

UNIVERSIDAD NACIONAL AUTÓNOMA DE HONDURAS
FACULTAD DE CIENCIAS SOCIALES
ESCUELA DE CIENCIAS PSICOLÓGICAS
MAESTRÍA EN PSICOMETRÍA Y EVALUACIÓN EDUCATIVA



TESIS

**Estudio de Validación para un Test de Razonamiento Mecánico para el Proceso de
Admisión a la Universidad Pedagógica Nacional de Honduras**

Presentada por

Ana Belly Ortiz Donaire

Para optar al grado de Máster en Psicometría y Evaluación Educativa

Asesor Temático: Leví Astul Castro Ordóñez, MSc

Asesor Metodológico: German Edgardo Moncada Godoy, PhD

Tegucigalpa, M.D.C., Honduras

Junio, 2022

Tegucigalpa, M.D.C., 30 de Junio de 2022

Señores

Universidad Nacional Autónoma de Honduras

Coordinadores Maestría en Psicometría y Evaluación Educativa

Estimados señores:

Por este medio autorizo a la UNAH para que publique en su REVISTA DIGITAL, WEB, CD Y CUALQUIER OTRO MEDIO, y uso de beneficio sociales y académicos de la sociedad hondureña mi trabajo de tesis titulado: “Estudio de Validación para un Test de Razonamiento Mecánico para el Proceso de Admisión a la Universidad Pedagógica Nacional de Honduras”, el cual he presentado como parte de los requisitos para optar por la Maestría en Psicometría y Evaluación Educativa.

Atentamente,

Ana Belly Ortiz Donaire

Cuenta No. PSE100220

AUTORIDADES UNIVERSITARIAS

Doctor Francisco José Herrera Alvarado
Rector Interino

Máster Belinda Flores de Mendoza
Vice-Rectora Académica

Doctora Jessica Patricia Sánchez Medina
Secretaria General

Doctora Leonarda del Carmen Andino
Directora de Docencia

Doctor Armando Euceda
Director Sistema de Estudios de Postgrado

Máster Carmen Julia Fajardo
Decana Facultad de Ciencias Sociales

Máster María José Irías Escher
Coordinadora General Postgrados Facultad de Ciencias Sociales

Máster Lina María Mendoza Recarte
Coordinadora Académica Maestría en Psicometría y Evaluación Educativa

Dedicatoria

Dedico este trabajo de investigación a mis amados padres, quienes me animaron en este arduo y largo proceso del cual muchas veces perdí esperanza y valor.

DEGT-UNAH

Agradecimiento

Agradezco a Dios mi Padre Celestial y a su Hijo Jesucristo, mi Señor y Salvador, en quien están escondidos todos los tesoros de la sabiduría y el conocimiento, que me ha concedido, en su infinita gracia, la porción de sabiduría necesaria para realizar esta tesis. Porque nada es imposible para Él, y al pedirle por sabiduría, concede abundantemente y sin reproche.

Agradezco a la Universidad Pedagógica Nacional de Honduras por brindarme la oportunidad de trabajar con la base de datos del test de admisión.

Agradezco a mi asesor temático Leví Castro, por su ayuda y orientación durante el desarrollo del trabajo.

Por último, agradezco a mi compañera Alejandra Cáceres, por su apoyo y ánimo constante, sin el cual hubiera sido un camino muy confuso y solitario.

Resumen

El Test de Razonamiento Mecánico fue diseñado con el propósito de precisar el perfil de ingreso y seleccionar los mejores aspirantes para la oferta académica de la UPNFM. En el diseño, el test original consta de 40 ítems de selección única con tres opciones de respuesta. La investigación fue de tipo cuantitativo y de diseño instrumental, en una muestra conformada por 287 estudiantes de decimo grado de bachillerato pertenecientes a instituciones gubernamentales de Tegucigalpa.

El objetivo fue realizar un estudio de validación a partir de los datos obtenidos en el estudio piloto. El trabajo consistió, primeramente, en un análisis del funcionamiento de las opciones de respuesta mediante el programa Rstudio, resultando 20 ítems descartados de análisis posteriores por presentar problemas en el funcionamiento de claves y distractores. Luego, se calcularon los parámetros psicométricos de los ítems desde la TCT y su confiabilidad por medio del alfa de Cronbach y Omega de MacDonald, mediante los programas Rstudio y Jamovi, revelando índices bajos de confiabilidad y discriminación para algunos ítems. Seguidamente, se evaluó la dimensionalidad del test e independencia local de los ítems mediante el programa Factor Analysis, IRT PRO y SPSS. Se realizó un AFE, utilizando la matriz de correlaciones tetracóricas y el método de extracción de Ejes Principales y Promin como método de rotación. El análisis paralelo afirmó la unidimensionalidad de la escala conformada por 10 ítems ($GFI=0.951$; $SRMR=0.0830$). Finalmente, se procedió a estimar parámetros desde la TRI aplicando el modelo logístico de dos parámetros y su función de información, mediante el programa IRT PRO, resultando 10 ítems que se ajustan a la escala. Tomando en cuenta los resultados anteriores, se concluye que solamente 10 ítems cumplen con los criterios métricos de calidad, por lo que se sugiere la inclusión de más ítems al Test de Razonamiento Mecánico para mejorar los índices y no comprometer la representatividad del constructo medido.

Palabras clave: Estudio de validación, Test de Razonamiento Mecánico, Aptitud Académica, Selección Universitaria.

