UNIVERSIDAD DE TALCA

FACULTAD DE PSICOLOGÍA

VALIDACIÓN DE UN TEST COMPUTARIZADO DE ROTACIÓN MENTAL A TRAVÉS DE LA TEORÍA DE RESPUESTA AL ÍTEM EN ESTUDIANTES UNIVERSITARIOS

Memoria

Para optar al Título de Psicólogo Mención Clínica

Alumno: Camila Valenzuela Fuentes

Profesores guía:

José Luis Ulloa Fulgeri

Agustín Martínez Molina

TALCA

Diciembre del 2018

CONSTANCIA

La Dirección del Sistema de Bibliotecas a través de su unidad de procesos técnicos certifica que el autor del siguiente trabajo de titulación ha firmado su autorización para la reproducción en forma total o parcial e ilimitada del mismo.

Talca, 2019

Agradecimientos

A mi mamá, que fue la persona que siempre creyó que sería capaz de lograr lo que me propusiera, y que dio la fuerza para lograrlo, aun sin estar físicamente presente.

A Marcelo, por todo su apoyo y su amor.

A mis profesores guía, el Dr. Agustín Martínez Molina, quien confió en mis capacidades cuando me quedaba más de la mitad de la carrera para egresar y al Dr. José Luis Ulloa, que apoyó mi trabajo en los momentos más complicados.

A mi familia, por las experiencias que compartimos juntos y que me convirtieron en la persona que actualmente soy.

Resumen

Introducción: La Rotación Mental (RM), se define como la habilidad de rotar representaciones mentales de imágenes u objetos (Shepard y Metzler, 1971). Al existir evidencia suficiente que demuestra que el mismo test no mide a los sujetos con la misma precisión (Muñiz, 1998), se presenta una alternativa para la validación del test computarizado de inteligencia, ERM a través de la Teoría de Respuesta al Ítem. ***Objetivo:*** Validar un test computarizado de Rotación Mental a través de la Teoría de Respuesta al Ítem y de la Teoría Clásica de los Test en estudiantes universitarios. ***Método:*** A través de la obtención de estadísticos descriptivos por dificultad según el concepto de discrepancia angular (Shepard y Metzler, 1971), su posterior análisis factorial exploratorio, AFE y la obtención de los indicadores del modelo 2PL de la TRI. ***Resultados:*** Existe una correspondencia entre los estadísticos descriptivos en aciertos y errores para cada dificultad propuesta, sin embargo, estos hallazgos no son reafirmados por el AFE. La TRI entrega información adicional sobre los parámetros de discriminación y dificultad de los ítems en sus clasificaciones. ***Discusión:*** El α de Cronbach de la escala general, se complementa con la TRI, donde la precisión varía a lo largo del rasgo medido, entregando indicadores que dan cuenta de las propiedades de cada ítem. La teoría discrepancia angular se diferencia de otros constructos con variables latentes de mayor sustento teórico.

Palabras clave: rotación mental, teoría de respuesta al ítem, análisis factorial exploratorio.

Abstract

Introduction: Mental Rotation (MR), can be defined as the ability to mentally rotate images or objects (Shepard y Metzler, 1971). Due there's enough evidence to show that the same test doesn't measure subjects with the same accuracy (Muñiz, 1998), a new alternative is presented to validate a computerized test of intelligence, ERM through Item Response Theory. ***Aim:*** Validating a computerized Mental Rotation Test using Item Response Theory and Classical Test Theory on college students. ***Method:*** Obtaining descriptive analysis per difficulty based on the concept of angular discrepancy (Shepard & Metzler 1971), its later exploratory factor analysis, EFA and the application of the IRT's 2PL model. ***Results:*** There's a correspondence between descriptive analysis in terms of success and failure per difficulty, however, these findings were not confirmed by EFA. TRI addresses additional information about items' difficulty and discrimination parameters and its classification. ***Discussion:*** Cronbach's α for the general scale is complemented with IRT, due its accuracy varies throughout the assessed trait, delivering indicators that account information for the properties of each item. The angular discrepancy theory is different from other constructs with latent variables with a greater theoretical support. The factorial structure is based on the theory of angular discrepancy, different from other psychological tests that measure latent variables with more theoretical support.

Key words: mental rotation, item response theory, exploratory factor analysis.

Índice

Introducción

El concepto de Rotación Mental (RM) se define como la habilidad de rotar representaciones mentales de imágenes u objetos (Shepard y Metzler, 1971). Esta habilidad es utilizada en tareas de la vida cotidiana, como la resolución de problemas y la comprensión del entorno (Meneghetti, Borella y Pazzaglia, 2014), y, si bien el empleo de las habilidades RM estaba apoyado en la literatura científica, ésta comenzó recién a ser evaluada a finales del siglo XIX (Boake, 2002). Importantes contribuciones comenzaron con el primer test de inteligencia desarrollado por Alfred Binet y Theodore Simon, que consistió en tareas visuoespaciales agrupadas en una subsecuente versión llamada Escala Stanford-Binet, la cual evalúa razonamiento verbal, razonamiento visual-abstracto, razonamiento cuantitativo y memoria de trabajo (Mora y Martín, 2007).

Shepard mostró que la RM podría ser estudiada a través de su Tarea (Peter y Battista, 2008), sin embargo, el estudio de la RM no ha estado basado solamente en las figuras compuestas por cubos publicadas por Shepard y Metzler (1971), sino que también en otras tareas como test computarizados y otras figuras objetivo en 3D y/o 2D (Kozhevnikov y Hegarty, 2001; Vandenberg y Kuse, 1978; Wright, Thompson, Ganis, Newcombe y Kosslyn, 2008).

Actualmente, la Rotación Mental ha sido estudiada cuidadosamente, debido a la posible transferencia de éstas habilidades, que, trasversalmente, podrían mejorar la inteligencia. Algunos estudios han mostrado una fuerte relación entre la RM con habilidades que implican el uso de ésta (Miyake, Friedman, Rettinger, Shah y Hegarty, 2001; Shah y Miyake, 1996), especialmente, en lo que respecta a la habilidad visual general (Gv), la cual es definida como la aptitud individual de almacenar, recuperar, transformar y generar imágenes mentales,

requiriendo la habilidad para codificar, recordar, transformar e identificar estímulos espaciales (McGrew, 2009; Lohman, 1979).

Por otra parte, en lo que respecta a la medición de los componentes de inteligencia—como lo es la Rotación Mental—, la Teoría de Respuesta al Ítem, TRI, es frecuentemente considerada como una aproximación psicométrica moderna en comparación a la Teoría Clásica de los Test, TCT, a pesar de haberse originado décadas atrás con el trabajo de Thurstone en los años 20, Rasch en los años 60 y de Lord en los 50 (Kean y Reilly, 2014), donde los aportes principales se centran en la manera de interpretar la medición, ya que la TRI muestra la relación entre la habilidad o el rasgo medido por el instrumento (θ), y la respuesta del sujeto dado cierto ítem, asumiendo, además, que no todos los ítems miden a los sujetos de la misma manera (Brzezinka, 2018)

A diferencia de la TRI, la teoría clásica de los test, TCT, se muestra relativamente simple y no cuenta con un modelo teórico complejo para relacionar la habilidad del sujeto para acertar cierto ítem. De esta forma, TCT, la TRI, por ejemplo, contempla que el éxito de un sujeto en relación a los elementos de la escala es conocido como el índice de dificultad (β). Por otra parte, la propiedad del ítem para discriminar entre niveles más altos de rasgo y niveles más bajos de rasgo en un sujeto se conoce como índice de discriminación y se denota α (Abedalaziz y Leng, 2018). En términos simples, la TRI busca entregar información sobre la manera que funciona un ítem para medir cierto rasgo o variable latente en los sujetos, comprendiendo que estos ítems entregarán información de distintos continuos del este rasgo o variable latente que se pretende medir, y que, además, algunos de ellos serán más discriminativos que otros, ya que entregarán

mayor variabilidad de respuesta, y que algunos ítems serán más difíciles que otros, es decir, se necesitarán mayores niveles de rasgo para alcanzar mayores probabilidades de acertar.

Según Reise (2014), desde una perspectiva aplicada, los modelos de la TRI fueron desarrollados con el objetivo de solventar problemas prácticos del mundo real, ya que, desde el inicio de este siglo que este enfoque de medición ha sido más utilizado en psicología y en constructos relacionados a la medicina.

En relación a los avances y las ventajas que implicarían el utilizar la teoría de respuesta al ítem, estudios basados en esta aproximación metodológica podrían contribuir a desarrollar nuevos modelos teóricos que podrían explicar mejor el funcionamiento de la RM y su relación con la inteligencia a través de herramientas más eficientes de medición. Por ejemplo, un Test Adaptativo Computarizado—en inglés Computer Adaptative Test, CAT— toma ventaja de la TRI para proveer solo los ítems necesarios para determinar el nivel de rasgo en un sujeto, seleccionando adaptativamente el ítem siguiente que debería ser administrado basándose en la respuesta del ítem anterior, examinando la habilidad del participante en menor tiempo (Kean y Reilly, 2014).

En lo que respecta a la importancia de la RM y su relación con otras habilidades, Newcombe (2010), después de una exhaustiva revisión, señaló que el razonamiento visuoespacial es una habilidad fundamental en las disciplinas de ciencia, tecnología, ingeniería y matemática (STEM). De acuerdo a su afirmación, una persona con un puntaje alto en pruebas visuoespaciales tiende a escoger una disciplina STEM para continuar estudios superiores.

Los test de RM pueden ser de naturaleza dinámica a diferencia de otros tipos de pruebas de inteligencia. En relación a lo anterior, Pellegrino y Hunt (1989), definieron el Razonamiento Espacial Dinámico—en inglés (DSR)—, estableciendo la diferencia entre los test dinámicos y

otros tipos de test. Ellos identificaron un factor que representa la habilidad de trabajar con estímulos de trayectoria de una dirección previamente establecida, es decir, en pruebas donde se posiciona al sujeto en el espacio y se espera que llegue a un objetivo, utilizando su capacidad para rotar imágenes.

En general, el estudio, tanto de la inteligencia como de sus componentes, además de la manera en la que éstos pueden ser medidos se ha mantenido vigente por larga data. Específicamente, la inteligencia ha intentado ser definida durante el transcurso de los años, y más que llegar a un consenso que abarque todas las posibles definiciones en una sola, las maneras de describir la inteligencia se han ido multiplicando con el paso del tiempo (Legg y Hutter, 2007).

1. Definiendo el concepto de inteligencia

Según Sternberg, parece ser que hay tantas definiciones de inteligencia como expertos a los que se les pide definirla (Gregory y Zangwill, 1987), y, efectivamente, a pesar de los años de investigaciones y debate en torno a esta temática, aun no hay una definición estándar o universal (Legg y Hutter, 2007).

Gardner la define como la habilidad de resolver problemas o crear productos que son valiosos en uno o más contextos culturales; Terman, como la habilidad de llevar el pensamiento abstracto; Thurstone como un rasgo mental, capaz de utilizar la abstracción para modificar el comportamiento, buscando ventajas individuales como un animal social; Wechsler, como un concepto global, que implica la habilidad del individuo para pensar racionalmente, para actuar a propósito y para lidiar efectivamente con el medio ambiente; Yerkes, como un complejo ensamblaje de funciones interrelacionadas que no son ni precisa ni completamente conocidas en el hombre (Legg y Hutter, 2007).

Por otra parte, la historia de la investigación de la inteligencia, ha puesto en evidencia que el éxito de una persona en una carrera, puesto de trabajo o la vida en general, no depende solamente de su CI, sino que también depende de otros factores personales, por lo que también existe la perspectiva de inteligencia que comprende tanto componentes cognitivos como emocionales y/o sociales (Derksen, Kramer y Katzko, 2002).

Galton (1816), en su publicación "Hereditary genius: an inquiry into its laws and consequences" hacía alusión a sus esfuerzos por querer demostrar que la inteligencia tenía un componente innato y hereditario. Desde su punto de vista, las personas consideradas como más inteligentes, eran quienes tenían mejor posición social y económica, puesto que la reputación, en cierto grado, podía ser una prueba de estas habilidades. Así, entendía la inteligencia como una manera de

clasificar a las personas, donde los resultados de sus estudios demostraron que ésta tenía una distribución normal, ya que quienes pertenecían a los valores extremos, es decir, niveles de inteligencia muy altos o muy bajos, siempre eran minorías, mientras que la mayor parte de las personas tenían un nivel de inteligencia muy cercano a la media.

De esta forma, las definiciones de inteligencia y el entendimiento de ésta, se han mantenido en constante desarrollo en diversos campos de estudio. Un ejemplo de ello, es la definición de la Inteligencia Competitiva en el ámbito organizacional, donde se entiende como una disciplina de gestión que permite que los ejecutivos tomen decisiones más exitosas, minimizando el riesgo y logrando un rendimiento óptimo desde la primera vez, con el objetivo de mantenerse vigentes en un ambiente competitivo (Nikolaos y Evangelia, 2012).

De acuerdo a Gardner (2000), cada sociedad busca su ser humano ideal, es decir, no podemos dejar ausente el concepto de la cultura para entender la inteligencia, así como tampoco los diversos contextos en los que, diferentes características darán cuenta de una persona inteligente. Durante los últimos siglos, en las culturas occidentales, se ha tenido la idea de lo que es la persona inteligente, la cual ha ido variando a medida que pasan los años. En lugares como las escuelas, una persona inteligente es quien es capaz de dominar áreas como la matemática, lenguaje, entre otros requerimientos propios del lugar. En contextos de negocios, el ideal de persona inteligente podría estar en alguien que anticipe oportunidades comerciales o que tome riesgos de forma mesurada. De cualquier manera, la inteligencia puede ser entendida de múltiples maneras (Gardner, 2000), y es por ello que siempre representa un campo en el que se puede incursionar y avanzar en el área investigativa.

Por otra parte, a pesar de la diversidad de conceptos relacionados, la inteligencia ha sido un constructo estudiado desde hace mucho tiempo atrás y que cuenta con una vasta historia, tanto en el desarrollo de test como en su aplicación e importancia en los últimos siglos.

2. Historia de las pruebas de inteligencia

La finalización del siglo XIX e inicios del siglo XX, trae consigo al menos dos corrientes en el estudio y medida de la inteligencia. La primera de ellas, iniciada por Galton (1869), donde el concepto de capacidad intelectual se entendía a través de procesos sensoriales y mentales simples. Lo anterior se desprende de su aporte en la elaboración de pruebas para medir parámetros psicofísicos como tiempos de reacción, agudeza visual, capacidad respiratoria, entre otros procesos. A partir de esto, intentaba encontrar una medida de inteligencia (Mora y Martín, 2007).

