a la elaboración de normas de interpretación y posibles baremos de las puntuaciones del test para su interpretación. Como ha sido comentado anteriormente, es necesario desarrollar estudios con población general y amplias muestras de participantes, que permitan cubrir estos propósitos.

En cuanto al ámbito aplicado, dada las excelentes propiedades psicométricas de la versión reducida, proponemos la utilización de esta versión de 20 ítems para los procesos de admisión de programas de talento académico. A diferencia de la versión completa, la versión reducida presentaría ítems con importantes niveles de dificultad al inicio de la aplicación. Al igual como se propone para la versión avanzada, es posible considerar la aplicación previa de una de las tres primeras subescalas a modo de entrenamiento, no siendo incluida en la puntuación total de la versión. Ello podría evitar el efecto de la comprensión de la tarea en los primeros ítems de la versión reducida. Finalmente, dado que no existe, por lo pronto, norma de interpretación específica para la versión reducida, proponemos para su aplicación en procesos de admisión de programas de talento, la utilización de una puntuación base de 36 puntos. Esta recomendación se basa en la menor dificultad de las subescalas A, B y C, en las excesivamente altas tasas de correctas presentadas por la muestra de estudiantes y en la ausencia de capacidad de discriminación de estas subescalas para este tipo de estudiantes.

La utilización de esta versión reducida disminuiría considerablemente los recursos y tiempos de aplicación de los procesos de admisión de programas de talento. Adicionalmente, aumentaría la fiabilidad de las respuestas, ya que contribuiría a disminuir la fatiga y el tiempo de atención necesario para su cumplimentación, y aumentaría la validez de las puntuaciones, ya que estas estarían basadas en ítems seleccionados a través de criterios psicométricos, obtenidos en el contexto en el cual estas son utilizadas.

REFERENCIAS BIBLIOGRÁFICAS

Abad, F. J., Colom, R., Rebollo, I. & Escorial, S. (2004). Sex differential item functioning in the Raven's Advanced Progressive Matrices: Evidence for bias. *Personality and individual differences, 36*(6), 1459-1470. doi: 10.1016/s0191-8869(03)00241-1

Alderton, D. L. & Larson, G. E. (1990). Dimensionality of Raven's Advanced Progressive Matrices items. *Educational and Psychological Measurement, 50*(4), 887–900. doi: 10.1177/0013164490504019

Alvino, J., McDonnel, R. C. & Richert, S. (1981). National Survey of Identification Practices in Gifted and Talented Education. *Exceptional Children, 48*(2), 124–132. doi: 10.1177/001440298104800205

Arthur, W. & Woehr, D. J. (1993). A confirmatory factor analytic study examining the dimensionality of the Raven's Advanced Progressive Matrices. *Educational and Psychological Measurement, 53*(2), 471–478. doi: 10.1177/0013164493053002016

Arthur, W. & Day, D. V. (1994). Development of a Short form for the Raven Advanced Progressive Matrices Test. *Educational and Psychological Measurement, 54*(2), 394–403. doi: 10.1177/0013164494054002013

Attorresi, H. F., Lozzia, G. S., Abal, F. J. P., Galibert, M. S. & Aguerri, M. E. (2009). Teoría de Respuesta al Ítem. Conceptos básicos y aplicaciones para la medición de constructos psicológicos. *Revista Argentina de Clínica Psicológica, 18*(2), 179-188.

Bentler, P. M. (1992). On the fit of models to covariances and methodology to the Bulletin. *Psychological Bulletin, 112*(3), 400–404. doi: 10.1037/0033-2909.112.3.400

Bernstein, I. H., Garbin, C. P. & Teng, G. K. (1988). Applied Multivariate Analysis. doi: 10.1007/978-1-4613-8740-4

Bilker, W. B., Hansen, J. A., Brensinger, C. M., Richard, J., Gur, R. E. & Gur, R. C. (2012). Development of Abbreviated Nine-Item Forms of the Raven's Standard Progressive Matrices Test. *Assessment, 19*(3), 354–369. doi: 10.1177/1073191112446655

Bors, D. A. & Stokes, T. L. (1998). Raven's Advanced Progressive Matrices: Norms for First-Year University Students and the Development of a Short Form. *Educational and Psychological Measurement*, *58*(3), 382–398. doi: 10.1177/0013164498058003002

Burgaleta, M., Head, K., Álvarez-Linera, J., Martínez, K., Escorial, S., Haier, R. & Colom, R. (2012). Sex differences in brain volume are related to specific skills, not to general intelligence. *Intelligence, 40*(1), 60–68. doi: 10.1016/j.intell.2011.10.006

Burt, C. L. & Moore, R. C. (1912). The mental differences between the sexes. *Journal of Experimental Pedagogy 1*, 355-388.