Índices de contenido

Dedicatoria.....	i
Agradecimiento.....	ii
Resumen.....	iii
Introducción.....	1
Capítulo 1. Planteamiento del problema.....	3
1.1 Situación Problema	3
1.2 Objetivos	6
1.2.1 General	6
1.2.2 Específicos.....	6
1.3 Justificación.....	7
Capítulo 2. Marco teórico	9
2.1 Razonamiento mecánico	9
2.2 Medición del razonamiento mecánico	11
2.3 Validez	13
2.4 Teorías de la medición	14
2.5 Teoría Clásica de los Test	14
2.5.1 Características del modelo TCT	15
2.5.2 Supuestos del modelo TCT	16
2.5.3 Cálculo de parámetros desde la TCT.....	17
2.6 Teoría de Respuesta al Ítem	20
2.6.1 Características de la TRI	21
2.6.2 Supuestos de la TRI.....	23
2.6.3 Modelos de la TRI.....	24

2.7 Análisis de opciones de respuesta	27
2.8 Análisis de dimensionalidad	29
2.8.1 Análisis factorial exploratorio.....	30
2.8.2 Supuestos del análisis factorial	31
2.8.3 Fases del análisis factorial.....	31
2.9 Independencia local.....	36
2.10 Fiabilidad.....	38
2.10.1 Fiabilidad desde la TCT	38
2.10.2 Fiabilidad desde la TRI	42
Capítulo 3. Marco contextual.....	44
3.1 Instrumentos de aptitud académica	44
3.2 Sistemas de selección universitaria.....	45
3.3 Instrumentos de selección para formación docente	49
Capítulo 4. Metodología	53
4.1 Tipo de investigación	53
4.2 Población y muestra	53
4.3 Variables	54
4.4 Instrumento de medición.....	54
4.5 Plan de análisis de datos.....	55
4.6 Procedimientos.....	55
Capítulo 5. Resultados de la investigación	57
5.1 Análisis de alternativas de respuestas	57
5.1.1 Análisis de proporción de respuestas	57
5.1.2 Análisis del poder discriminativo de las opciones de respuesta.....	58

5.2 Cálculo de parámetros psicométricos desde la TCT y su confiabilidad	59
5.3 Análisis de dimensionalidad	61
5.4 Análisis de independencia local	65
5.5 Estimación de parámetros psicométricos desde la TRI y su función de información	65
Capítulo 6. Discusión de los resultados	68
6.1 Análisis desde la TCT	69
6.2 Análisis desde la TRI	70
6.3 Limitaciones	72
6.4 Recomendaciones para estudios futuros	73
Capítulo 7. Conclusiones	74
Capítulo 8. Referencias.....	76
Declaración de ética de la investigación.....	85
Capítulo 9. Apéndice	86
Apéndice 1. Funcionamiento de las opciones de respuesta de los ítems	86
Apéndice 2. Curvas Características de los ítems y su nivel de información.....	89
Apéndice 3 . Curva característica del test	91
Apéndice 4. Curva de información total	92

Índice de tablas

Tabla 1. Formulación del modelo TCT	15
Tabla 2. Matriz de Operacionalización de Variables.....	54
Tabla 3. Análisis de proporción de respuesta	58
Tabla 4. Análisis del poder discriminativo	59

Tabla 5. Parámetros TCT.....	60
Tabla 6. Matriz de correlaciones tetracóricas	62
Tabla 7. Proporción de varianza explicada.....	63
Tabla 8. Análisis paralelo	63
Tabla 9. Matriz de estructura	64
Tabla 11. Parámetros psicométricos desde la TRI.....	67

Índice de figuras

Figura 1. Curva Característica de un ítem	22
Figura 2. Función de información.....	43

Introducción

El Test de Razonamiento Mecánico diseñado por la Coordinación del Examen de Admisión de la Universidad Pedagógica Nacional Francisco Morazán (UPNFM), procura evaluar en los aspirantes, la aptitud para comprender sistemas mecánicos en torno a la fuerza, movimiento, energía, operación de máquinas y herramientas. La inclusión del Test de Razonamiento Mecánico al Examen de Admisión se justifica sobre la base que permitirá identificar el nivel de aptitud y determinar el futuro desempeño de los postulantes en áreas relacionadas al mismo.

El razonamiento mecánico se define como la capacidad para identificar el funcionamiento de los componentes de un sistema mecánico, lo que implica la aplicación de principios físicos y matemáticos, simulaciones mentales e inferencias de sistemas mecánicos (González et al., 2008; Hegarty, 2004; Velásquez Ospina, 2008). Esta aptitud ha sido evaluada desde inicios del siglo pasado, mediante test como el Test de Habilidad Mecánica de Stenquist (Wesley, 1937), el Test de bloques ondulados de O'Connor (Remmers y Smith, 1936), el Test de Aptitudes Mecánicas de Mac Quirre (Rosales López, 2012), el Test de Comprensión Mecánica de Bennett (Bennet, 2008), entre otros.

La ciencia que estudia la medición de constructos psicológicos es la psicometría, de la que destacan dos grandes teorías de los test: La Teoría Clásica de los Test y la Teoría de Respuesta al Ítem, las cuales son el marco de referencia que guía y condiciona el diseño, análisis y utilización de los test (Güler et al., 2014).

Por las consideraciones anteriores, el presente trabajo tuvo como propósito realizar un estudio de validación del Test de Razonamiento Mecánico a partir de los datos obtenidos en el estudio piloto. Para esto se analizó el funcionamiento de las opciones de respuesta, se determinó los parámetros psicométricos de los ítems desde la TCT y su confiabilidad, asimismo, para la estimación de los parámetros desde la TRI, se analizó la dimensionalidad e independencia local del test. Se utilizó el modelo logístico de dos parámetros y se estimó la función de información.

Los siguientes capítulos presentan primeramente un panorama del marco teórico basado en la literatura actual en el campo de la psicometría, particularmente en

procedimientos metodológicos para la validación de test. Asimismo, el marco contextual en el que se desarrollan los test de aptitud, de razonamiento mecánico y sistemas de selección a nivel internacional, regional y nacional. Más adelante, se detalla la metodología utilizada en el desarrollo del estudio. Por último, los resultados son presentados y discutidos, cerrando con unos párrafos de conclusión.