En Estados Unidos, James McKeen Cattell (1890), acuña el término "test mental", adaptando las pruebas desarrolladas por Galton con objetivos investigativos con alumnos universitarios estadounidenses. Algunas de las pruebas de retención de dígitos, que fueron utilizadas en estudios de Galton en Inglaterra, demuestran que este tipo de pruebas—las que, actualmente están presentes en las escalas de Wechsler—, se remontan a periodos anteriores a las que fueron desarrolladas por Binet y Simon (Boake, 2002).

Por otra parte, Binet es quien desarrolla el concepto de inteligencia desde una mirada distinta, ya que intenta explorar procesos mentales de orden superior como las imágenes mentales, la comprensión, la memoria, entre otros (Mora y Martín, 2007), los cuales se entienden como procesos de mayor complejidad cognitiva.

En el año 1905, Alfred Binet y Theodore Simon publican su escala, la que incluye tanto, sub-pruebas desarrolladas por otros—como, por ejemplo, la retención de dígitos—, más otras pruebas desarrolladas por Binet y colaboradores, donde la validez de esta escala fue demostrada de la misma manera en la que se evidenció la validez de las subpruebas en los estudios de Jacobs (1887)—quien era un historiador que trabajó con Galton en Inglaterra—, principalmente, a

través de la observación del incremento de puntajes a medida que aumentaba la edad y por las propiedades de la escala para poder diferenciar a niños normales de niños que presentaban daño cognitivo (Peterson, 1925).

En el año 1908, Binet y Simon hacen una revisión de su escala agrupando las pruebas en niveles etarios—procedimiento que, años más tarde, es conocido como escala etaria—, donde el rendimiento de la mayor parte de los niños fue satisfactorio. Un ejemplo de esto es, el acomodar las versiones de retención de dígitos de diferente duración por edad, donde las versiones más largas se le administraron a niños mayores, mientras que las versiones más cortas fueron administradas a niños de menor edad (Boake, 2002).

La inteligencia del niño era cuantificada en términos de su nivel intelectual, que se definía como el nivel de edad más alto en el que el niño completaba las pruebas de manera satisfactoria, lo que más tarde fue llamado edad mental (Boake, 2002).

Goddard, quien era el director de investigación de una escuela de entrenamiento en Vineland—el cual, a su vez, era un centro residencial en Nueva Jersey para niños con trastornos cognitivos—, supo de la escala creada por Binet y Simon luego de un viaje que hizo por Europa, encargándose más tarde de traducirla al inglés, contribuyendo a la popularización de la administración de escalas de inteligencia, lo que fue un determinante para el uso de la escala de Binet y Simon en instituciones estadounidenses (Zenderland, 1998).

Poco tiempo antes de que Estados Unidos entrara a la Primera Guerra Mundial, se hicieron dos principales revisiones a la escala desarrollada por Binet y Simon. La primera revisión decantó en lo que fue llamado la Escala de Examinación de Puntajes Yerkes-Bridges (Yerkes, Bridges y Hardwick, 1915), la que se caracterizaba por agrupar ítems similares en contenido en un

número más corto de sub-escalas, empezando por los ítems más fáciles, procediendo con el orden en base a la dificultad, hasta completar el test.

Por otra parte, la segunda revisión estuvo a cargo de Lewis Terman (1916), de la Universidad de Stanford, quien extendió el rango de aplicación de la escala a la adultez, a través del reemplazo de la edad mental por el coeficiente intelectual—IQ en inglés, CI en español—, lo cual provocó que, rápidamente, fuese la versión dominante para medir la inteligencia en Estados Unidos, en lo que se terminó conociendo como la escala Stanford-Binet.

La escala Stanford-Binet, fue, por consiguiente, una versión modificada de la escala que originalmente crearon Alfred Binet y Theodore Simon, ya que el gobierno francés le había solicitado a Binet que pudiese desarrollar una forma de detectar niños con retraso mental o con capacidades intelectuales que se encontrasen bajo la media, creando por primera vez una línea base de la inteligencia (Boake, 2002). Lewis Terman (1916), también probó métodos adicionales para publicar oficialmente la versión modificada del test, llamada "La Medida de la Inteligencia: Una explicación y una guía completa para el uso de la revisión de Stanford y Ampliación de la Escala de Inteligencia de Binet-Simon".

La noción de la inteligencia desde procesos de orden superior se refleja entonces, tanto en la escala desarrollada por Binet y Simon, como en la escala de inteligencia Standford-Binet (1916). Esta escala, fue construida con el objetivo medir éstas capacidades cognitivas que se entienden como de nivel superior, diagnosticando deficiencias en el desarrollo de los niños a nivel intelectual. Esta prueba medía cinco factores: procesamiento visuoespacial, memoria de trabajo, conocimiento, razonamiento cuantitativo y razonamiento fluido.

La necesidad por desarrollar pruebas de inteligencia surge desde diversas barreras, una de ellas es la idiomática, o la falta de habilidades lingüísticas por parte de sujetos que presentan daño a

nivel cognitivo o retraso mental. Y, en vista de que este aspecto podía ser criticado en algunas sub-pruebas utilizadas en la escala Stanford-Binet, Healy y Fernald (1911), argumentan que sus test fueron construidos para poder apartar las capacidades mentales de las experiencias del individuo en el entrenamiento formal a nivel lingüístico, o de cualquier lenguaje. En relación a lo anterior, enfatizaron que sus sub-pruebas fueron diseñadas para para ser resueltas simplemente con las capacidades de razonamiento. A este método de medir la inteligencia, a través del uso de tareas no verbales, se le conoce con el término de Prueba de Rendimiento.

Otra de las pruebas de larga data, que se centra en el desempeño sin que algún tipo de verbalización tenga alguna implicancia, y que, actualmente se encuentra en las Escalas de Wechsler, es la prueba de Claves, la cual probablemente desciende de un test de substitución desarrollado por la Universidad de Wisconsin, en el que nueve dígitos eran asociados a diferentes símbolos, siendo estas pruebas incluidas en baterías para niños (Healy y Fernald, 1911; Pyle, 1913).

El uso de las pruebas de inteligencia durante la Primera Guerra Mundial en el reclutamiento de soldados en Estados Unidos, fue, probablemente uno de los hitos más importantes dentro de la importancia de las baterías de inteligencia, a través de lo que fue el desarrollo de los test Alfa y Beta, impusado por Yerkes (1921), quien era el presidente de la Asociación Americana de Psicología durante la época en la Estados Unidos entró a Primera Guerra, y que además, fue el encargado de crear un comité de expertos para decidir si los reclutas del ejército eran adecuados para el servicio militar. Los test Alfa y Beta, fueron administrados utilizando el formato de la selección múltiple en dos grupos de examinación. Por una parte, se encontraba el grupo Alfa, el cual fue designado para la evaluación de angloparlantes, mientras que, por otra parte, el grupo Beta fue designado para la minoría de reclutas que o bien, eran inmigrantes, o que no sabían

leer, por lo que existió un paralelismo entre las Pruebas de Rendimiento y los roles establecidos por las escalas de Binet y Simon en las pruebas individuales de inteligencia, donde ambas escalas puntuaban una serie de sub-test que podrían ser administrados en menos de una hora. Se estimó que desde 1917 a 1919, las examinaciones de los tests Alfa y Beta se administraron a casi dos millones de reclutas (Yerkes, 1921; Boake, 2002). Los grupos examinados fueron la fuente mayor de subescalas e ítems utilizados en la escala Wechsler-Bellevue (Wechsler, 1939), entre las que destacan, del test Alfa, el subtest de aritmética donde verbalmente se reproducen problemas matemáticos; el subtest de comprensión y la subprueba de Información.

Durante el tiempo de la guerra, Wechsler termina su tesis en la Universidad de Columbia y comienza a trabajar en un campamento del ejército en Long Island, donde se sometió al test Alfa. Posteriormente, se enlista en el ejército de los Estados Unidos en el verano de 1918, para ser parte de los examinadores psicológicos en Georgia. Estas experiencias fueron fuente de inspiración para la escala que desarrolló más tarde, debido a que fue el encargado de examinar a los reclutas que habían fallado ambos test—Alfa y Beta—. Fue allí cuando relata que, piensa en combinar sub-pruebas verbales y no verbales en una sola escala de inteligencia (Wechsler, 1979).

Wechsler, a través de su experiencia como examinador, había notado varios casos en los que la escala de Stanford-Binet había arrojado una edad mental mucho menor para algunos reclutas que se habían desenvuelto de buena manera antes de enlistarse al ejército, lo cual atribuía a el énfasis que les daba la escala a las habilidades verbales adquiridas a través de la educación formal (Boake, 2002).

Después de la guerra, la credibilidad de los test de inteligencia se consolidó, provocando que muchos psicólogos fuesen formados, al menos en Estados Unidos, en interpretación y aplicación

de escalas de inteligencia. Además, se crearon otros tantos test como la prueba de Diseño de Bloques (Kohs, 1923), que consistía en reconstruir un diseño con los cubos de colores, luego de separarlo en unidades. Esta tarea fue correlacionada, tanto con la escala Stanford-Binet, así como con otros indicadores de inteligencia general, y a pesar de que Kohs menciona estos antecedentes, nunca se habla de algún concepto como percepción visuoespacial.

Para el año 1932, Wechsler se convierte en el psicólogo jefe en un hospital psiquiático de Nueva York y esta experiencia lo convence aún más sobre la necesidad de desarrollar una escala alternativa a la de Binet, que estuviese estandarizada y que fuese más adecuada para su uso con adultos (Wechsler, 1979), por lo que remplaza la manera en la que se calculaba el CI con el puntaje de desviación, a través de la transformación de la suma de los puntajes de los test en un puntaje estándar, usando la media y la desviación estándar en cada nivel etario, cambiando la edad mental cronológica en un puntaje estandarizado (Thorndike y Lohman, 1990).

Otro aporte importante de Wechsler para la escala Wechsler-Bellevue, fue la incorporación de sub-test verbales y pruebas de desempeño dentro de la misma, permitiendo la exploración de las dos aproximaciones contemporáneas principales de la aplicación de las pruebas de inteligencia. Esta escala contaba con pruebas extraídas de los test utilizados en el ejército. La batería de pruebas contaba con cinco sub-pruebas tradicionales más una sexta prueba, Construcción con cubos. Por otra parte, se eliminó de la sub-prueba de Analogías y Vocabulario, ya que él mismo había criticado la gran influencia de los aprendizajes adquiridos en educación formal en la escala Stanford-Binet. El sub-test de Vocabulario fue añadido más tarde, luego de desarrollar ítems extra, como un sub-test verbal alternativo, conocido actualmente como la sub-prueba de Comprensión (Wechsler, 1939).

Comparando la escala Wechsler-Bellevue con la escala Stanford-Binet, las ventajas presentadas por la primera de ellas fueron reconocidas por la comunidad de psicólogos como válidas, ya que las sub-pruebas estaban organizadas en escalas verbales y de desempeño, permitiendo su interpretación y administración por separado. Además, el uso de puntajes de desviación proporcionó una base estadística para interpretar los sub-test (Boake, 2002).

Años más tarde, Estados Unidos entra a la Segunda Guerra Mundial, necesitando una escala individual para poder hacer la selección, por lo que una forma alternativa de la escala Wechsler-Bellevue fue creada, conocida como la Escala de habilidad mental de Wechsler (1946).

En el año 1949, una revisión de la escala Wechsler-Bellevue fue publicada como la Escala de Inteligencia de Wechsler para Niños, WISC (Wechsler, 1949), la cual contó con algunos cambios como la incorporación de versiones más tempranas de las sub-pruebas de la escala para adultos. La muestra utilizada para la estandarización fue de 200 sujetos—100 hombres y 100 mujeres—, de edades entre los 5 y los 15 años.

La escala Wechsler-Bellevue ha pasado por cuatro revisiones a la fecha, culminando con su última versión, la Escala de Inteligencia de Wechsler para Adultos, WAIS-IV (Wechsler, 2008), la cual actualmente se compone de cuatro índices—Comprensión Verbal, Razonamiento Perceptual, Memoria de Trabajo y Velocidad de procesamiento— con sus respectivas sub-pruebas, y contando, además, con una validación chilena (Rosas, Tenorio, Pizarro, Cumsille, Bosch, Arancibia, Carmona-Halty, Pérez-Salas, Pino, Vizcarra y Zapata-Sepúlveda, 2014).

El campo de los test de inteligencia se fue ampliando con el tiempo, a través del uso de recursos y nuevas tecnologías que emergieron, como, por ejemplo, el uso masivo de computadoras, lo cual permitió contar con herramientas que facilitaron tanto la recogida de datos como la precisión en su medición. Un ejemplo de lo anterior son las versiones del Test Dinámico

Espacial en todas sus versiones (Colom, Contreras, Botella, y Santacreu, 2002; Colom, Contreras, Shih, y Santacreu, 2003; Santacreu, 1999), para medir Rotación Mental.

En relación a lo anteriormente mencionado, Mead (1993), luego de un meta-análisis descubrió que las correlaciones para los test cognitivos que un límite de tiempo o velocidad para su medición, contaban con correlaciones moderadas. Lo anterior se podía comprender debido a que las habilidades motoras de el o los examinadores, son importantes al momento de aplicar el test, así como también del sujeto que ejecuta, debido a que, por ejemplo, marcar con un círculo la alternativa de respuesta correcta, podría implicar más tiempo que presionar una tecla. En relación a lo anterior, la Teoría de Respuesta al Ítem, TRI, entrega información complementaria a los modelos que aporta la teoría clásica, dando cuenta de cómo la medición de un rasgo y sus observaciones no se comportan de la misma manera en los sujetos, sino que más bien, pueden ser adecuados o no para medir cierto grado de habilidad (Leenen, 2014).

3. Teoría de respuesta al ítem y Teoría Clásica: Aportes para el estudio de la RM

Si bien es sabido que indicadores que dan cuenta tanto de la fiabilidad como la validez de los instrumentos utilizados en psicología, nos entregan información acerca de la manera en la que debemos entender su aplicación y su fundamento teórico, alguno de los índices utilizados en la metodología, como, por ejemplo, el alfa de Cronbach, bajo el supuesto de la Teoría Clásica (TCT)—en inglés Classical Test Theory—, tienen importantes limitaciones, asumiendo que la fiabilidad es la misma para cada sujeto. Respecto a este tema, hay evidencia suficiente que demuestra que el mismo test no mide a los sujetos con la misma precisión (Muñiz, 1998).