Carpenter, P. A., Just, M. A. & Shell, P. (1990). What one intelligence test measures: A theoretical account of the processing in the Raven Progressive Matrices Test. *Psychological Review*, *97*(3), 404–431. doi: 10.1037/0033-295x.97.3.404

Cattell, R. B. (1971). *Abilities: Their Structure, Growth and Action*. Boston: Houghton Mifflin.

Chiesi, F., Ciancaleoni, M., Galli, S., Morsanyi, K. & Primi, C. (2012). Item Response Theory analysis and Differential Item Functioning across age, gender and country of a short form of the Advanced Progressive Matrices. *Learning and Individual Differences, 22*(3), 390–396. doi: 10.1016/j.lindif.2011.12.007

Colom, R., & Abad, F. J. (2007). Advanced progressive matrices and sex differences: Comment to Mackintosh and Bennett (2005). *Intelligence*, *35*(2), 183–185. doi: 10.1016/j.intell.2006.06.003

Colom, R., García, L. F., Juan-Espinosa, M. & Abad, F. J. (2002). Null Sex Differences in General Intelligence: Evidence from the WAIS-III. *The Spanish Journal of Psychology, 5*(1), 29–35. doi: 10.1017/s1138741600005801

Colom, R., Juan-Espinosa, M., Abad, F. y García, LF (2000). Diferencias de sexo insignificantes en inteligencia general. Inteligencia, *28*(1), 57–68. doi: 10.1016/s0160-2896(99)00035-5

Colom, R. & Lynn, R. (2004). Testing the developmental theory of sex differences in intelligence on 12–18 year olds. *Personality and Individual Differences, 36*(1), 75–82. doi:10.1016/s0191-8869(03)00053-9

Čvorović, J. & Lynn, R. (2014). Sex differences in intelligence: Some new data from Serbia. *Mankind Quarterly 55*(1/2), 101-109.

Dahl, G. (1972). *Reduzierter Wechsler-Intelligenztest* [Short version of the Wechsler Intelligence Test]. Meisenheim/Glan, Federal Republic of Germany: Main.

Deary, I. J., Whiteman, M. C., Starr, J. M., Whalley, L. J. & Fox, H. C. (2004). The Impact of Childhood Intelligence on Later Life: Following Up the Scottish Mental Surveys of 1932 and 1947. *Journal of Personality and Social Psychology, 86*(1), 130–147. doi: 10.1037/0022-3514.86.1.130

DeShon, R. P., Chan, D. & Weissbein, D. A. (1995). Verbal overshadowing effects on Raven's Advanced Progressive Matrices: Evidence for multidimensional performance determinants. *Intelligenc*e, *21*(2), 135–155. doi: 10.1016/0160-2896(95)90023-3

Dillon, R. F., Pohlmann, J. T. & Lohman, D. F. (1981). A factor analysis of Raven's Advanced Progressive Matrices freed of difficulty factors. *Educational and Psychological Measurement, 41*(4), 1295–1302. doi: 10.1177/001316448104100438

Escurra-Mayaute, L. M. & Delgado-Vásquez, A. E. (2010). Análisis psicométrico del Test de Matrices Progresivas Avanzadas de Raven mediante el Modelo de Tres Parámetros de la Teoría de la Respuesta al Ítem. *Persona, 0*(13), 71. doi: 10.26439/persona2010.n013.265

Eysenck, H. J. & Kamin, L. (1981). Rejoinder to Eysenck. *Intelligence: The Battle for the Mind.* 173–187. doi: 10.1007/978-1-349-05958-4

Fernández-Liporace, M., Ongarato, P., Saavedra, E. & Casullo, M. M. (2004). El Test de Matrices Progresivas, Escala General: un análisis psicométrico. *Evaluar, 4*(1), 50-69.