Capítulo 1.

Planteamiento del problema

El presente apartado expone los argumentos que justifican la necesidad del estudio de validación del Test de Razonamiento Mecánico, el cual ha sido diseñado para integrarse al Test de Admisión de la UPNFM. La estructura lógica en que se desarrollan las ideas es la siguiente: primeramente, se describe la situación problemática donde se identifica la situación que da origen al estudio, con relación a la cantidad de aspirantes y en contraste con la importancia de contar con un sistema de admisión válido. Además, se subraya que, el Test de Admisión actualmente utilizado no cuenta con un Test de Razonamiento Mecánico, y que, para ser incorporado requiere un estudio de validación. Asimismo, se señala que, identificar el grado de razonamiento mecánico de los aspirantes, contribuirá a una caracterización más detallada, lo cual dará soporte a las decisiones de selección y de esta forma garantizar éxito académico de los futuros estudiantes de la carrera docente.

1.1 Situación Problema

La UPMFM, es la única universidad en Honduras formadora de docentes del sistema educativo. Esta institución, regularmente atiende una gran demanda de aspirantes a nivel nacional. Según la Memoria Estadística del año 2019, la cantidad de estudiantes matriculados en el primer periodo académico fue de 21,851 a nivel nacional, de los cuales 2,614 corresponde a estudiantes de primer ingreso (UPNFM, 2019). La Coordinación de Examen de Admisión reporta que, la cantidad de aspirantes que se sometieron al Test en el año 2019, fue más de 6,000 personas, de los cuales 3,272 fueron preseleccionados para iniciar estudios en el año 2020 (CEA, 2019a).

Como lo establece el Reglamento del Régimen Académico para programas de pregrado, el Test de Admisión es un requisito de ingreso, bajo la responsabilidad de la Vicerrectoría Académica (UPNFM, 2005). Como parte de esta labor, la UPNFM ha diseñado el Examen de Admisión, siendo “una de las pocas instituciones de educación superior que, a nivel nacional, es propietaria del Test de Admisión que utiliza” (CEA, 2019b, p. 3). Este Test sigue procedimientos psicométricos internacionales en cuanto a su diseño, mantenimiento y estudios de validación.

El Test de Admisión, constituye una prueba de conocimientos y/o aptitud que corresponde a las competencias y habilidades generales del aspirante a ingresar a la universidad (UPNFM, 2005). El Test actual “es una batería o conjunto de subtests, construido tomando como referencia el Test de Aptitudes Diferenciales (TAD) de G.K. Bennett, H.G. Seashore y A.G. Wesman, la Teoría Tipológica de John Holland, y el Inventario de Intereses Ocupacionales de Lee-Thorpe” (CEA, 2019b, p. 4).

Sin embargo, hasta la fecha, el Test de Admisión de la UPNFM, no contempla un Test de Razonamiento Mecánico, por lo que no se puede identificar el nivel de aptitud o habilidades de los aspirantes en áreas que requieren comprensión de principios de fuerzas físicas y mecánicas (Costa Neiva, 1996).

Sobre la base de las consideraciones anteriores, la Coordinación del Examen de Admisión ha diseñado el Test de Razonamiento Mecánico, para poder ser incorporado en la batería del Test de Admisión, actualmente se encuentra en la etapa de estudio piloto, por lo que se requiere un estudio de validación. Este estudio deberá contemplar el análisis de las opciones de respuesta, parámetros psicométricos de los ítems, composición factorial del test, independencia local, entre otros.

Al incorporar el Test de Razonamiento Mecánico al Sistema de Admisión de la UPNFM, se podrá medir acertadamente la capacidad para comprender principios básicos y mecánicos en torno al movimiento, las propiedades de las herramientas y el funcionamiento de las máquinas por medio de las ilustraciones que componen los ítems del Test (Velásquez Ospina, 2008).

Estudios afirman que existen perfiles aptitudinales claramente diferenciados por áreas de conocimiento (González et al., 2008). Además, Liu y Schunn (2017) muestran que niveles altos de razonamiento mecánico se asocian con el uso frecuente y complejo de estrategias matemáticas. Estas habilidades son implementadas en áreas como ciencias, computación, matemáticas, tecnología (Injoque-Ricle et al., 2019).

En tal sentido, resulta oportuna la incorporación del Test de Razonamiento Mecánico al Test de Admisión a la UPNFM, ya que se podrá medir y conocer la aptitud específica en esta área y atraer los mejores perfiles estudiantiles, asegurando un óptimo cumplimiento en el perfil de egreso; y por lo tanto, se logrará la formación de individuos

aptos para el ejercicio de la docencia que contribuyan a la calidad educativa del país (Aguirre-Gómez et al., 2015; CEA, 2019b).

Todo lo anterior, evidencia la necesidad de establecer un sistema de selección válido para los aspirantes a la carrera docente, que logre identificar las aptitudes que permitan predecir su efectividad futura como profesionales (Elacqua et al., 2017). Ya que, “la calidad de un sistema educativo no puede ser mejor que la de su profesorado” (Barber y Mourshed, 2007).

DEGT-UNAH

1.2 Objetivos

1.2.1 General

Realizar un estudio de validación del Test de Razonamiento Mecánico a partir de los datos obtenidos en el estudio piloto.

1.2.2 Específicos

Analizar el funcionamiento de las opciones de respuesta de los ítems del Test de Razonamiento Mecánico.

Determinar los parámetros psicométricos de los ítems del Test de Razonamiento Mecánico.

Analizar la dimensionalidad del Test de Razonamiento Mecánico por medio del análisis factorial.