El Análisis Factorial, que se basa en la teoría clásica de medición, y se comprende como una amplia variedad de procedimientos estadísticos que buscan determinar dimensiones a partir de conjuntos de información—definidos como observaciones o espacio de atributos—, tienen como objetivo definir una problemática científica específica, explorar y definir nuevos conocimientos (López-Roldán y Fachelli, 2016). En este sentido, los factores son entendidos como variables latentes, las cuales son constructos o conceptos que, si bien tienen fundamentación en la literatura y que se entienden como fenómenos existentes en las ciencias sociales, no son posibles de medir de la misma manera que se miden los fenómenos de las ciencias naturales, los cuales cuentan, en muchas ocasiones con herramientas especializadas para su medición (Muthén y Asparouhov, 2015). A diferencia de las observaciones en ciencias naturales, las variables latentes adquieren su nombre porque no pueden ser observadas directamente, sino que son inferidas a través de modelos matemáticos estructurados, los cuales varían dependiendo de los requerimientos de la investigación (Mai, Zhang y Wen, 2018).

Por otra parte, la TRI, o Teoría de Respuesta al Ítem, es un modelo probabilístico que intenta explicar cómo es la respuesta de un sujeto *j,* con habilidad θ_j que responde un ítem cualquiera *I* con una dificultad b_i, condicionada a la dificultad de cada ítem y también, a las capacidades de la persona (Hambleton, Swaminathan & Rogers, 1991). Este modelo puede ser entendido como una función de la habilidad del sujeto θ, el cual tiene tres parámetros. Primero, el parámetro de dificultad b_1, el cual controla la posición de la respuesta en la escala θ. El ítem con mayor dificultad está localizado a la derecha de θ, y requiere mayor capacidad o mayores habilidades para resultar en la misma probabilidad de logro. Segundo, el parámetro de discriminación α_I, el cual controla la pendiente de respuesta de la función. Mientras más discriminante es el ítem, la pendiente se muestra más empinada. Y, por último, el tercer parámetro es el de adivinación, el que indica la probabilidad de seleccionar la respuesta correcta de forma aleatoria (Van der Linden, 2010). En otras palabras, la TRI entrega información sobre la manera que funciona un ítem para medir cierto rasgo o variable latente en los sujetos, comprendiendo que estos ítems u observaciones entregarán información de distintos continuos del este rasgo o variable latente que se pretende medir, y que, además, algunos de ellos serán más discriminativos que otros, ya que entregarán mayor variabilidad de respuesta, y que algunos ítems serán más difíciles que otros, es decir, se necesitarán mayores niveles de rasgo para alcanzar mayores probabilidades de acertar.

En relación a lo anterior, existen tres modelos logísticos para los indicadores previamente mencionados. El primero de ellos, conocido como 1PL o modelo de Rasch, que asume que la adivinación es parte del rasgo y que todos los ítems tienen discriminaciones equivalentes ($\alpha=1$), por lo que todos los ítems son descritos solo con el parámetro θ. Por otra parte, está el modelo

2PL, que no contempla el parámetro de adivinación, ya que asume que es parte del rasgo como en el modelo de Rasch, pero sí considera los parámetros de discriminación y dificultad. Por último, está el modelo 3PL que contempla los parámetros de discriminación, dificultad y adivinación (Karadavut, 2017).

Los modelos de TRI pueden generar curvas características. La curva característica del ítem grafica la probabilidad de que un sujeto responda correctamente a un ítem en función de la variable latente medida por el test. Los valores del eje X en una curva característica del ítem, CCI, representan a la variable latente o a lo que se conoce como el rasgo medido, lo que se puede observar en la Figura 1. El eje Y representa la probabilidad de éxito de un sujeto que responda el test, por lo que, a medida que el rasgo necesario incrementa, la probabilidad del sujeto de responder correctamente aumentará (Bichi y Talib, 2018).

Lo que se observa en el ejemplo graficado en la Figura 1, es que este ítem en particular va aumentando la probabilidad de acierto—lo cual se evidencia en el eje Y— para niveles más altos del rasgo o variable latente observada para los sujetos—lo que se muestra en el eje X—. Se observa que, aproximadamente en el nivel medio del rasgo—lo cual se representa por el valor cero en el eje X—, los valores del continuo de la variable latente medida o de la habilidad de los sujetos permiten una posibilidad de acierto aproximada del 50%, la cual, además, aumenta rápidamente al aumentar los valores del rasgo, lo que da cuenta de que, este ítem es sensible a mayores cantidades de habilidad observada especialmente a partir desde la media.

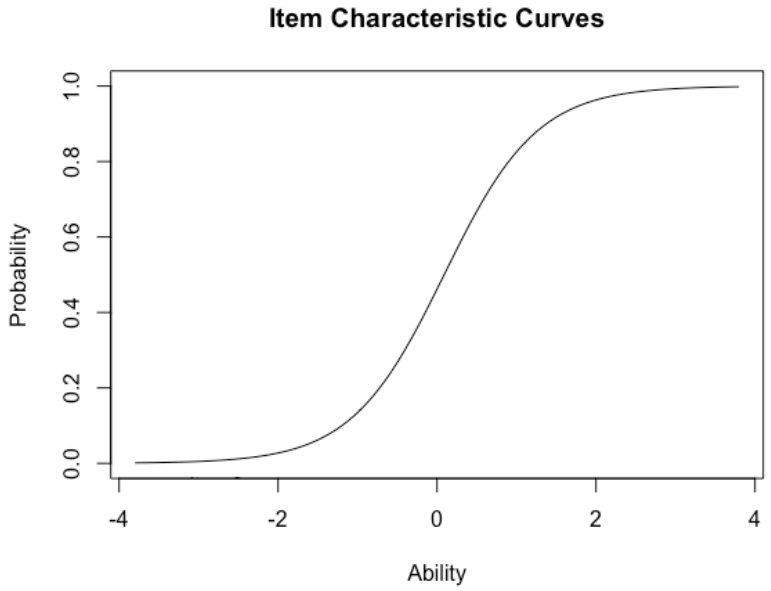

Figura 1. Ejemplo de curva característica del ítem.

Según algunos autores como Hidalgo-Montesinos y French (2016), para que un ítem funcione bien, debe discriminar a los sujetos en todos los niveles del rasgo, lo que se traduciría en un valor relativamente alto para el parámetro de discriminación (α), y, además, deben reunir cumplir con el un umbral mínimo de α para que el ítem funcione de manera aceptable, una revisión de expertos para los ítems que no cumplan el criterio o que se encuentren cercano a cumplirlos, la evaluación de las curvas características de información—las cuales no deberían ser planas—, entre otras consideraciones, como la teoría en la que se basa la escala.

Siguiendo los criterios de evaluación de Baker (2001), el umbral mínimo de discriminación debería ser α>,65, donde α>1,34 sería un ítem con un funcionamiento elevado y α>1,69 indicaría un ítem con un funcionamiento muy elevado. De la misma manera, un ítem que no cumpla con los requerimientos necesarios o esté a punto de alcanzar el umbral mínimo, puede ser revisado por un experto en el tema, luego de evaluar las curvas de información para así,

procurar conservar ítems que no se grafiquen planos o con forma de línea recta paralela al eje X.

La ventaja de usar TRI es que podría mejorar, por ejemplo, la precisión de la evaluación en las diferencias de sexo en las tareas de rotación mental, debido a que éstas diferencias han encontrado grandes tamaños de efecto (Halpern y LaMay, 2000), con resultados que también han sido sistemáticamente respaldados por investigadores de 40 países diferentes (Silverman, Choi y Peters, 2007; Silverman y Eals, 1992). Si bien las diferencias entre sexos pueden no deberse específicamente a una estructura que delimite diferencias en las capacidades y/o en la inteligencia, sino que al contexto y a la estimulación que se recibe en base al género y las expectativas en diversas culturas, a través de la educación (Gunderson, Ramirez, Levine y Beilock, 2012), está claro que existe la necesidad de tratar una variable como esta, al menos en este momento, considerando un método justo de comparación, teniendo en consideración la información entregada por los parámetros de la TRI.

Por otra parte, la Teoría de Respuesta al Ítem podría ser de mucha utilidad en contextos educativos, ya que permitiría obtener mayor información sobre la manera en la que los ítems se comportan tanto en las pruebas estandarizadas—las cuales representan la oportunidad de acceder a estudios superiores—, como en las pruebas realizadas por los académicos en colegios y/o universidades (Bernard-Brak, Lan y Yang, 2018).

Cabe destacar que, la emergencia de la ingeniería de la computación, la informática y su uso en los test psicológicos ha sido muy útil en lo que respecta a la evaluación de RM. Las mediciones de test visuoespaciales que incluyen traslación mental son el tipo de test denominado “test dinámicos”. El SODT—por sus siglas en inglés para Test Dinámico de Orientación Espacial—, creado por Santacreu (1999), han demostrado ser un buen instrumento para medir las Aptitudes

Espaciales en de todas sus versiones (Colom, Contreras, Botella, y Santacreu, 2002; Colom, Contreras, Shih, y Santacreu, 2003; Santacreu, 1999). En los ítems del test SODT, el objeto en movimiento es manipulado por un panel de control que está localizado en la parte superior de la pantalla, el cual es utilizado de forma imaginaria para rotar mentalmente la figura. Una flecha blanca en el centro del panel de control muestra la dirección actual del objeto en movimiento y una flecha negra en ambos lados de la flecha blanca son usados para cambiar la dirección hacia la izquierda o hacia la derecha. El objetivo de este test es llegar al punto de destino lo más rápido posible. El SOTD también ha utilizado refuerzos. Esto quiere decir que retroalimentación es proveída para cada respuesta, comprendiendo puede ser distinta dependiendo de si el ítem fue respondido correcta o incorrectamente (Contreras, Martínez-Molina, Manzanero y Peña, 2009; Contreras, Martínez-Molina y Santacreu, 2012). Considerando teoría psicométrica nueva, podría ser posible encontrar mejoras para futuros hallazgos en la investigación de RM, utilizando escalas estandarizadas no sólo con buenas propiedades de fiabilidad y validez, sino que también, actualizadas. Por otra parte, se podrían superar problemas que afectan a la precisión de la medición.

Estudios basados en la TRI podrían, además, contribuir a desarrollar nuevos modelos teóricos que podrían explicar mejor el funcionamiento de la RM y su relación con la inteligencia. Por ejemplo, un Test Adaptativo Computarizado—en inglés Computer Adaptative Test, CAT—toma ventaja de la TRI para proveer solo los ítems necesarios para determinar el nivel de rasgo en un sujeto, seleccionando adaptativamente el ítem siguiente que debería ser administrado basándose en la respuesta del ítem anterior. Como resultado, un Test Adaptativo Computarizado podría contar con un banco que incluyera cientos de ítems para cubrir el continuo de desempeño,

pero administrar de 5 a 8 para medir precisamente un individuo determinado (Kean y Reilly, 2014).

La importancia de utilizar herramientas y métodos que ayuden a medir componentes de inteligencia, es que existe evidencia de que estas habilidades pueden ser transferidas (Uttal, Meadow, Tipton, Hand, Alden, Warren y Newcombe, 2012), lo cual sería ventajoso para mejorar el rendimiento en otros procesos asociados, como las disciplinas STEM (Newcombe, 2010).

4. Entrenamiento en Rotación Mental: Antecedentes de su transferencia

La habilidad espacial es entendida como una de las habilidades cognitivas se manifiestan, tanto en varios aspectos de la vida cotidiana, como en variadas disciplinas (Shriki, Barkai y Patkin, 2017).

De acuerdo a lo señalado por algunos autores, como, por ejemplo, Nora Newcombe (2010)—quien se basó en una amplia revisión bibliográfica—, señalaba que las habilidades viso-espaciales eran fundamentales en las disciplinas STEM (Ciencia, Tecnología, Ingeniería y Matemáticas). Específicamente las habilidades espaciales predicen logros en las disciplinas STEM de acuerdo a diversos autores, así como la especialización en estas disciplinas en la educación superior (Uttal y Cohen, 2012; Wai, Lubinski, y Benbow, 2009; Uttal, Meadow, Tipton, Hand, Alden, Warren y Newcombe, 2012).

En la revisión bibliográfica de Uttal y colaboradores (2012), se establecieron una serie de criterios para determinar la efectividad, durabilidad y transferencia del entrenamiento. El primer criterio implicaba que los efectos del entrenamiento deben ser fiables y al menos de un tamaño moderado, observándose en los resultados, que los grupos entrenados mostraban un tamaño de efecto en promedio de ,62.

El segundo criterio, implicaba que el efecto debía ser duradero: Si bien, la mayor parte de los estudios no contaba con esta información, los resultados descritos por los autores indican que el entrenamiento puede dejar consecuencias mantenidas en el tiempo, debido a que la magnitud de los efectos de entrenamiento fue estadísticamente similar entre los post-test aplicados inmediatamente después del entrenamiento y luego de un lapso de tiempo después de terminado el entrenamiento. En relación a lo anterior, tienen claro que es posible que esos estudios que incluyeron evaluaciones después de un lapso de tiempo fueran diseñados para mejorar los

efectos o la durabilidad del entrenamiento, y es por esto que señalan que, se necesitan futuras investigaciones para especificar qué tipo de entrenamiento es más efectivo para dirigir efectos duraderos (Uttal, Meadow, Tipton, Hand, Alden, Warren y Newcombe, 2012).

El tercer criterio, fue que el entrenamiento debía ser transferible. En relación a lo anterior, en el meta-análisis de Uttal y Cohen (2012), los autores concluyen que las habilidades espaciales son una puerta de entrada para el campo STEM. Además, en las revisiones Uttal, Meadow, Tripton, Hand, Alden, Warren y Newcombe (2012), se probó que el entrenamiento tiene efectos fuertes de transferencia de una tarea a otra, y es por esto que los investigadores concluyen que, para poder incrementar las posibilidades de éxito de los aprendices de disciplinas STEM, deberían someterse a un entrenamiento dirigido (Shriki, Barkai y Patkin, 2017).

Cabe destacar—en relación al mejoramiento de las habilidades—, que en la investigación desarrollada por Uttal et al. (2012), se evidenció que quienes comenzaron a entrenar en niveles más bajos de desempeño, tuvieron un incremento mayor de mejora que los que empezaron en niveles más altos, por lo que se presume la posible existencia de un "techo" para el mejoramiento. Desde allí se puede entender que la maleabilidad de ciertas habilidades espaciales, al menos en lo que respecta a las pruebas en las que se evidenciaron, es limitada. Lo anterior lleva a preguntarse si es que la capacidad espacial mejora después de periodos largos de tiempo desde el entrenamiento y si el entrenamiento de una habilidad espacial en un contexto determinado afecta a otras habilidades espaciales, así como la edad en la que es más efectivo entrenar una habilidad espacial.