Ferrando, P. J. y Lorenzo, U. (1992). Extracción del componente de dificultad en la evaluación de escalas basadas en ítems dicotómicos. *Psicothema*, *4*(1), 269-276.

Flynn, J. R. (1984). The mean IQ of Americans: Massive gains 1932 to 1978. *Psychological Bulletin, 95*(1), 29–51. doi: 10.1037/0033-2909.95.1.29

_____. (1987). Massive IQ gains in 14 nations: What IQ tests really measure. *Psychological Bulletin, 101*(2), 171–191. doi: 10.1037/0033-2909.101.2.171

_____. (1998). IQ gains over time: Toward finding the causes. En U. Neisser (Ed), *The rising curve: Long term gains in IQ and related measures* (pp. 25-66). Washington, DC: American Psychological Association.

García Cepero, M. C., Proestakis, A. N., Lillo Olivares, A., Muñoz, E. M., López Valladares, C. & Guzmán Garay, M. I. (2012). Caracterización de estudiantes desde sus potencialidades y Talentos Académicos en la región de Antofagasta, Chile. *Universitas Psychologica, 11*(4), 1340. doi: 10.11144/javeriana.upsy11-4.cept

Gignac, G. E. (2015). Raven's is not a pure measure of general intelligence: Implications for g factor theory and the brief measurement of G. *Intelligence, 52*, 71–79. doi: 10.1016/j.intell.2015.07.006

Haier, R. J. (2007). Brains, bias and biology: Follow the data. In: Ceci, S. J. & Williams, W. M. (eds), *Why Aren't More Women in Science?* Washington, D. C.: American Psychological Association.

Halpern, D. (2012). *Sex Differences in Cognitive Abilities*, 4th edition. New York: Psychology Press.

Hambleton, R. K. Swaminathan. (1985). *Item response theory. Principles and applications*. Norwell, MA: Kluwer Academic Publishers.

Hayes, T. R., Petrov, A. A. & Sederberg, P. B. (2015). Do we really become smarter when our fluid-intelligence test scores improve? Intelligence, *48*, 1–14. doi: 10.1016/j.intell.2014.10.005

Herrnstein, R. & Murray, C. (1994). *The Bell Curve*. New York: Random House.

Irwing, P. & Lynn, R. (2005). Sex differences in means and variability on the progressive matrices in university students: A meta-analysis. *British Journal of Psychology, 96*(4), 505–524. doi: 10.1348/000712605x53542

Jensen, A. R. (1998). *The G Factor: The Science of Mental Ability*. Westport, CT: Praeger/Greenwood.

Lohman, D. F., Korb, K. A. & Lakin, J. M. (2008). Identifying Academically Gifted English-Language Learners Using Nonverbal Tests. *Gifted Child Quarterly, 52*(4), 275–296. doi: 10.1177/0016986208321808

Lord, F. M. (1980). *Application of Item Response Theory to Practical Testing Problems*. Hillsdale, N. J., Lawrence Erlbaum Ass.

Lubinski, D. (2000). Measures of intelligence: Intelligence tests. *Encyclopedia of Psychology, 5*. 139–144. doi: 10.1037/10520-069

Lynn, R. (1994). Sex differences in intelligence and brain size: A paradox resolved. *Personality and Individual Differences, 17*(2), 257–271. doi:10.1016/0191-8869(94)90030-2

_____. (2017). Sex Differences in Intelligence: The Developmental Theory. *Mankind Quarterly, 58*(1), 9–42. doi: 10.46469/mq.2017.58.1.2

Lynn, R. & Irwing, P. (2004). Sex differences on the progressive matrices: A meta-analysis. *Intelligence, 32*(5), 481–498. doi: 10.1016/j.intell.2004.06.008

Mackintosh, N. J. (1996). Sex differences and IQ. *Journal of Biosocial Science, 28*(4), 558–571. doi:10.1017/s0021932000022586

_____. (2011*). IQ and Human Intelligence.* 2nd edition. Oxford: Oxford University Press.

Mansilla, C., Vásquez, D. & Estrada, C. (2012). Pertinencia normativa del Raven para la evaluación de población infantojuvenil socialmente vulnerable. *Terapia Psicológica, 30*(1), 73–80. doi: 10.4067/s0718-48082012000100007

Martínez Arias, R. (1995). *Psicometría: Teoría de los tests psicológicos y educativos*. Madrid: Síntesis.