Analizar la independencia local de los ítems del Test de Razonamiento Mecánico.

Determinar la fiabilidad de las puntuaciones del Test de Razonamiento Mecánico.

1.3 Justificación

En este espacio se describe el valor del conocimiento que aportará el estudio de validación del Test de Razonamiento Mecánico, además de su utilidad en el fortalecimiento del sistema de selección de aspirantes, asegurando la calidad de la formación docente. Asimismo, aportando científicamente en la aplicación de procedimientos y metodologías de validación de instrumentos ya establecidos.

Se plantea entonces, desde la perspectiva más general que, la calidad de los docentes es la variable más determinante para la calidad del sistema educativo (Aguilar Parra et al., 2016). Por tal razón, la búsqueda de procedimientos adecuados para la selección de los futuros docentes es una materia de vital importancia en todos los sistemas educativos (Egido Gálvez, 2010).

Además, dado que el Test de Admisión es una evaluación de alto impacto para los aspirantes, puesto que, los resultados tienen implicaciones económicas, sociales, educativas e incluso físicas, se requiere un estudio de evidencia de validez, para que las inferencias que se obtengan de los resultados sean apropiadas (Mendiola et al., 2020).

Sobre esta idea se justifica la utilidad del estudio de validez, ya que: Se incorporará un test a la batería del Test de Admisión, que eficazmente identifique la aptitud en cuanto al razonamiento mecánico de los aspirantes a la UPNFM. Esto incrementará la probabilidad de éxito de culminar la formación docente en carreras específicas para las cuales el razonamiento mecánico es determinante (CEA, 2018). Por otro lado, los resultados del Test permitirán la formulación de investigaciones referentes al ingreso, permanencia y egreso de estudiantes, así como la toma de decisiones y definición de políticas institucionales (CEA, 2019b). Esto a su vez, beneficiará la calidad de la enseñanza, ya que, conocer el nivel de aptitud sobre razonamiento mecánico encaminará metodologías de instrucción en respuesta a sus debilidades por medio del entrenamiento directo (Velásquez Ospina, 2008).

El estudio de evidencias de validez del Test de Razonamiento Mecánico es relevante ya que: Disciplinas como ciencias naturales, computación, matemáticas, tecnología son muy importantes para la sociedad actual y están vinculadas a la capacidad de los sujetos para razonar acerca del funcionamiento de sistemas mecánicos. Por lo que contar

con un Test de Razonamiento Mecánico, permitirá identificar aspirantes más competentes para estas áreas (Injoque-Ricle et al., 2019).

Este estudio aportará teórica y metodológicamente: procedimientos claros para llevar a cabo estudios de validez por medio de un análisis psicométrico del Test de Razonamiento Mecánico. Ya que la validez es el principio más importante de la psicometría (Martínez Arias et al., 2006). Como afirman Cizek et al. (2010) “La validez es una de las mayores deidades en el panteón de la psicometría” (p.733).

Capítulo 2.

Marco teórico

Este apartado pretende describir el fundamento teórico que sustenta el estudio de validación del Test de Razonamiento Mecánico. Los temas tratados se irán desarrollando a partir de la información teórica recogida de diversas fuentes científico-académicas, presentándose desde los planos más globales, hasta los más concretos. El primero de ellos es el razonamiento mecánico y su medición, seguido por la validez. Posteriormente, se abordarán las dos grandes Teorías de Medición (Teoría Clásica del Test y Teoría de Respuesta al Ítem) que sustentan el estudio psicométrico, en torno a sus características, supuestos, cálculo de parámetros y estimación de modelos. Por último, se presentarán la base teórica de los procedimientos desarrollados en el estudio de validación: Análisis de opciones de respuesta, dimensionalidad, independencia local y fiabilidad.

2.1 Razonamiento mecánico

El razonamiento mecánico es un proceso psicológico de orden superior a través del cual una persona puede identificar la relevancia y funcionamiento de los componentes principales de un sistema mecánico, así como inferir cómo interactúan entre ellos (Injoque-Ricle et al., 2019). Esta habilidad supone la comprensión y aplicación los principios mecánicos y físicos en torno al movimiento, las propiedades de las herramientas y el funcionamiento de las máquinas en situaciones conocidas (González et al., 2008; Velásquez Ospina, 2008).

El razonamiento mecánico es considerado como un elemento de la inteligencia general (Bennet, 2008; Juliá y Antolí, 2018). En este respecto, Hegarty (2004) declara que el razonamiento mecánico se deriva particularmente de la inteligencia espacial, ya que la mecánica es la ciencia del movimiento y el movimiento es una propiedad de la inteligencia espacial, la cual es definida como la capacidad para construir, mantener y transformar representaciones mentales.

En esta línea, el razonamiento mecánico está basado en representaciones mentales de sistemas mecánicos y depende de información visual como ubicación de objetos, forma,

conectividad y relaciones entre objetos. Procesos cognitivos como la memoria de trabajo visoespacial y la velocidad de procesamiento están implicados en la habilidad para entender los sistemas mecánicos (Hegarty, 2004; Injoque-Ricle et al., 2019).

Injoque-Ricle et al. (2019) Clasifica los sistemas mecánicos en dos tipos:

- Sistema de engranajes: habilidad para inferir el movimiento de una pieza de un engranaje a partir del movimiento de otra pieza adyacente.
- Sistema de rueda: habilidad para determinar el movimiento del conjunto a partir de la combinación de los movimientos individuales de las partes que lo componen.

Por otro lado, el desempeño en tareas que requieren habilidad mecánica involucra el uso de ciertas estrategias como la elaboración de inferencias mecánicas, simulaciones mentales, diagramas y aplicación de principios matemáticos y físicos. Una inferencia mecánica es un proceso cognitivo que permite deducir información sobre el movimiento y funcionamiento de mecanismos, mediante representaciones o modelos mentales. Al realizar una inferencia mecánica, las personas conscientemente hacen una simulación mental interna donde se representan los sistemas mecánicos (Hegarty, 2004).