Los investigadores señalan que, en la mayoría de los estudios que se centraron en el impacto que provoca el entrenamiento a largo plazo, se encontró un efecto de retención y que demostraron que existe transferencia de las habilidades que se practicaron en otras, lo que

entrega información a favor de la mejora de otras habilidades asociadas con la capacidad espacial luego de la proporción suficiente de entrenamiento o de experiencia (Shriki, Barkai y Patkin, 2017).

Pregunta de investigación

¿Cuáles son los aportes de la aplicación de TRI en la validación de un instrumento computarizado de rotación mental aplicado a estudiantes universitarios?

Objetivo General:

Validar un test de Rotación Mental mediante el uso de la TRI y de TCT con una muestra de estudiantes universitarios.

Objetivos específicos:

1. Aportar teoría sobre el uso de la TRI en pruebas estandarizadas.
2. Desarrollar una versión computarizada, modificada y validada de La Tarea de orientación Dinámica Revisada (Conocida en por sus siglas en inglés como SODT-R). Esta nueva versión es llamada Ermental (ERM).
3. Obtener índices de dificultad y discriminación a partir de la aplicación del test en una muestra de estudiantes de nivel de estudios superiores.
4. Obtener índices de ajuste y de error para el Análisis Factorial Exploratorio (AFE).

Hipótesis:

H1: La fiabilidad para cada uno de los indicadores supera o es igual a los criterios considerados desde aceptables a buenos (≥,70 aceptable, adecuado y bueno ≥ ,90 a ,95) según Martínez-Molina, Arias y Garrido (2015).

H2: La nueva medida de RM contará con indicadores de ajuste adecuados para el análisis factorial exploratorio (EFA), según dificultad, mediante la técnica análisis paralelo (Horn, 1965; Garrido, Abad y Ponsoda, 2013; Velicer, Eaton y Fava, 2000) y según los intervalos recomendados por Schreiber, Stage, King, Nora y Barlow (2005): CFI ≥ ,95, TLI ≥ ,95, RMSEA < ,05.

H3: El índice de dificultad será mayor para los ítems que fueron catalogados como más difíciles que los que fueron considerados con menos dificultad, basándose en los puntos de refuerzo para cada ítem.

H4: Los ítems que fueron considerados como discriminantes (valores intermedios de refuerzo, es decir, 2, 3), muestran una mayor discriminación, que los que pueden ser respondidos con mayor facilidad o con demasiada dificultad (valor de refuerzo de 1 y 4).

Método

1) Diseño y tipo de estudio*:* Se realizará una validación del instrumento a través de Análisis Factorial Exploratorio (AFE) y Teoría de Respuesta al Ítem en una muestra no probabilística por conveniencia de estudiantes universitarios.

2) Instrumentos: Una versión alternativa para el Test Dinámico de Orientación Espacial revisada (Ver Figura 1)—en inglés The Spatial Orientation Dynamic Task-Revised, SODT-R (Santacreu y Rubio, 1998)—, llamada Ermental (ERM).

El SODT-R es un test computarizado en el que a los participantes se les pide dirigir un punto rojo y un punto azul hacia una dirección específica presionando las flechas de dirección en el teclado, basándose en el panel de control (Ver Figura 2).

El test ERM tiene originalmente tres intentos de práctica y dieciséis ítems que cuentan para el análisis de datos. El tiempo de cada intento es de 5 segundos más 2,5 segundos de presentación de estímulo. Los participantes son previamente advertidos en las instrucciones que deben responder lo más rápido posible. Los participantes no saben exactamente la cantidad de tiempo que tienen para responder los ítems.

El SODT-R reúne información a partir de múltiples variables como la frecuencia de respuesta, la latencia de la respuesta, la calidad de la primera presión y el tiempo invertido—el resultado es obtenido a través de la fórmula: tiempo del intento menos el tiempo de la latencia de respuesta—. La posición final para cada punto en movimiento es utilizada para computar la desviación promedio hacia el objetivo, que es la variable de logro para el SODT-R (Quiroga, Martínez-Molina, Lozano y Santacreu, 2011).

Los resultados del alfa de Cronbach para las mediciones de Rendimiento en el SODT-R fueron: 1) Frecuencia de Respuesta (FR)—total de presiones por intento—, α = ,87; 2) Calidad de la

primera presión (CPP)—definida como el número total de primeras presiones que reducen la desviación hacia el objetivo—, α = .71; y el Tiempo Invertido (TI) en cada intento—definido como el tiempo entre el primera y la última presión de las teclas—, α = ,63 (Quiroga, Martínez-Molina, Lozano, & Santacreu, 2011).

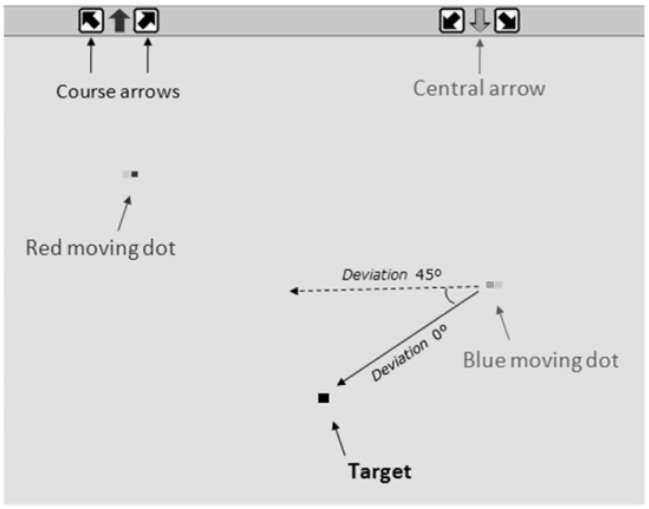

Figura 2. Ejemplo de un intento del Test Dinámico de Orientación Espacial-Revisado, SODT-R.

A diferencia de su versión previa, el test ERM ha añadido algunas modificaciones para los propósitos de este estudio, donde la flecha reemplaza la función del punto rojo, introduciendo una analogía en las instrucciones. Así, la tarea es descrita como una situación donde el sujeto debe imaginar que está conduciendo un automóvil, representado por la flecha. El objetivo, que en este caso es el cuadrado negro pequeño, y para poder llegar a él, el sujeto debe decidir la dirección basándose en el panel de control; presionando "z", si es que es necesario doblar hacia la izquierda, o "m", si es que es necesario doblar hacia la derecha (Ver Figura 3).

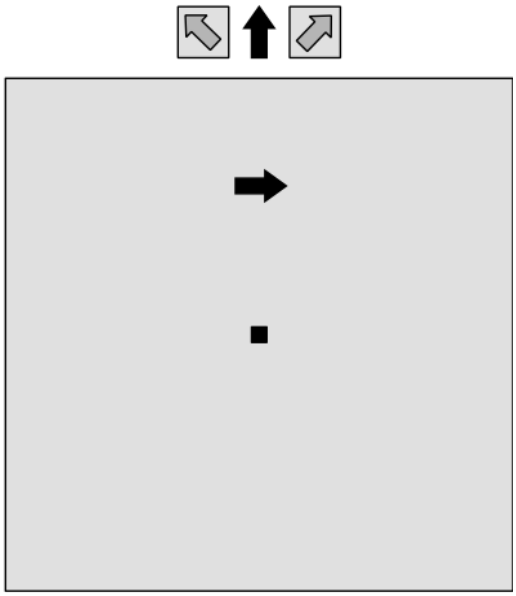

Figura 3. Ejemplo de un ensayo del test ERM.

El test Ermental contiene 16 estímulos organizados en dos diferentes formas del mismo test. Dependiendo de la dificultad de cada ítem es que se designa el refuerzo, que se refleja en puntajes del 1 al 4, si es que la respuesta fue correcta. Por otro lado, si la respuesta fue incorrecta, se presentará una diapositiva gris sin ningún refuerzo (Ver figura 4). Esto también se explica en las instrucciones.

En lo que respecta a la designación de los puntajes de refuerzo para las dificultades, Shepard y Metzler (1971), encontraron una relación lineal positiva entre los grados de discrepancia en la orientación de la figura en un ítem y el tiempo de resolución de la tarea, conocida como discrepancia angular. Estos estudios, se relacionan paralelamente a la conservación de las propiedades físicas debido a que esta relación se mantiene al momento de demostrar el paralelismo entre el tiempo que toma realizar en la vida real el hacer un determinado juicio y el tiempo que se tarda al hacer un juicio visuoespacial en la memoria; resultados que han sido replicados en niños (Quiroga, Martínez-Molina, Lozano y Santacreu, 2011).

Figura 4. Ejemplo de refuerzo en el test ERM.

2) Descripción de la muestra: 213 estudiantes de educación superior, donde n=115 corresponde al número de mujeres, lo que da cuenta de un 54% de la muestra total y n=98 corresponde al número de hombres, lo que da cuenta del 46% de la muestra total. La edad promedio de los participantes de este estudio bordea los 24 años (M=23,84).

2.1) Justificación: Para determinar el tamaño muestral de este estudio, se requiere tener en consideración los análisis a realizar. En este caso, se requieren estudios de Análisis Factorial Exploratorio (AFE) y Teoría de Respuesta al Ítem (TRI), por lo que primero se debe establecer el número de ítems que se analizarán y las características de los datos, como lo son la cantidad de variables por factor y las comunalidades (Velicer y Fava, 1998).

Según Fabrigar, Wegener, MacCallum y Strahan (1999), en situaciones de comunalidad moderada, el tamaño muestral no debería ser menor a 200. Por otro lado, la relación entre observaciones e ítems no debería ser menor de 5:1 ó 10:1(Costello y Osborne, 2005; Fabrigar

et al., 1999; Floyd y Widaman, 1995). Se necesita calcular la matriz de correlaciones tetracóricas para realizar AFE, debido a que los ítems utilizados para la medición son dicotómicos. Estas correlaciones tienen mayor error muestral que las de Pearson, y es por esto que se sugiere utilizar el ratio más conservador de 10:1. En el caso particular de este test, el cual cuenta con 16 ítems resultantes de la distribución de los estímulos en dificultades, se necesitaría un tamaño de la muestra de 10*16, es decir, de 160 alumnos voluntarios de educación superior. Debido a la naturaleza de la fuente en la que se recopilarán los datos y, asumiendo que, por distintos motivos, habrá casos excluidos, se tiene en consideración que las observaciones para este estudio deberían ser de, al menos el doble, es decir, 320 sujetos que respondan el test.

3) Descripción de variables:

Las variables utilizadas en este estudio, tanto para el análisis descriptivo para los posteriores análisis contemplados en el plan se detallan a continuación en la siguiente tabla (Ver tabla 1)

Tabla 1: Descripción de las variables y sus respectivos valores en la base de datos.

Variable	Valores	Descripción
Usuario	1 a 213	Indica el número del participante
Sexo	1; 0	Indica si el participante se identificó como hombre o mujer, donde 1 se designó para las mujeres de la muestra, y 0 para las mujeres de la muestra.
Edad	N	Indica la edad del participante en años.
Aciertos	1 a N	Sumatoria de aciertos de las observaciones.
Errores	0 a N	Sumatoria de errores de las observaciones.

Puntos	1 a 4	Indica el puntaje asignado como refuerzo para cada uno de los estímulos. Este puntaje refleja el nivel de dificultad del ítem, donde a mayor dificultad, mayor puntaje obtenido.
I1 a I16	1;0	Indica si el sujeto acierta (1) o erra (0) un ítem.
Puntos	1 a 4	Indica el puntaje asignado como refuerzo para cada uno de los estímulos. Este puntaje refleja el nivel de dificultad del ítem, donde a mayor dificultad, mayor puntaje obtenido.

N = Número correspondiente al valor máximo para la variable descrita en la base de datos.

4) Procedimiento:

Se convoca a estudiantes de educación superior, ya sea de universidades o de institutos técnico-profesionales, a completar el test ERM a través de redes sociales—Facebook, Instagram, Whatsapp, correo electrónico, etcétera—, donde se inicia la difusión del link que contiene la prueba, la cual se encuentra disponible en una plataforma online en la dirección www.ermental.cl. Los participantes acceden a colaborar en el da través de la aceptación de un contrato ético en el que los propósitos de este estudio son brevemente explicados.

A) Análisis descriptivo inicial: El objetivo es poder eliminar los casos no válidos. La muestra inicialmente contaba con 474 observaciones en la base de datos de ERM, de los cuales, 213 casos fueron válidos, es decir, se pudo dar cuenta de todas las variables implicadas en el estudio para 213 estudiantes de estudio superior.

B) Análisis de fiabilidad: Se obtiene la fiabilidad de los ítems subdivididos de acuerdo a su dificultad con el objetivo de obtener información acerca de la precisión con la que mide el conjunto de ítems designado según modelos propuestos. En el caso de la TRI, esto se obtiene a través de las curvas de información.

C) Análisis de descriptivos por dificultad: Se obtendrán índices de correlaciones bivariadas entre los ítems.

D) Análisis paralelo: Se procederá a realizar un Análisis Factorial Exploratorio (AFE) a través del programa Jamovi, para obtener los índices de ajuste del modelo en una primera instancia; posteriormente, se obtendrán índices de ajuste a través del programa Factor Analysis. La extracción de la matriz factorial.

F) Teoría de Respuesta al Ítem (TRI): A través del software R Commander, para obtener comparaciones respecto a la información extra que este análisis podría entregar para la comprensión del comportamiento de las variables en términos de índices de dificultad, discriminación y adivinación.

4.1) Obtención de indicadores esperados:

En cuanto a la fiabilidad: Para que el modelo sea aceptado con un buen índice de fiabilidad, debe igualar o superar a los indicadores considerados desde aceptables a buenos según (≥,70 aceptable, adecuado y bueno ≥ ,90 a ,95) según Martínez-Molina, Arias y Garrido (2015).

En cuanto a la validez: Supera adecuadamente los criterios del análisis factorial exploratorio (AFE) mediante la técnica de determinación análisis paralelo (Horn, 1965; Garrido, Abad y Ponsoda, 2013; Velicer, Eaton y Fava, 2000) y los límites recomendados por Schreiber, Stage, King, Nora y Barlow en 2005 (CFI ≥ 0.95, TLI ≥ 0.95, RMSEA < 0.05).

En lo que respecta a la Teoría de Respuesta al Ítem (TRI), se espera obtener índices de discriminación y dificultad, tal y como se describe en las hipótesis.

5) Consideraciones éticas: Para participar en el estudio, los participantes deben leer y aceptar un consentimiento informado que aparece en la página del estudio al momento de ingresar su participación, el cual, además, explica brevemente los propósitos del estudio y objetivos de éste.