Matthews, D. J. (1988). Raven's matrices in the identification of giftedness. *Roeper Review, 10*(3), 159–162. doi: 10.1080/02783198809553115

Maydeu-Olivares, A. & Joe, H. (2005). Limited- and Full-Information Estimation and Goodness-of-Fit Testing in 2nContingency Tables. *Journal of the American Statistical Association, 100*(471), 1009–1020. doi: 10.1198/016214504000002069

_____. (2006). Limited Information Goodness-of-fit Testing in Multidimensional Contingency Tables. *Psychometrika, 71*(4), 713–732. doi: 10.1007/s11336-005-1295-9

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

_____. (2010). Las teorías de los tests: teoría clásica y teoría de respuesta a los ítems. *Papeles del psicólogo, 31*(1), 57-66.

Orlando, M. & Thissen, D. (2000). Likelihood-Based Item-Fit Indices for Dichotomous Item Response Theory Models. *Applied Psychological Measurement, 24*(1), 50–64. doi: 10.1177/01466216000241003

_____. (2003). Further Investigation of the Performance of S - X2: An Item Fit Index for Use With Dichotomous Item Response Theory Models. *Applied Psychological Measurement, 27*(4), 289–298. doi: 10.1177/0146621603027004004

Paul, S. M. (1986). The Advanced Raven's Progressive Matrices. *The Journal of Experimental Education, 54*(2), 95–100. doi: 10.1080/00220973.1986.10806404

Raven, J. C. (1936). Raven Standard Progressive Matrices. PsycTESTS Dataset. doi: 10.1037/t07027-000

_____. (1991). *Test de matrices progresivas para la medida de la capacidad intelectual (de sujetos de 12 a 65 años)*, Manual. Buenos Aires: Paidós.

Raven, J., Raven, J. C. & Court, J. H. (1998). *Raven Manual. General Overview*. Oxford: Oxford Psychologists Press.

Renzulli, J. S. (2016). The three-ring conception of giftedness, In S. M. Reis (Ed.). *Reflections On Gifted Education* (pp. 55-86). Waco, TX: Prufrock Press.

Rosas, R., Tenorio, M., Pizarro, M., Cumsille, P., Bosch, A., … Zapata-Sepúlveda, P. (2014). Estandarización de la Escala Wechsler de Inteligencia Para Adultos-Cuarta Edición en Chile. *Psykhe, 23*(1), 1–18. doi: 10.7764/psykhe.23.1.529

Savage-McGlynn, E. (2012). Sex differences in intelligence in younger and older participants of the Raven's Standard Progressive Matrices Plus. *Personality and Individual Differences, 53*(2), 137–141. doi: 10.1016/j.paid.2011.06.013

Spearman, C. (1927). *The abilities of man*. New York: McMillan.

Steiger, J. H. & Lind, J. M. (1980). Statistically based tests for the number of common factors. *Paper presented at the annual meeting of the Psychometric Society*, Iowa City, 1A.

Van der Ven, A. H. G. & Ellis, J. (2000). A Rasch analysis of Raven's standard progressive matrices. *Personality and Individual Differences, 29*(1), 45–64. doi: 10.1016/s0191-8869(99)00177-4

Waschl, Nettelbeck, T., Jackson, S. & Burns, N. (2016). Dimensionality of the Raven's Advanced Progressive Matrices: Sex differences and visuospatial ability. *Personality and Individual Differences, 100,* 157-166.

Wytek, R., Opgenoorth, E. & Presslich, O. (1984). Development of a New Shortened Version of Raven's Matrices Test for Application Rough Assessment of Present Intellectual Capacity within Psychopathological Investigation. *Psychopathology, 17*(2), 49–58. doi:10.1159/000284003