La simulación mental es una representación mental del proceso secuencial de un sistema mecánico, donde se analizan cada uno de sus componentes e interacciones, lo cual implica un conocimiento implícito sobre su funcionamiento (Hegarty, 2004). De igual forma, Anderson et al. (2002) señalan que los diagramas ayudan a entender los sistemas mecánicos, ya que representan relaciones espaciales mentales.

Otra estrategia que favorece la resolución de problemas de índole mecánico es hacer uso de principios matemáticos. Niveles altos de razonamiento mecánico se asocian con el uso frecuente y complejo de principios matemáticos. A su vez, la aptitud mecánica puede favorecer el aprendizaje y aplicación de las matemáticas (Liu y Schunn, 2017). Además, la capacidad de razonamiento mecánico requiere un conocimiento tácito o implícito sobre principios físicos (Hegarty, 2004). Personas con alto nivel de comprensión mecánica tienen mayor facilidad para entender y aprender los principios físicos de operación y reparación de sistemas complejos (Bennet, 2008; Juliá y Antolí, 2018).

De igual forma, el razonamiento mecánico puede ser potenciado mediante el aprendizaje de conceptos mecánicos, favoreciendo el establecimiento de conexiones entre aspectos concretos y abstractos, mecanismos físicos y análisis matemáticos que permitan aplicar apropiadamente ese conocimiento para la resolución de diferentes problemas (McKenna y Agogino, 2004).

En síntesis, el razonamiento mecánico es una habilidad que permite identificar los componentes de un sistema mecánico y su interacción, lo que supone la aplicación de principios mecánicos y físicos. El razonamiento mecánico se relaciona estrechamente con la inteligencia espacial debido a que se basa en el establecimiento de representaciones espaciales y simulaciones mentales de sistemas mecánicos.

2.2 Medición del razonamiento mecánico

La evaluación de la aptitud mecánica permite potenciar su desarrollo y el aprendizaje de nuevos conocimientos, creando nuevas capacidades para desenvolverse eficazmente en áreas como ciencia, tecnología, ingeniería y matemáticas (Juliá y Antolí, 2018; Lázaro Tortosa, 2015). El aprendizaje de conceptos mecánicos, principios básicos sobre fuerza, movimiento y energía son requeridos para dichas carreras (McKenna y Agogino, 2004). En su estudio, González et al. (2008) encontró que estudiantes de carreras de ciencias exactas e ingenierías presentaron mayores puntajes en razonamiento mecánico, en contraste con estudiantes de ciencias sociales y psicología.

Liu y Schunn (2017) afirman que, la aptitud mecánica puede favorecer el aprendizaje y aplicación de principios matemáticos, asimismo influye positivamente en la resolución de problemas matemáticos. También, el dominio de la habilidad tecnológica involucra la comprensión de conceptos mecánicos, habilidades manuales y habilidad para identificar y solventar problemas mecánicos (McKenna y Agogino, 2004).

Sobre la base de las consideraciones anteriores, la medición del razonamiento mecánico es muy importante, ya que permite identificar esta aptitud en los individuos y así potenciar su desarrollo para asegurar un óptimo desempeño. Entre algunos de los test que evalúan esta aptitud se encuentran el Test de Aptitud Diferencial diseñado en 1947 por George K. Bennett, Harold G. Seashore y Alexander G. Wesman y la Batería para la

Evaluación de Aptitudes diseñado por D. Arribas, P. Santamaría, F. Sánchez-Sánchez, I. Fernández-Pinto y C. Arellano. Cabe mencionar que estos test evalúan otras aptitudes, pero incorporan un subtest sobre el razonamiento mecánico. Por otro lado, existen test enfocados completamente en el razonamiento mecánico como el Test de Habilidad Mecánica de Stenquist, el Test de bloques ondulados de O'Connor, el Test de Aptitudes Mecánicas de Mac Quirre, el Test de Comprensión Mecánica de Bennett, entre otros.

El Test de Habilidad Mecánica de Stenquist, creada en 1914, fue utilizado en la evaluación militar durante la segunda guerra mundial. Consiste en dos partes con ítems basados en imágenes para evaluar percepción mecánica y habilidad para razonar problemas mecánicos. En la primera parte se presentan una imagen y el evaluado debe responder como esta se relaciona con otras 5 imágenes. La segunda parte consiste en problemas mecánicos basados en piezas mecánicas. Tiene una duración de 30 minutos (Wesley, 1937).

De igual forma, el Test de bloques ondulados de O'Connor, diseñado en 1927 por Johnson O'Connor. Es un test dirigido a niños y adultos que mide la habilidad para visualizar y construir objetos en tercera dimensión. Está compuesto por nueve bloques con superficie curvas que al unirse forman un rectángulo. Se asume que el tiempo requerido para organizar y unir las piezas es un buen indicador de aptitud mecánica innata. La habilidad para resolver esta tarea correlaciona con éxito en áreas de ingeniería y mecánica por lo que se utiliza en la selección de personal industrial y mecánico donde se requiere la visualización estructural. Este test resulta un eficaz predictor en el desempeño profesional que requiera análisis y solución de problemas mecánicos, creación y reparación de máquinas (Remmers y Smith, 1936; Wesley, 1937).