Resultados

A) Estadísticos descriptivos:

De acuerdo a lo evidenciado por la Tabla 2, en relación a los ítems que fueron clasificados dentro de la misma dificultad, basándose en los supuestos acerca de la discrepancia angular, existirían relaciones positivas entre los ítems 1 y 7, siendo estas significativas (p<,01), entre los ítems 11 y 1 (p<,05), y entre los ítems 11 y 7 (p<,01), lo que da cuenta de una relación que sustentaría la agrupación de estos ítems en una dimensión.

En relación a la media, se observa que, en general, los ítems tienen una alta tasa de acierto, siendo en promedio M=,784, lo que indicaría que es la mayor tasa de acierto en relación al resto de dificultades. En relación a la moda, por cada ítem se observa que en todos los ítems es más común el acertar que el errar.

En cuanto a la varianza, se observa, además, que, en relación al promedio, es la variabilidad de respuesta más baja.

Tabla 2: Tabla de correlaciones y estadísticos descriptivos de los ítems pertenecientes a la dificultad 1.

D1	i1	i7	i11	i13	x̄
i1	1,00				
i7	,274**	1,00			
i11	,158*	,264**	1,00		
i13	,144	-,044	-,009	1,00	
N	213	213	213	213	
Media	,840	,826	,817	,657	,785
DE	,367	,380	,399	,476	
Varianza	,135	,144	,150	,226	0,163
Asimetría	-1,87	1,02	,731	-1,57	
Curtosis	1,52	1,02	,731	-1,57	
Moda	1	1	1	1	

Por otra parte, según lo evidenciado en la Tabla 3, para la segunda dimensión, la cual también se basa en los supuestos de discrepancia angular, se observan correlaciones positivas entre los ítems, siendo estas significativas (P<0,05) entre los ítems 8 y 6.

En lo que respecta a la tasa de respuestas, se observa que, en promedio, esta es peor que para la dificultad 1, pero mejor que para las dificultades 3 y 4.

En relación a la media, se observa que la tasa de respuestas es la segunda mejor en comparación al resto de dificultades. Este comportamiento se replica para la varianza, que demuestra ser la segunda peor variabilidad.

En relación a la moda, se observa que en todos los ítems es más común el acertar que el errar.

Tabla 3: Tabla de correlaciones y estadísticos descriptivos de los ítems pertenecientes a la dificultad 2

D2	i2	i4	i6	i8	$\bar{x}$
i2	1,00				
i4	,123	1,00			
i6	,045	,083	1,00		
i8	,099	,126	,175*	1,00	
N	213	213	213	213	
Media	,695	,784	,723	,596	,700
DE	,462	,412	,449	,492	
Varianza	,213	,170	,201	,242	0,206
Asimetría	-,852	-1,39	-1,00	-1,86	
Curtosis	-1,29	-0,067	-1,00	-1,86	
Moda	1	1	1	1	

Según lo evidenciado por la Tabla 4, se observan relaciones positivas para los ítems designados para la dificultad 3, siendo estas significativas entre los ítems 3 y 5 (P<0,01); 3 y 9

(P<0,05); 3 y 15 (P<0,01); 9 y 5 (P<0,01); 15 y 5 (P<0,01) y por último 15 y 9 (P<0,01), lo que da cuenta de una relación que sustentaría la agrupación de estos ítems en una dimensión.

En lo que respecta a la tasa de respuestas, se observa que, en promedio, es la peor en relación al resto de dificultades (M=,456). Además, según los resultados obtenidos, se obtuvieron más errores que aciertos para estos ítems.

En relación a la moda, se observa que para todos los ítems es más común el errar que el acertar.

En cuanto a la varianza, se observa que es mayor que D1 y D2, indicando una mayor variabilidad de respuestas.

Tabla 4: Tabla de correlaciones y estadísticos descriptivos de los ítems pertenecientes a la dificultad 3.

D3	i3	i5	i9	i15	x̄
i3	1,00				
i5	,203**	1,00			
i9	,167*	,301**	1,00		
i15	,298**	,221**	,378**	1,00	
N	213	213	213	213	
Media	,408	,480	,474	,465	,456
DE	,493	,493	,501	,500	
Varianza	,243	,243	,251	,226	,240
Asimetría	,375	,375	-2,01	-2,00	
Curtosis	-1,88	-1,88	-2,01	-2,00	
Moda	0	0	0	0	

De acuerdo a lo evidenciado por la Tabla 5, se observa que existen correlaciones positivas entre los ítems, siendo ésta una relación significativa entre los ítems 2 y 4 (P<0,05). Teniendo en consideración que son sólo 4 ítems, estos análisis no proporcionarían información

suficiente para subdividir estos ítems en un solo factor además de lo sugerido por la teoría respecto a la discrepancia angular.

En relación a la tasa de respuestas, se observa que, si bien no es tan alta como para las dificultades 1 y 2, es más alta que para la dificultad 3. Es por esto que se realiza una prueba T para muestras relacionadas, tal y como se observa en la Tabla 6, donde el resultado es estadísticamente significativo (P=,03), dando cuenta de una media más alta para la dificultad 4.

En relación a la moda, se observa que, específicamente en el ítem 12, que es el ítem con mayor tasa de respuestas, fue más probable acertar que errar. En cambio, para los otros ítems fue más probable el errar que acertar.

Tabla 5: Tabla de correlaciones y estadísticos descriptivos de los ítems pertenecientes a la dificultad 4.

D4	i10	i12	i14	i16	$\bar{x}$
i10	1,00				
i12	,044	1,00			
i14	,051	,146*	1,00		
i16	,108	,014	,117	1,00	
N	213	213	213	213	
Media	,474	,512	,498	,498	,496
DE	,501	,501	,501	,501	
Varianza	,251	,251	,251	,251	,251
Asimetría	-2,01	-2,02	-2,02	-2,02	
Curtosis	-2,01	-2,02	-2,02	-2,02	
Moda	0	1	0	0	

Tabla 6: Prueba T para muestras relacionadas entre D3 y D4.

	Media	Desviación estándar	t	Sig (bilateral)
Par 1 (D3-D4)	-,225	1,509	-2,179	,030

B) Análisis Factorial Exploratorio, AFE:

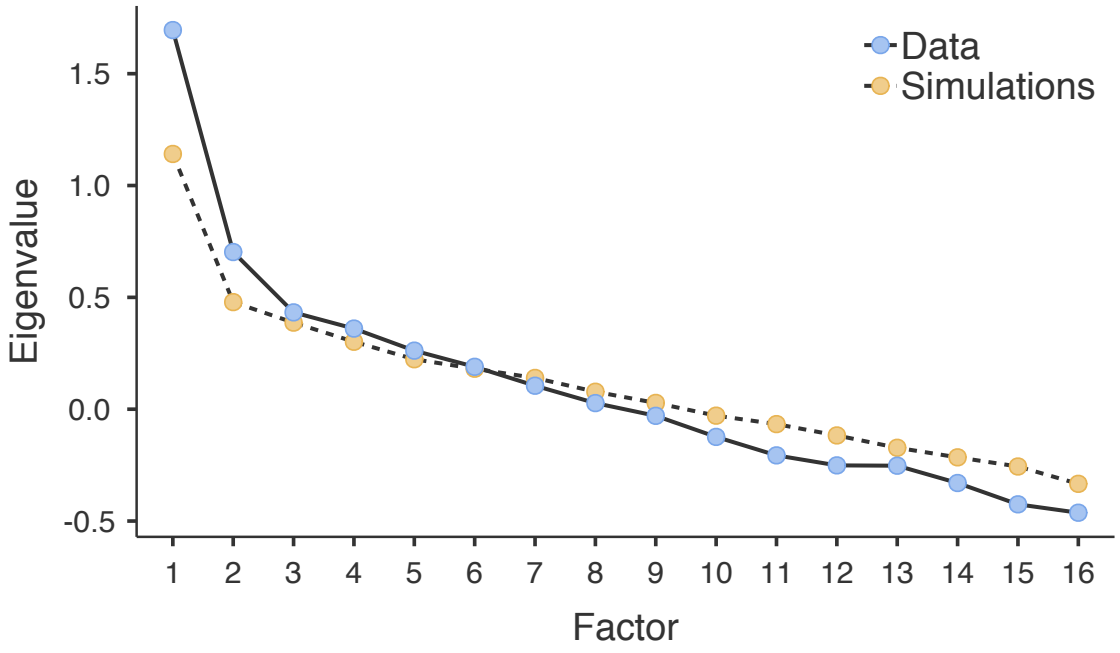

Figura 5. Gráfico de sedimentación resultante del análisis paralelo.

En lo que respecta a los factores que sugiere retener el análisis, Factor arrojó 5, mientras que el gráfico de sedimentación obtenido de Jamovi (Ver Figura 5), muestra que al menos 4 factores deberían ser retenidos.

Tabla 7: Alfa de Cronbach según dificultad.

	Alfa si se elimina el elemento	α de Cronbach
Escala total		,565
Dificultad 1		,346
i1	,151	
i7	,229	
I11	,270	
I14	,475	
Dificultad 2		,328
I2	,309	
I4	,265	
I6	,279	
I8	,212	
Dificultad 3		,587
I3	,563	
I5	,540	
I9	,488	
I15	,464	
Dificultad 4		,258
i10	,234	
i12	,234	
i14	,149	
i16	,207	

En lo que respecta a la fiabilidad de la escala (α=,565), se observa que, al igual que en todas las sub-escalas propuestas a partir de las dificultades basadas en la discrepancia angular, se evidencia que los indicadores no cumplen con los requisitos mínimos para ser considerados aceptables según Martínez-Molina, Arias y Garrido (2015).

Por otra parte, se observa que el Alfa de la escala general no puede ser mejorada, aunque se quiten ítems.

Tabla 8: Matriz de cargas por factor.

Matriz de cargas rotadas				
Variable	F 1	F 2	F 3	F 4
i1			,397	,583
i2	-,302	,558		,482
i3		,368		
i4			,600	
i5	,709			
i6	,688			
i7				,752
i8				
i9	,794			
i10				
i11		-,728		
i12				
i13			,499	
i14			,583	-,309
i15	,353	,579		
i16				
Criterio			Índice	Valor P
RMSR			,0724	
RMSEA			,051	
Kelley			,0687	
CFI			,967	
GFI			,954	
Bartlett				
X^2	df			
361	120			< .001

A partir de los resultados obtenidos en la matriz de cargas por factor, —con rotación oblimin—, se obtienen cargas adecuadas para el primer factor, que implicaría los ítems 5, 6 y 9, de acuerdo a la regla 30-40-20 (Martínez-Molina, Arias y Garrido, 2015). Para el segundo factor, se obtienen cargas adecuadas en los ítems 2 y 1. Para el tercer factor, se obtienen cargas adecuadas

en los ítems 4, 13 y 14. Por último, para el cuarto factor, se obtienen cargas adecuadas para el ítem 7, ya que se evidencian cargas cruzadas para el ítem 1 entre el tercer y el cuarto factor.

En lo que respecta a los índices de ajuste, de acuerdo a CFI y GFI, indicaría que el ajuste es adecuado, y el RMRSEA tiene un buen índice de ajuste. En relación al RMSR, se observa que este es mayor que el estadístico de Kelley, por lo que no entrega antecedentes para un ajuste adecuado.

El estadístico de Barlett es significativo, por lo que se desprende que la muestra cuenta con un N adecuado.

C) Teoría de Respuesta al ítem

Tabla 9: Parámetros de dificultad (b) y discriminación (a), según dificultad con sus respectivos ítems.

Ítem	Dificultad (β)	Discriminación (α)	x̄ discriminación
1	-2,3512	,7899	,4109
7	-2,6414	,6379	
11	-24,1707	,0619	
13	-4,2472	,1541	
Ítem	Dificultad	Discriminación	
2	-4,0514	,2051	,5405
4	-2,3016	,6024	
6	-,8947	,4952	
8	-,5257	,8593	
Ítem	Dificultad	Discriminación	
3	,4882	,8852	1,3892
5	,3450	1,5184	
9	,0900	1,7085	
15	,1345	1,4450	
Ítem	Dificultad	Discriminación	
10	,2947	,3618	,2945
12	-,1385	,3488	
14	,0226	,4342	
16	,2818	,0333	

A partir de la información proporcionada por la Tabla 9, se evidencia que, tal y como se hipotetizaba, los ítems más discriminantes pertenecen, a los niveles intermedios, según las comparaciones de media del índice de discriminación. Se observa, además, que el ítem 16 cuenta con el nivel más bajo de discriminación (α=,0333), junto al ítem 11 (α=,0619). En el caso del ítem 11, queda evidenciado que se necesita menos rasgo (β) para acertar los ítems —lo que en este caso se traduciría en capacidad de rotación mental—. Esto se refleja, además, en sus bajos índices de dificultad βi_{11}=-24,1707. Respecto a esto mismo, en la curva característica del ítem (ver Figura 6), se observa que la probabilidad de acertar del ítem 11, particularmente es muy alta incluso desde niveles bajos de rasgo, en comparación al resto de curvas de los otros ítems. Ambos ítems pertenecen a las dificultades 4 y 1, respectivamente, lo que indica que estos ítems tienen poca capacidad discriminativa estando en el primer nivel de dificultad y en el cuarto nivel de dificultad.

En relación a lo anterior, el ítem 16 presenta una dificultad alta en relación al resto de los ítems (βi_{16}=,2818), lo cual corresponde al nivel en el que fue clasificado a priori, siendo éste el tercer ítem más difícil después del ítem 10 (βi_{10}=,2947), el cual, sería el segundo más difícil—perteneciente, también, a la dificultad máxima— y el ítem 3 (βi_3=,4882)—el cual pertenece a la tercera dificultad—, ya que, como se observa, a pesar de que se aumente el nivel del rasgo, la probabilidad de acertar sigue sin alcanzar niveles tan altos como otros ítems, que incluso logran probabilidades más altas que otros con niveles bajos de rasgo.

En lo que respecta a la curva de información del ítem, se evidencia que el ítem 11 aporta muy poca información en relación a la capacidad de rotación mental respecto grado de habilidad con la que cuentan los sujetos, mientras que en el caso del ítem 16 sucede lo mismo. Se evidencia

que esto ocurre con ítems que tienen poca capacidad de discriminar a los participantes o a los sujetos.

Por otra parte, se evidencia que el ítem con mayor discriminación es el ítem 9 (ver Tabla 9), lo que se condice con la curva de información del mismo ítem (ver Figura 13), la cual es muy parecida a una curva normal, donde la mayor parte de la información estaría en la media, lo que quiere decir que es un ítem apto para medir el rasgo de rotación mental en la mayor parte de los participantes. Esto además coincide con la hipótesis de dificultades intermedias, ya que son éstas las que, teóricamente son las con mayor capacidad discriminativa.