El Test de Comprensión Mecánica de Bennett, publicado por primera vez en 1969, es un instrumento de evaluación para medir la habilidad para percibir y entender la relación de fuerzas físicas y elementos mecánicos en situaciones prácticas. El test permite predecir el desempeño en tareas que requieren la aplicación de principios mecánicos como en ingeniería y mecánica. Es utilizado en la selección de personal, asimismo es utilizado en contextos académicos para identificar el grado de comprensión mecánica de los estudiantes con relación a programas de instrucción específicos. Está compuesto por 68 ítems con ilustraciones simples frecuentemente presentes en tareas mecánicas, basados problemas de

engranaje, polea y ruedas. Su aplicación es individual y colectiva. Puede ser administrado en papel y lápiz y también en computadora. Tiene una duración de 30 minutos (Bennet, 2008).

Por último, el Test de Aptitudes Mecánicas, diseñado por Mac Quarrie en 1996, evalúa diversos aspectos de la inteligencia técnica, habilidades de rapidez y precisión manual. Generalmente se utiliza en la predicción de éxito en tareas profesionales y escolares de naturaleza mecánica. Está dirigido a personas mayores de 10 años. Su aplicación es individual o colectiva con una duración de 30 minutos aproximadamente. La prueba consta de siete subtests: trazado, marcado, punteado, copiado, localizado, recuento y laberinto (Rosales López, 2012).

En suma, identificar el razonamiento mecánico permite formar una idea del desempeño futuro de los individuos en áreas de estudio que requieren esta habilidad, como las matemáticas, ciencias y tecnología. A lo largo de la historia se han desarrollado test que evalúan el razonamiento mecánico, los cuales son de mucha utilidad para identificar y potenciar su desarrollo.

Habiendo descrito el significado del razonamiento mecánico y su medición, a continuación, se desarrollarán temáticas enfocadas en el análisis psicométrico que sustenta el estudio.

2.3 Validez

La validez es definida por los Estándares para Pruebas Educativas y Psicológicas como “el grado en que toda la evidencia disponible relevante, es acumulada para respaldar la interpretación prevista de los puntajes de una prueba para el uso propuesto” (APA, AERA, NCME, 2018, p. 14). La validez se ocupa del significado de las medidas obtenidas al aplicar un test y de los errores sistemáticos que se pueden producir en esas medidas, por lo que la validez es un criterio métrico de la calidad del test (Navas Ara, 2001).

Martínez Arias et al. (2006) añaden que la validez es el grado en que el uso del test y su utilidad puede ser justificado. Por tanto, la legitimidad y eficiencia de los usos de las mediciones dependen de la validez (Prieto y Delgado, 2010). El proceso de validación implica la recolección de evidencias significativas, delimitada por su uso, propósito, interpretación propuesto para cada prueba y los posibles errores sistemáticos involucrados en

el proceso (Elousa, 2003). Dichas evidencias pueden ser obtenidas de fuentes de información provenientes del contenido, proceso de respuesta, estructura interna, relación con otras variables y consecuencias del test (APA, AERA, NCME, 2018).

2.4 Teorías de la medición

La ciencia que permite la medición de fenómenos psicológicos es la psicometría, ésta se fundamenta en modelos de medición, basados en la estadística y la probabilidad, para poder establecer relación entre las puntuaciones de un test, la habilidad o rasgo psicológico evaluado y el error de medida (Martínez Arias et al., 2006).

Las teorías de los test son el marco teórico de referencia que guían y condicionan el proceso de diseño, análisis y utilización de los test conforme a los avances en la psicometría y estadística (Güler et al., 2014). Estas abordan los problemas de medición en la evaluación permitiendo una utilización rigurosa y científica (Muñiz, 2010). El objetivo principal de dichas teorías es diseñar un modelo estadístico que fundamente las puntuaciones de los test y la estimación de los errores de medida que afectan dichas puntuaciones (Muñiz, 2018).

En la actualidad, las teorías de los test contienen diferentes modelos psicométricos que permiten evaluar la calidad de medida (Attorresi et al., 2009). Normalmente responden a dos enfoques principales: Los que se basan en la puntuación total obtenida a partir del conjunto de ítems y los que se basan en los ítems individualmente y en los patrones de respuesta (Martínez Arias et al., 2006). Los modelos más comunes son la Teoría Clásica de los Test (TCT) y la Teoría Respuesta al Ítem (TRI) (Montero Rojas, 2013). Ambas teorías plantean un modelo o distintos modelos y un conjunto de supuestos, que al cumplirse garantizan la precisión de la medida. Vale aclarar que, la TCT y la TRI son “teorías compatibles y se complementan en la práctica psicométrica para un análisis más profundo de la calidad y funcionamiento del test” (Attorresi et al., 2009, p.3).

2.5 Teoría Clásica de los Test

La TCT fue formulada casi en su totalidad bajo el fundamento de los trabajos pioneros de Spearman a principios del siglo XX. Esta teoría fue expuesta por primera vez de forma completa en la obra de Guilford en 1936 *Psychometric methods* (Navas Ara, 2001).

Posteriormente en 1950, Guilliksen sintetizó y amplió el enfoque clásico en su libro *Theory of Mental Test* (Muñiz, 2010). Una década más tarde, Gerorge Rasch publicó su famosa obra sobre el modelo logístico de un parámetro. En 1968 Lord y Novick reformularon y reanalizaron críticamente la TCT, abriendo nuevas perspectivas a las teorías de los test (Muñiz, 2018).

2.5.1 Características del modelo TCT

El modelo de la TCT parte de que las respuestas dadas a los ítems del test se combinan en una única puntuación total (X) para cada sujeto (Martínez Arias et al., 2006). Se asume que dicha puntuación (X) está conformada por dos componentes aditivos: la puntuación verdadera (V) y el error de medida (e) que inevitablemente va asociado a todo proceso de medición (Martínez Arias et al., 2006; Muñiz, 2010; 2018). Según la TCT la función que relaciona la puntuación observable en el test (X) y el nivel del sujeto en la variable inobservable (V), habilidad o rasgo medido es una función lineal (Navas Ara, 2001).

Tabla 1. Formulación del modelo TCT

Modelo	$X = V + e$
Supuestos	1. $E(E_i) = 0$
	2. $\sigma_{V_i E_i} = 0$
	3. $\sigma_{E_i E_j} = 0$
	4. $\sigma_{V_i E_j} = 0$

Nota. Fuente: Elaboración propia

Donde:

X = Puntuación empírica

V = Puntuación verdadera

e = error de medida

E = Esperanza matemática de los errores de medida

2.5.2 Supuestos del modelo TCT

La puntuación verdadera se calcula basandose en los siguientes supuestos (Martínez Arias et al., 2006; Muñiz, 2010; 2018):

- La esperanza matemática o valor esperado de los errores de medida es igual a cero, para una población determinada medida con el mismo test.
- Se asume que no existe relación entre la magnitud de la puntuación verdadera y el tamaño de los errores de medida de esas puntuaciones. Es decir, el valor de la puntuación verdadera no tiene que ver con el error que afecta a esa puntuación.
- Los errores de medida de las personas en un test no están relacionados con los errores de medida de otro test distinto.
- Los errores de medida de un test no están correlacionados con las puntuaciones verdaderas de otro test.

De estos supuestos, se formulan las siguientes deducciones inmediatas del modelo:

- El error de medida es la diferencia entre la puntuación empírica y la verdadera $e = X - V$.
- La media de las puntuaciones empíricas es igual a la media de las verdaderas $\mu_x = \mu_v$.
- Las puntuaciones verdaderas no covarían con los errores $\text{cov}(V, e) = 0$.
- La covarianza entre las puntuaciones empíricas y las verdaderas es igual a la varianza de las verdaderas $\text{cov}(X, V) = \text{var}(V)$.
- La covarianza entre las puntuaciones empíricas de dos test es igual a la covarianza entre las verdaderas $\text{cov}(X_j, X_k) = \text{cov}(V_j, V_k)$.
- La varianza de las puntuaciones empíricas es igual a la varianza de las verdaderas más la de los errores $\text{var}(X) = \text{var}(V) + \text{var}(e)$.
- La correlación entre las puntuaciones empíricas y los errores es igual al cociente entre la desviación típica de los errores y la de las empíricas $\rho(X, e) = \sigma_e / \sigma_x$ (Muñiz, 2018).

Los supuestos de la TCT son perfectamente razonables y sensatos, pero son débiles ya que se acomodan a la mayoría de situaciones prácticas de medición (Martínez Arias et al.,

2006). Además, no son comprobables empíricamente, lo único que se puede hacer es someter a prueba las últimas 3 deducciones de los supuestos citados previamente, para analizar si el modelo lineal es confirmado (Navas Ara, 2001).

La mayor ventaja de la teoría clásica es su simplicidad práctica y sus débiles supuestos teóricos que facilitan su cumplimiento y estimación de la puntuación verdadera (Matas, 2010). Sin embargo, el enfoque clásico presenta dos limitantes teóricas básicas:

- La medición de las variables no es independiente del instrumento utilizado, es decir, la puntuación del individuo depende del test utilizado (Güler et al., 2014).
- Los instrumentos de medida no son invariantes respecto de las personas evaluadas; es decir, las propiedades de los test (la dificultad y la discriminación) son dependientes del tipo de personas utilizadas para estimarlas y del número de ítems que compone el test (Muñiz, 2010; 2018).

2.5.3 Cálculo de parámetros desde la TCT

Los parámetros más importantes para evaluar la calidad métrica del ítem dentro de la TCT son la dificultad y la discriminación (Muñiz, 2018).

La dificultad de un ítem es, primeramente, una construcción teórica en base al contenido y complejidad de tareas demandadas para acertar el ítem (Navas Ara, 2001). Empíricamente, el índice de dificultad es la proporción de personas que aciertan el ítem de aquellas que han intentado resolverlo. Cabe mencionar, que este índice solamente es aplicable a pruebas cognitivas (Muñiz, 2018). La forma más común para calcular la dificultad es mediante el índice p , el cual se expresa matemáticamente de la siguiente forma:

$$p = \frac{A}{N}$$

Donde:

A = Número de personas que aciertan el ítem.

N = Número de personas que responden el ítem.

Este parámetro oscila entre un valor mínimo de 0 y máximo de 1. Ítems con valores menores que 0.40 se consideran como difíciles, entre 0.40-0.50 medianamente difíciles, entre 0.51-0.80 dificultad media, entre 0.81-0.90 medianamente fáciles y entre 0.91-1.0 fáciles

(Ortiz Romero et al., 2015). En otras palabras, valores cercanos a cero, indican que el ítem es muy difícil debido a que ninguno o pocos sujetos responden correctamente. Valores cercanos a 1 indican que el ítem es muy fácil, ya que muchos han respondido correctamente. Ítems con valores extremos ($p=0$, $p=1$) deben ser eliminados ya que no contribuyen a la medir las diferencias entre los sujetos (Navas Ara, 2001).

Para evitar que sujetos con poca o nula aptitud respondan al azar, se puede corregir el efecto azar en el cálculo de la dificultad de ítems de elección múltiple. Esta corrección se utiliza cuando se percibe que un considerable número de individuos responden por pura adivinación (Navas Ara, 2001).

El ajuste del índice de dificultad se realiza de la siguiente manera:

$$p_c = p - p_{azar} = \frac{A}{N} - \frac{\frac{E}{k-1}}{N} = \frac{A - \frac{E}{k-1}}{N}$$

Donde:

p_c = dificultad del ítem corregida

p = dificultad del ítem sin corregir

p_{azar} = proporción de respuestas correctas al azar

E = número de respuestas incorrectas

K = número de alternativas del ítem

Una limitación de este índice es que su valor depende del tipo de personas a las que se aplique. Si son muy competentes, resultará un ítem fácil y como resultado muchos lo acertarán. En cambio, si son incompetentes, el mismo ítem podría resultar difícil (Muñiz, 2018).