En relación a la discriminación, se observa que, la dificultad que mayormente podría clasificar a los sujetos es la tercera, ya que cuenta con los niveles más altos de discriminación, lo que además coincide con niveles intermedios de dificultad, que van desde un rango de β=-24,1707 a β=,4882, siendo el valor mínimo un valor extremo, ya que el segundo grado de dificultad más bajo es β=-4,2472 y pertenece al ítem 13.

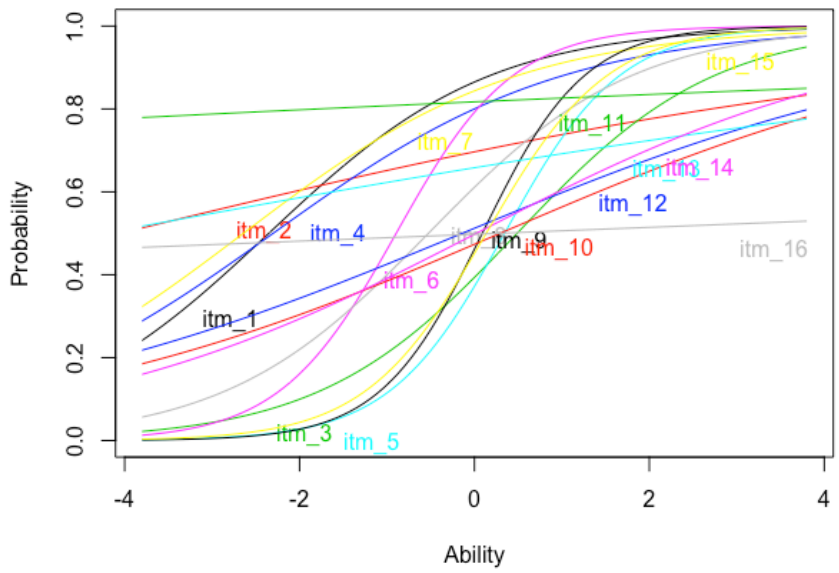

Figura 6. Curva característica de los 16 ítems de ERM.

De acuerdo a lo evidenciado por las curvas características de los ítems pertenecientes a la primera dificultad (ver Figura 7), se observa en el gráfico que el grado de rasgo que se necesita para acertar con un 50% de probabilidad, es relativamente bajo, lo que da cuenta de niveles bajos de dificultad. En el caso particular del ítem 11, se observa que, a pesar de contar con niveles muy bajos de rasgo, la posibilidad de acertar es casi de un 80%.

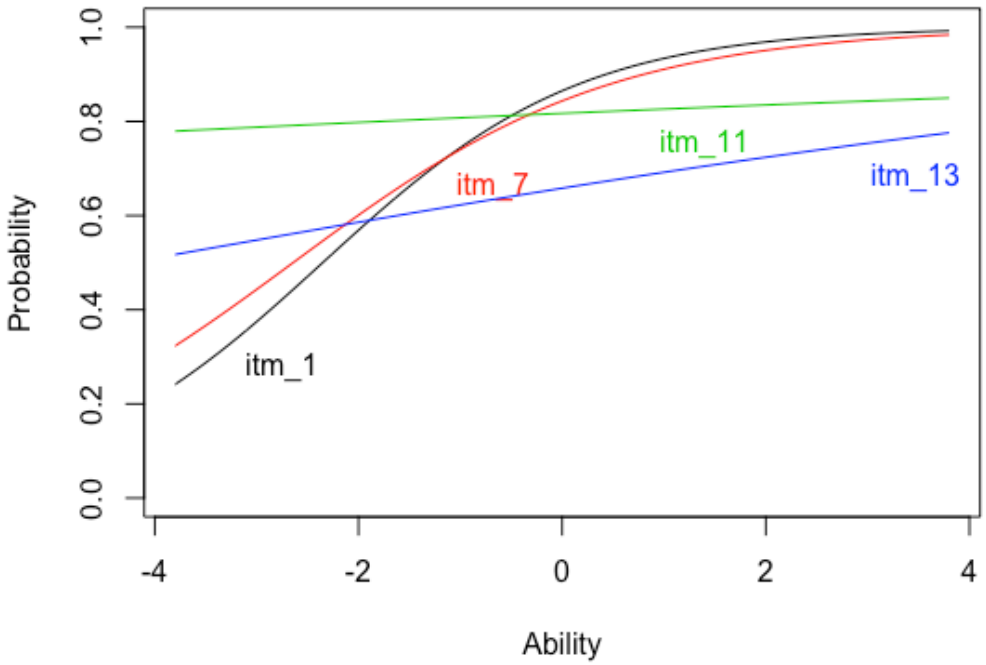

Figura 7. Curva característica de los ítems pertenecientes a la dificultad 1.

De acuerdo a lo evidenciado por las curvas características de los ítems pertenecientes a la segunda dificultad (ver Figura 8), se observa que los ítems tienen propiedades muy distintas entre sí, ya que, por ejemplo, para el ítem 8, se evidencia que se necesita una capacidad de rotación mental aproximadamente en la media para obtener un 50% de probabilidad de acertar, mientras que en el caso del ítem 2, se observa que se necesita una capacidad de rotación mental de aproximadamente dos puntos menos de desviación estándar para obtener un 50% de

probabilidades de acertar.

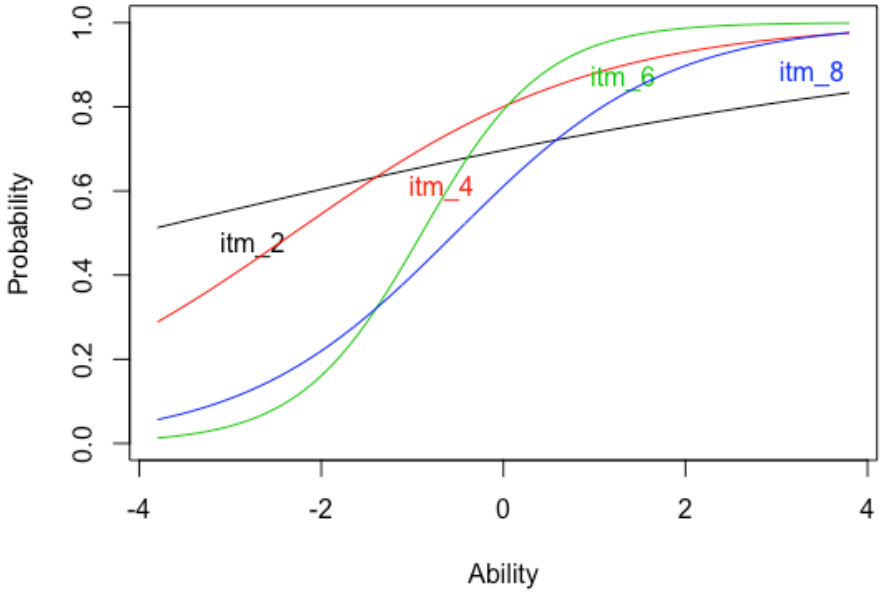

Figura 8. Curva característica de los ítems pertenecientes a la dificultad 2.

De acuerdo a lo evidenciado por la curva característica de los ítems pertenecientes a la tercera dificultad (ver Figura 9), se observa que las propiedades de los ítems en términos de discriminación y dificultad se reflejan de manera similar en la forma de "S" de las curvas. Esto refiere que los ítems van aumentando su probabilidad de respuesta en la medida en que la capacidad de rotación mental es mayor. Se observa, particularmente, que para el ítem 15, se necesita cercano a una desviación estándar de habilidad para obtener un 50% de probabilidad de acertar.

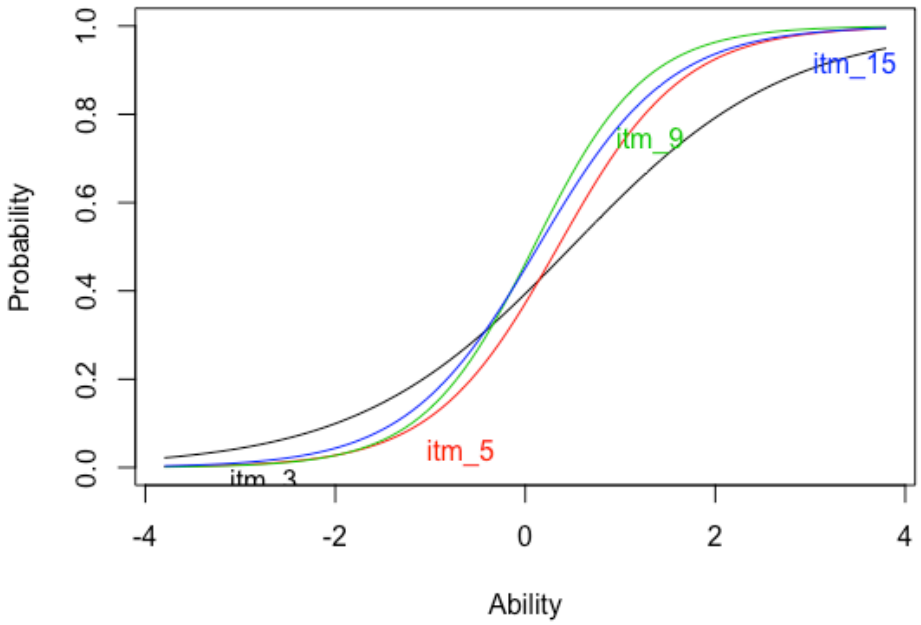

Figura 9. Curva característica de los ítems pertenecientes a la dificultad 3.

De acuerdo a lo evidenciado por las curvas características de los ítems pertenecientes a la dificultad 4 (ver Figura 10), se observa que son más similares a una línea recta, lo que refleja que a medida que aumenta la capacidad de rotación mental, aumenta rápidamente la probabilidad de acertar, a excepción del ítem 16, que se muestra más difícil que el resto, ya que a pesar de que la capacidad de rotación mental aumente, la probabilidad de acertar no aumenta mucho.

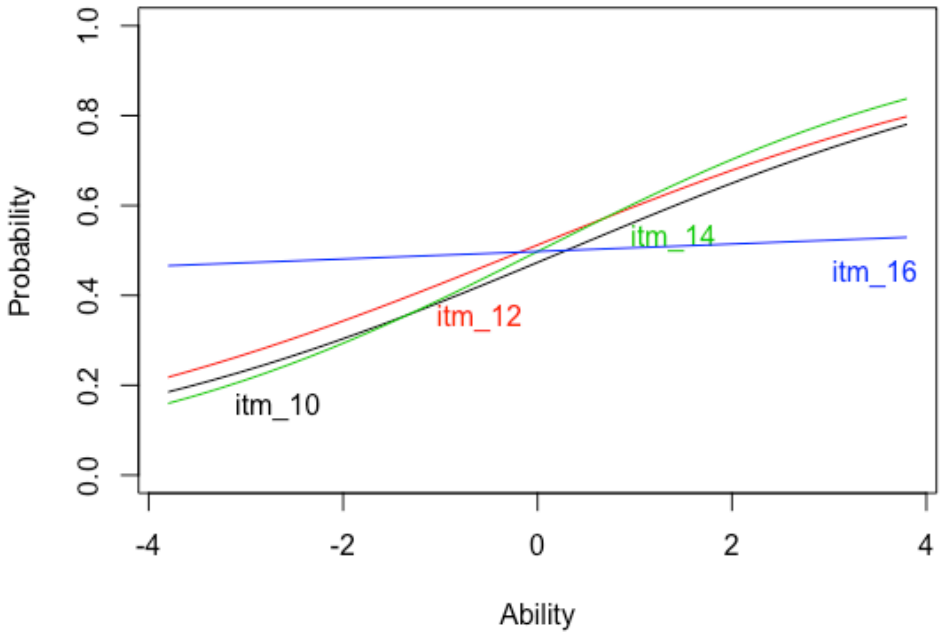

Figura 10. Curva característica de los ítems pertenecientes a la dificultad 4.

En lo que respecta a las curvas de información para la dificultad 1 (ver Figura 11), se observa que los ítems 11 y 13 entregan poca información sobre la capacidad de rotación mental de los sujetos que responden el test, mientras que la mayor cantidad de información se concentra en los grados más bajos de habilidad, lo que da cuenta de que la utilidad de estos ítems está en entregar información sobre los grados más bajos de rasgo. Esto coincide, además, con la asignación de estos ítems en el nivel más bajo de dificultad.

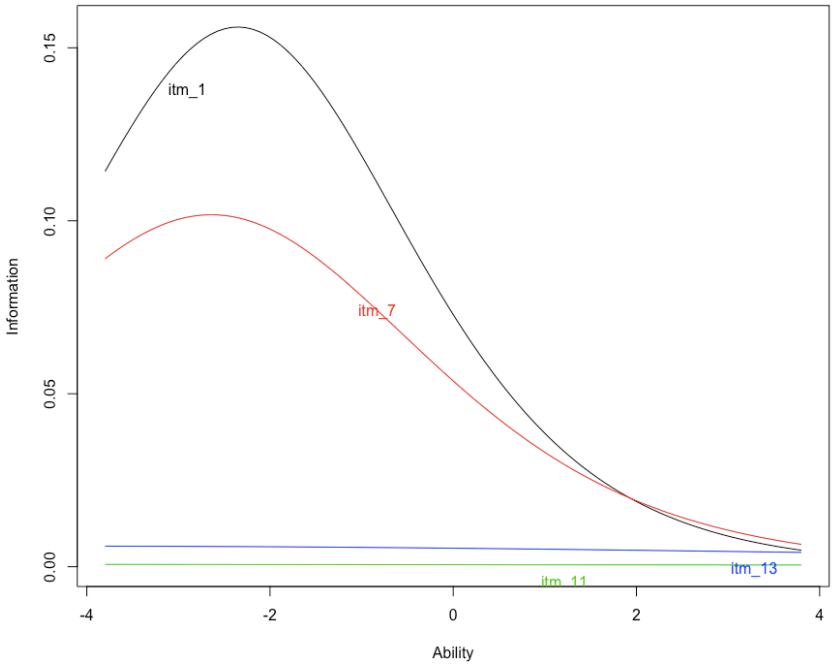

Figura 11. Curva de información de los ítems pertenecientes a la dificultad 1.

En lo que respecta a las curvas de información para la dificultad 2 (ver Figura 12), se observa que se moviliza la concentración de información hacia la derecha, dando cuenta de que, al menos los ítems 6 y 8 serían más pertinentes para la obtención de información en la segunda dificultad. Por otra parte, el ítem 2 entrega muy poca información respecto al rasgo observado y la información concentrada para el ítem 4 se encuentra más posicionada a la izquierda.

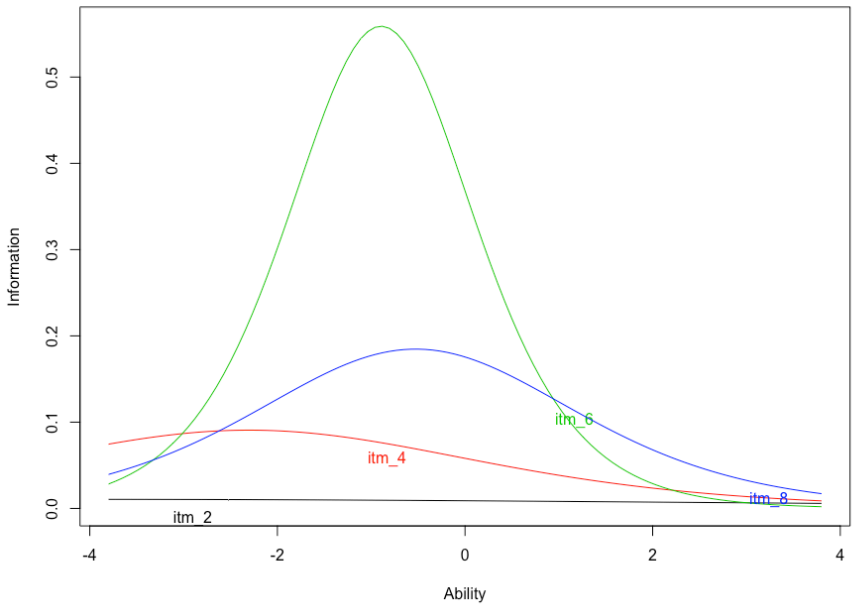

Figura 12. Curva de información de los ítems pertenecientes a la dificultad 2.

De acuerdo a las curvas de información para la dificultad 3 (ver Figura 13), se observa que, en general, la información se concentra en torno a la media, lo que indica particularmente que estos ítems entregan más información sobre la mayor parte de los sujetos, siendo el ítem 3 el que entrega la menor cantidad de información en comparación al resto. Por otra parte, se observa que el ítem 9 entrega la mayor cantidad de información respecto a los sujetos que responden en torno a la media.

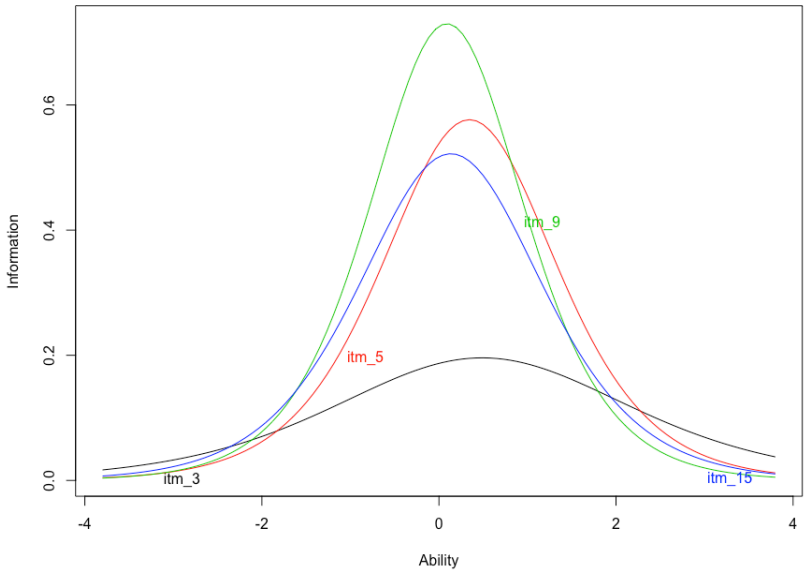

Figura 13. Curva de información de los ítems pertenecientes a la dificultad 3.

Según lo evidenciado por las curvas de información de los ítems de la cuarta dificultad (ver Figura 14), se observa que, el ítem 14 es el ítem que mayor información entrega, mientras que el 12 y el 10 se encuentran en un nivel intermedio de entrega de información en comparación a los ítems 14 y 16. En el caso particular del ítem 16, se observa que entrega muy poca información, lo que coincide con que es el ítem menos discriminante.

En general, esta dificultad es la que presenta en promedio la menor discriminación ($\alpha_{\bar{x}}$=,2945), por lo que los niveles de información son relativamente altos en casi todos los ítems, a excepción del ítem 16.

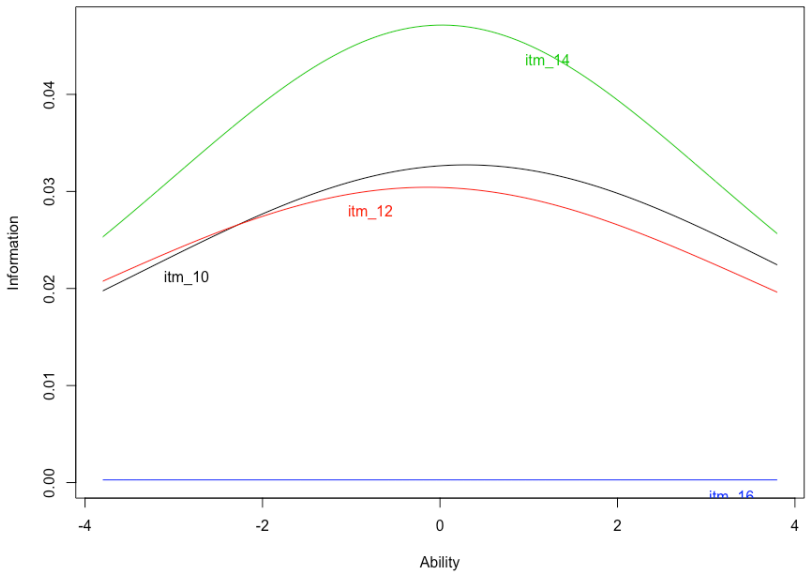

Figura 14. Curva de información de los ítems pertenecientes a la dificultad 4.

Discusión

Según lo hipotetizado para la fiabilidad—que además, es uno indicadores de la TCT—, se esperaba que fuera sobre ,70 (Martínez-Molina, Arias y Garrido, 2015), lo cual no fue logrado en general, sin embargo, la manera en la que se comprende precisión desde la TRI, se entiende a través de la curva de información del ítem, debido a que, entre los supuestos se asume que la precisión puede variar en la distribución o el continuo de habilidad, que en este caso se entendía como la capacidad de rotación mental para una dificultad determinada. Desde allí se explica que no haya habido una precisión adecuada para la escala general (Hidalgo-Montesinos y French, 2016), sin embargo, lo anterior no quiere decir que no existan ítems que cuenten con precisión adecuada para medir cierto espectro del rasgo, como en el caso de los ítems 1 y 7 de dificultad uno, donde la curva de información muestra que la precisión de los ítems se concentra hacia la izquierda (ver Figura 10), dando cuenta de una mayor adecuación para medir niveles bajos de rasgo, y no así para otros posibles niveles lo que indica que no son precisos para la mayoría de los participantes. Por otra parte, y, respecto a lo anterior, las curvas de información para la dificultad 3, centran la información hacia los niveles medios de habilidad, dando cuenta de niveles de precisión en el continuo del rasgo donde se desenvuelve la mayor parte de los sujetos.

En lo que respecta al análisis factorial exploratorio, AFE, se obtienen índices de ajuste adecuados (Schreiber, Stage, King, Nora y Barlow, 2005), CFI=,967; GFI=,954 y en cuanto al error, el indicador es excelente, con un RMSEA=,051. Por otra parte, el criterio de Kelley no es

menor al RMSR, por lo que no indica un buen ajuste. En cuanto a la adecuación de la muestra, de acuerdo al índice de Barlett, se observa que ésta es adecuada (p<,001)

A partir de los resultados obtenidos en la matriz de cargas por factor (ver Tabla 7), se obtienen cargas adecuadas para ciertos factores, sin embargo, esta distribución no coincide con la disposición teórica que se tenía de los ítems en base a la dificultad propuesta por la discrepancia angular (Shepard y Metzler, 1971). Lo anterior, puede deberse a que no todos los ítems funcionaban bien para las dificultades establecidas a priori de acuerdo a los resultados, ya que éstos mostraron que, incluso, algunos de ellos no entregaron mayor información acerca del desempeño de la capacidad de rotación mental. Sumado a lo anterior, eliminando ciertos ítems que no funcionaban correctamente, no se cumplía el criterio mínimo de ítems por factor de 3:1 (Costello y Osborne, 2005). Además, la manera de resolver los ítems respecto a los grados que se deben rotar y la dirección de la rotación, es diferente para cada uno de ellos, por lo que, posiblemente el añadir alguna repetición de los ítems a lo largo de la escala podría mejorar la precisión de ésta.

De acuerdo a los resultados obtenidos a partir de la TRI, se observa que, tal y como se hipotetizó en un comienzo, las dificultades de los ítems que fueron catalogados en una primera instancia como más fáciles (ver Tabla 8) fueron mucho menores, obteniéndose dificultades negativas para todos los ítems de los dos primeros niveles de dificultad. Lo anterior, se refuerza en por las curvas características de los ítems, de las cuales se desprende que, con un nivel más bajo de rasgo para las primeras dificultades, la probabilidad de acertar es mayor que en el caso de las dificultades posteriores (ver Figuras 6, 7, 8 y 9).

En base a lo anterior, además, se obtiene en los resultados que, la discriminación promedio es menor en los niveles intermedios de dificultad, es decir, en los niveles 2 y 3—con índices de discriminación promedio de ,5405 y 1,3892, respectivamente—, que en el nivel más fácil y en el nivel más difícil de dificultad (ver Tabla 8).

A partir de los resultados obtenidos, se podría entonces pensar en propuestas para utilizar los ítems de la escala, debido a que no todos cuentan con propiedades adecuadas para discriminar correctamente a los sujetos. Respecto a lo anterior, algunos autores como Hidalgo-Montesinos y French (2016), afirman que para que un ítem funcione bien, debe ser capaz de discriminar a los sujetos en todos los niveles del rasgo, lo que se traduciría en un valor relativamente alto para el parámetro de discriminación (α), y que, además, deben reunir información suficiente. Estas fuentes de información se obtendrían a través del cumplimiento de un umbral mínimo de α para que el ítem funcione de manera aceptable, además de una revisión de expertos para los ítems que no cumplan el criterio, la evaluación de las curvas características de información—las cuales no deberían ser planas—, entre otras consideraciones que puedan influir en la elegibilidad del ítem, como, por ejemplo, la teoría a la base.

Siguiendo los criterios de evaluación de Baker (2001), si el objetivo es poder discriminar a los sujetos de acuerdo a su capacidad de rotación mental, se deberían conservar los ítems que cumplan con el criterio α>,65, donde α>1,34 sería un ítem con un funcionamiento elevado y α>1,69 indicaría un ítem con un funcionamiento muy elevado. En este caso, los ítems pertenecientes a la dificultad 3, cumplen en su totalidad con el requerimiento mínimo de discriminación propuesto por este autor, con un α=8852 para el ítem 3; α=1,5184 para el ítem 5; α=1,7085 para el ítem 9; y, finalmente α=1,4450 para el ítem 15. Además, las curvas de

información indican que se concentra en un nivel medio del rasgo, es decir, en torno a las respuestas promedio de los sujetos, donde se concentra el desempeño de la mayoría de éstos.

En caso de que se quisiera medir mínimamente la capacidad de rotación mental, ítems con niveles de dificultad baja podrían ser utilizados para este propósito. Así mismo, si se quisiera obtener información de los sujetos que tienen puntajes más altos en rotación mental, podrían utilizarse los ítems con mayores niveles de dificultad para estos efectos.

Cabe destacar, de todas formas, que los resultados muestran que la escala es relativamente fácil de responder, debido a que los índices de dificultad no son muy elevados para ninguno de los ítems, siendo más difíciles los ítems 3 (βi_3=,4882) y 16 (βi_{16}=,2818), respectivamente. En el caso del ítem 16, se observa que la curva característica del ítem no aumenta mucho en términos de probabilidad a pesar de que se aumente el rasgo, lo que podría dar cuenta de que este ítem es, más bien, respondido al azar, ya que los valores de la probabilidad rondan el 50% (ver Figura 9).

Entre los aportes de la TRI para la validación de esta escala, se destaca que esta teoría entrega información complementaria a la TCT. Esto, principalmente porque la TRI cuenta con indicadores que dan cuenta de las propiedades de cada ítem por separado, más que ver el funcionamiento general de una escala, y no solamente tiene en consideración la manera en la que los elementos que la componen se comportan, sino al o a los sujetos que la responden, entendiendo que no todos responderán con el mismo nivel de rasgo y que, por lo tanto, no todos los ítems serán igual de adecuados para todos los sujetos. En este caso en particular, la estructura factorial de los ítems en términos de la adecuación daba cuenta de una manera de poder organizar la información proporcionada por los ítems, sin embargo, la lógica de las dificultades,

aun respondiendo a un fundamento teórico, no contiene mayores conceptos a la base más que el de la discrepancia angular, distinto a otros tipos de test psicológicos que miden variables latentes con un desarrollo teórico mayor y más profundo, que se corresponde con un número determinado de ítems teóricamente respaldados para describir a cada factor, los cuales, a su vez, se establecen en base a constructos con amplias referencias bibliográficas.

En base a la información entregada por la TRI, se puede hacer sugerencia del uso de ítems de esta escala dependiendo del objetivo que se tenga, sobre todo después de los antecedentes entregados en términos de dificultad y discriminación. Se podría proponer, por ejemplo, una versión corta de ERM para screening con los ítems que componen la tercera dificultad—es decir, los ítems 3, 5, 9 y 15—, debido a que, como se ha evidenciado, cuentan con buenos índices de discriminación y dificultades que permiten que la información se concentre en el comportamiento de la mayor parte de los sujetos.

La extensión del uso de la TRI en la validación de instrumentos puede entregarnos mayor información a futuro sobre los ítems a depurar en las escalas, no solamente a través de indicadores como las cargas factoriales, alfa de Cronbach o a través de los índices de ajuste/error, sino que en la información entregada por los índices de discriminación y dificultad, conjugándose así con los propósitos de una posible investigación, ampliando el uso del instrumento, incluso como herramienta de selección, además de contribuir en futuros aportes teóricos para la medición de la inteligencia. Ejemplo de lo anterior podría ser el uso de la TRI en Test Adaptativos Computarizados, CAT, donde la dificultad de los ítems es adaptada al desempeño de los sujetos que realizan las pruebas, por lo que se puede obtener la misma información e incluso más administrando menos ítems, ya que una aplicación extendida de una prueba de inteligencia implica alta demanda cognitiva (Kean y Reilly, 2014).

Otro de los usos a los que se podría extender la teoría de respuesta al ítem es en el ámbito educacional, ya que la medición es un elemento importante para asegurar la calidad de los alumnos en el sistema educativo y a través del uso de esta técnica se podrían obtener mejores estimadores respecto a la dificultad de los ítems administrados así como de la utilización de éstos para la medición de un objetivo determinado de rasgo (Veldkamp y Matteucci, 2013)

Referencias

Attorresi, H. F., Lozzia, G. S., Abal, F. J. P., Galibert, M. S., y 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).

Abedalaziz, N., & Leng, C. H. (2018). The relationship between CTT and IRT approaches in Analyzing Item Characteristics. *MOJES: Malaysian Online Journal of Educational Sciences*, *1*(1), 64-70.

Barnard-Brak, L., Lan, W. Y., & Yang, Z. (2018). Differences in mathematics achievement according to opportunity to learn: A 4pL item response theory examination. *Studies in Educational Evaluation*, *56*, 1-7.

Bichi, A. A., y Talib, R. (2018). Item Response Theory: An Introduction to Latent Trait Models to Test and Item Development. *International Journal of Evaluation and Research in Education (IJERE)*, *7*(2).

Binet, A., y Simon, T. (1905). Méthodes nouvelles pour le diagnostic du niveau intellectuel des anormaux. L'année Psychologique, 11(1), 191-244.

Binet, A., y Simon, T. (1916). The development of intelligence in children. Recuperado de https://ia902609.us.archive.org/13/items/developmentofint00bineuoft/developmentofint00bineuoft.pdf

Boake, C. (2002). From the Binet–Simon to the Wechsler–Bellevue: Tracing the history of intelligence testing. Journal of clinical and experimental neuropsychology, 24(3), 383-405.

Brzezinska, J. (2018). Item Response Theory Models in the Measurement Theory with the Use

of ltm Package in R. *Econometrics*, *22*(1), 11-25.

Colom, R., Contreras, M. J., Botella, J., y Santacreu, J. (2002). Vehicles of spatial ability. Personality and Individual Differences, 32, 903-912.

Contreras, M. J., Colom, R., Hernández, J. M., y Santacreu, J. (2003). Is static spatial performance distinguishable from dynamic spatial performance? A latent-variable analysis. Journal of General Psychology, 130, 277-288.

Colom, R., Contreras, M. J., Shih, P. C., y Santacreu, J. (2003). The assessment of spatial ability with a single computerized test. European Journal of Psychological Assessment, 19, 92-100.

Contreras, M. J., Martínez-Molina, A., Manzanero, A., y Peña, D. (2009). ¿Mejora el rendimiento espacial por efecto de la práctica? Anales de psicología, 25(2), 351.

Contreras, M. J., Martínez-Molina, A., y Santacreu, J. (2012). Do the sex differences play such an important role in explaining performance in spatial tasks?. Personality and individual differences, 52(6), 659-663.

Costello, A. B., y Osborne, J. W. (2005). Best practices in exploratory factor analysis: Four recommendations for getting the most from your analysis. *Practical assessment, research & evaluation*, *10*(7), 1-9.

Derksen, J., Kramer, I., y Katzko, M. (2002). Does a self-report measure for emotional intelligence assess something different than general intelligence? *Personality and individual differences*, *32*(1), 37-48.

Fabrigar, L. R., Wegener, D. T., MacCallum, R. C., y Strahan, E. J. (1999). Evaluating the use of exploratory factor analysis in psychological research. *Psychological methods*, *4*(3), 272.

Galton, F. (1869). *Hereditary genius: An inquiry into its laws and consequences* (Vol. 27). Macmillan.

Gardner, H. E. (2000). *Intelligence reframed: Multiple intelligences for the 21st century*. Hachette UK.

Gregory, R. L., y Zangwill, O. L. (1987). *The Oxford companion to the mind*. Oxford University Press.

Halpern, D. F., y LaMay, M. L. (2000). The smarter sex: A critical review of sex differences in intelligence. Educational Psychology Review, 12(2), 229-246.

Hambleton, R. K., Swaminathan, H., y Rogers, H. J. (1991). Fundamentals of item response theory. Newbury Park, CA: Sage.

Healy, W., y Fernald, G.M. (1911). Tests for practical mental classification. Psychological Monographs, *13*.

Hidalgo-Montesinos, M. D. H., y French, B. F. (2016). Una introducción didáctica a la Teoría de Respuesta al Ítem para comprender la construcción de escalas. *Revista de Psicología Clínica con Niños y Adolescentes*, *3*(2), 13-21.

Holmes, J., Adams, J.W. y Hamilton, C.J. (2008). The relationship between visuospatial sketchpad capacity and children's mathematical skills. European Journal of Cognitive Psychology, 20, 272- 289.

Jaeggi, S. M., Buschkuehl, M., Jonides, J., y Perrig, W. J. (2008). Improving fluid intelligence with training on working memory. Proceedings of the National Academy of Sciences, 105(19), 6829-6833.

Jensen, A. R. (1998). *The g factor: The science of mental ability* (Vol. 648). Westport, CT: Praeger.

Karadavut, T. (2017). Estimation of ítem response theory models when ability is uniformly distributed, 7, 30-37.

Kean, J., y Reilly, J. (2014). Item response theory. *Handbook for Clinical Research: Design, Statistics and Implementation (pp. 195-198). New York, NY: Demos Medical Publishing*.

Kozhevnikov, M., y Hegarty, M. (2001). A dissociation between object manipulation spatial ability and spatial orientation ability. Memory & Cognition, 29, 745–756. doi:10.3758/BF03200477.

Leenen, I. (2014). Virtudes y limitaciones de la teoría de respuesta al ítem para la evaluación educativa en las ciencias médicas. *Investigación en educación médica*, *3*(9), 40-55.

Legg, S., y Hutter, M. (2007). A collection of definitions of intelligence. *Frontiers in Artificial Intelligence and applications*, *157*, 17.

Lohman, D. (1979). Spatial ability: a review and reanalysis of the correlational literature. Technical Report, N.8. Stanford University, Aptitude Research Project, School of Education.

López-Roldán, P., y Fachelli, S. (2016). Análisis factorial. En P. López-Roldán y S. Fachelli, Metodología de la Investigación Social Cuantitativa. Bellaterra (Cerdanyola del Vallès): Dipòsit Digital de Documents, Universitat Autònoma de Barcelona. 1ª edición, versión 3. Edición digital: http://ddd.uab.cat/record/142928.

Mai, Y., Zhang, Z., & Wen, Z. (2018). Comparing Exploratory Structural Equation Modeling and Existing Approaches for Multiple Regression with Latent Variables. *Structural Equation Modeling: A Multidisciplinary Journal*, 1-13.

McGrew, K. (2009). Editorial: CHC theory and the human cognitive abilities project: Standing on the shoulders of the giants of psychometric intelligence research. Intelligence, 37, 1-

10.

Mead, A. D., y Drasgow, F. (1993). Equivalence of computerized and paper-and-pencil cognitive ability tests: A meta-analysis. *Psychological Bulletin*, *114*(3), 449.

Meneghetti, C., Borella, E., y Pazzaglia, F. (2016). Mental rotation training: transfer andmaintenance effects on spatial abilities. Psychological research, 80(1), 113-127.

Miyake, A., Friedman, N. P., Rettinger, D. A., Shah, P., y Hegarty, M. (2001). How are visuospatial working memory, executive functioning, and spatial abilities related? A latent-variable analysis. Journal of Experimental Psychology; General, 130, 621-640.

Mix, K.S. y Cheng, Y.L. (2012). The relation between space and math: Developmental and educationa implications. Advances in Child Development and Behavior, 42, 197-243.

Mora, J.A., y Martín, M.L. (2007). La Escala de la Ingeligencia de Binet y Simon (1905) su recepción por la Psicología posterior. Revista de Historia de la Psicología, 28(2), 307-313.

Muñiz, J. (1998). La medición de lo psicológico. Psicothema, 10(1), 1-21.

Newcombe, N. S. (2010). Picture this: Increasing math and science learning by improving spatial thinking. American Educator, 34, 29–35, 43. doi:10.1037/A0016127.

Nikolaos, T., y Evangelia, F. (2012). Competitive intelligence: concept, context and a case of its application. *Science Journal of Business Management*, *2012*.

Pellegrino, J. W., y Hunt, E. B. (1989). Computer-controlled assessment of static anddynamic spatial reasoning. In R. F. Dillon, & J. W. Pellegrino (Eds.), Testing: Theoretical and applied perspectives. (pp. 174-198). New York, NY, England: Praeger Publishers.

Peters, M., y Battista, C. (2008). Applications of mental rotation figures of the Shepard and Metzler type and description of a mental rotation stimulus library. Brain and cognition,

66(3), 260-264.

Peterson, J. (1925). Early conceptions and tests of intelligence. Yonkers-on-Hudson, NY: World Book.

Pyle, W.H. (1913). The examination of school children: A manual of directions and norms. New York: Macmillan.

Quiroga, M. Á., Martínez-Molina, A., Lozano, J. H., & Santacreu, J. (2011). Reflection-impulsivity assessed through performance differences in a computerized spatial task. Journal of Individual Differences.

Reise, S. P. (2014). Item response theory. *The Encyclopedia of Clinical Psychology*, 1-10.

Rosas, R., Tenorio, M., Pizarro, M., Cumsille, P., Bosch, A., Arancibia, S., Carmona-Halty, M., Pérez-Salas, C., Pino, E., Vizcarra, B., y Zapata-Sepúlveda, P. (2014). Estandarización de la Escala Wechsler de Inteligencia para Adultos: cuarta edición en Chile. *Psykhe (Santiago)*, *23*(1), 1-18.

Santacreu, J., y Rubio, V. (1998). SODT-R and SVDT: Dynamic computerized tests for the assessment of spatial ability. Technical Report. Madrid: Universidad Autónoma de Madrid.

Schiff, W., y Oldak, R. (1990). Accuracy of judging time to arrival: Effects of modality, trajectory and gender. Journal of Experimental Psychology: Human Perception & Performance, 16, 303-316.

Shah, P., y Miyake, A. (1996). The separability of working memory resources for spatial thinking and language processing: An individual differences approach. Journal of Experimental Psychology: General, 125, 4-27.

Shea, D. L., Lubinski, D. y Benbow, C. P. (2001). Importance of assessing spatial ability in

intellectually talented young adolescents: A 20-year longitudinal study. Journal of Educational Psychology, 93, 604–614.

Shepard y Metzler (1971). "Mental Rotation of Three-Dimensional Objects". Science. 171 (3972): 701–703. doi:10.1126/science.171.3972.701

Shriki, A., Barkai, R., y Patkin, D. (2017). Developing mental rotation ability through engagement in assignments that involve solids of revolution. *The Mathematics Enthusiast*, *14*(1), 541-562.

Silverman, I., Choi, J., y Peters, M. (2007). The hunter-gatherer theory of sex differences in spatial abilities: Data from 40 countries. Archives of Sexual Behavior, 36, 261-268.

Silverman, I., y Eals, M. (1992). Sex differences in spatial abilities: Evolutionary theory and data. In J. H. Barkow, L. Cosmides, & J. Tooby (Eds.), The adapted mind: Evolutionary psychology and the generation of culture (pp. 531–549). New York: Oxford Press.

Terman, L. M. (1916). *The measurement of intelligence: An explanation of and a complete guide for the use of the Stanford revision and extension of the Binet-Simon intelligence scale*. Houghton Mifflin.

Thorndike, R.M., y Lohman, D.F. (1990). A century of ability testing. Chicago: Riverside.

Uttal, D. H., y Cohen, C. A. (2012). Spatial thinking and STEM education: When, why, and how? Psychology of Learning and Motivation, 57, 147-181.

Uttal, D. H., Meadow, N. G., Tipton, E., Hand, L. L., Alden, A. R., Warren, C., y Newcombe, N. S. (2012). The malleability of spatial skills: A meta-analysis of training studies. Psychological Bulletin. Available at: http://www.spatialintelligence.org/publications_pdfs/uttal_etal_june-2012.pdf

Vandenberg, S. G., y Kuse, A. R. (1978). Mental rotations, a group test of 3-dimensional spatial

visualization. Perceptual and Motor Skills, 47, 599–604. doi:10.2466/pms.1978.47.2.599.

Van Der Linden, W. J., y Hambleton, R. K. (1997). Item response theory: Brief history, common models, and extensions. In Handbook of modern item response theory (pp. 1-28). Springer New York.

Veldkamp, B. P., y Matteucci, M. (2013). Bayesian computerized adaptive testing. *Ensaio: Avaliação e Políticas Públicas em Educação*, *21*(78), 57-82.

Velicer, W. F., & Fava, J. L. (1998). Affects of variable and subject sampling on factor pattern recovery. *Psychological methods*, *3*(2), 231.

Wai, J., Lubinski, D., y Benbow, C. P. (2009). Spatial ability for STEM domains: Aligning over 50 years of cumulative psychological knowledge solidifies its importance. *Journal of Educational Psychology*, *101*(4), 817.

Wechsler, D. (1939). The measurement of adult intelligence. Baltimore: Williams & Wilkins.

Wechsler, D. (1946). The Wechser-Bellevue Intelligence Scale. Form II. Manual for administering and scoring the test. New York: Psychological Corporation.

Wechsler, D. (1949). Wechsler Intelligence Scale for Children. Manual. New York: Psychological Corporation.

Wechsler, D. (1979). The psychometric tradition: Developing the Wechsler Adult Intelligence Scale. Paper presented at the 87th Annual Meeting of the American Psychological Association, New York.

Wechsler, D. (2008). *Wechsler Adult Intelligence Scale–Fourth Edition (WAIS–IV)*. San Antonio, TX: The Psychological Corporation.

Wolfgang, C., Stannard, L., y Jones, I. (2003). Advanced constructional play with LEGOs

among preschoolers as a predictor of later school achievement in mathematics. Early Child Development and Care, 173(5), 467-475.

Wright, R., Thompson, W. L., Ganis, G., Newcombe, N. S., y Kosslyn, S. M. (2008). Training generalized spatial skills. Psychonomic Bulletin & Review, 15(4), 763-771

Yerkes, R.M. (Ed.) (1921). Psychological examining in the United States Army. Memoirs of the National Academy of Sciences, 15 (Parts 1-3), Washington DC: Government Printing Office.

Yerkes, R.M., Bridges, J.W., y Hardwick, R.S. (1915). A point scale for measuring mental ability. Baltimore: Warwick & York.

Zenderland, L. (1998). Measuring minds: Henry Herbert Goddard and the origins of American intelligence testing. New York: Cambridge University Press.