En cuanto a la discriminación del ítem, este parámetro permite diferenciar adecuadamente a los sujetos con diferentes niveles de habilidad o rasgo que evalúa el test. Este parámetro es utilizado en pruebas cognitivas y no cognitivas. Existen dos maneras para calcular la discriminación:

- Índice de discriminación D : se basa en la comparación de las respuestas al ítem de dos grupos extremos de sujetos, conforme a su puntuación alta o baja en el test. Un ítem tendrá un buen índice de discriminación si las personas que aciertan el

ítem son en su mayoría los mismos individuos que tienen una alta puntuación total (Navas Ara, 2001).

$$D = \frac{A_{alto} - A_{bajo}}{N_g}$$

Donde:

D = índice de discriminación

A_{alto} = Grupo con alta puntuación

A_{bajo} = Grupo con baja puntuación

N_g = número de sujetos en el grupo con mayor número de personas.

El ítem presentará un alto poder discriminativo si D es mayor o igual a 0.40, una discriminación aceptable si $0.30 \leq D \leq 0.39$, y una discriminación pobre si $0.20 \leq D \leq 0.29$, el ítem no es adecuado y debe ser eliminado si es $0.0 < D \leq 0.19$ (Navas Ara, 2001).

- Índice basado en la correlación ítem-test: se basa en la asociación entre las respuestas al ítem y las puntuaciones en el test. Supone que los individuos con alta puntuación en el test contestan correctamente los ítems (Navas Ara, 2001). Este índice se puede medir mediante correlación biserial-puntual, correlación biserial, coeficiente phi y correlación tetracórica (Muñiz, 2018). El tipo de correlación utilizada dependerá de la forma de medición de las preguntas y la prueba (si ambas variables son dicotómicas, dicotomizadas, continuas o una combinación de ellas) (Hurtado, 2018).

La correlación biserial-puntual, es una aplicación del coeficiente de correlación de Pearson cuando los ítems son dicotómicos y la puntuación del test es cuantitativa continua. Es importante señalar que al realizar el cálculo se debe descontar el ítem objeto de estudio para evitar correlaciones espurias (Muñiz, 2018). Los valores del índice de discriminación pueden oscilar entre -1 y $+1$. Cuanto mayor sea esta correlación, mayor es la discriminación (Hurtado, 2018). La ventaja de este cálculo es que considera a todos los sujetos de la muestra, a diferencia del índice D (Navas Ara, 2001).

$$\rho_{bp} = \frac{\mu_p - \mu_x}{\sigma_x} \sqrt{\frac{p}{q}}$$

Donde:

μ_p = Media en el test de las personas que aciertan el ítem.

μ_x = Media del test.

σ_x = Desviación típica del test.

p = Proporción de personas que aciertan el ítem.

$q = (1 - p)$.

La discriminación y la dificultad de un ítem se relacionan directamente, siendo que cuando un ítem tiene valores extremos de dificultad, que son respondidos de igual forma por todos o la mayoría de los evaluados, evidencia que el ítem tiene poco o nulo poder discriminativo. Por otra parte, ítems con dificultad intermedia ($p=.50$), muestran mayor poder discriminativo (Martínez Arias et al., 2006).

2.6 Teoría de Respuesta al Ítem

La TRI emerge de estudios pioneros de Thurstone en la segunda década del siglo XX (Muñiz, 2010). Sin embargo, el nacimiento de la TRI se atribuye a los trabajos de Frederick M. Lord a mitad del siglo XX. Particularmente, su trabajo de tesis doctoral *A theory of test scores* en 1953 dirigida por Gulliksen, dio el fundamento robusto a la TRI. De igual forma, Birnbaum en 1957 propuso la aplicación de modelos logísticos para la evaluación psicológica. Seguidamente, en 1960 Georg Rasch publicó su libro sobre el modelo logístico de un parámetro. Más adelante en 1968 Lord y Novick publican un libro en que presentan aplicaciones prácticas a los modelos de la TRI, dando mayor fundamento a la teoría (Muñiz, 2018).

Sin embargo, debido a que los modelos de la TRI implicaban complejos procesos matemáticos, fue hasta el desarrollo de software apropiado, en los años 80 (como BICAL, LOGISTIC, BILOG, MULTILOG, MICROCAT) que comenzó su aplicabilidad (Martínez Arias et al., 2006). Para finales del siglo XX, los modelos basados en TRI experimentaron una acelerada expansión y consolidación. Finalmente, se resalta el libro publicado por Frederic Lord en 1980 *Aplicaciones de la Teoría de Respuesta a los Ítems a problemas prácticos del uso de los test* en el que se sintetizan los desarrollos teóricos y modelos de la

TRI (Muñiz, 2018). En la actualidad, la TRI se ha popularizado y ha tomado gran preponderancia en el mundo de la medición psicológica y educativa.

2.6.1 Características de la TRI

La TRI resuelve importantes problemas en la medición psicológica que no encontraban solución en la Teoría Clásica (Muñiz, 2010; 2018). Su gran contribución y objetivo principal es obtener mediciones invariantes respecto de las personas implicadas y más allá de los ítems o instrumentos utilizados (Attorresi et al., 2009; Muñiz, 2018).

La razón de llamarse TRI es porque su enfoque es en el análisis del ítem individual y sus propiedades, más que en el test (Muñiz, 2018). La TRI busca establecer una relación entre el comportamiento de un sujeto al contestar un ítem y el rasgo latente o habilidad subyacente a esa conducta (Attorresi et al., 2009). Los principios básicos de esta teoría son:

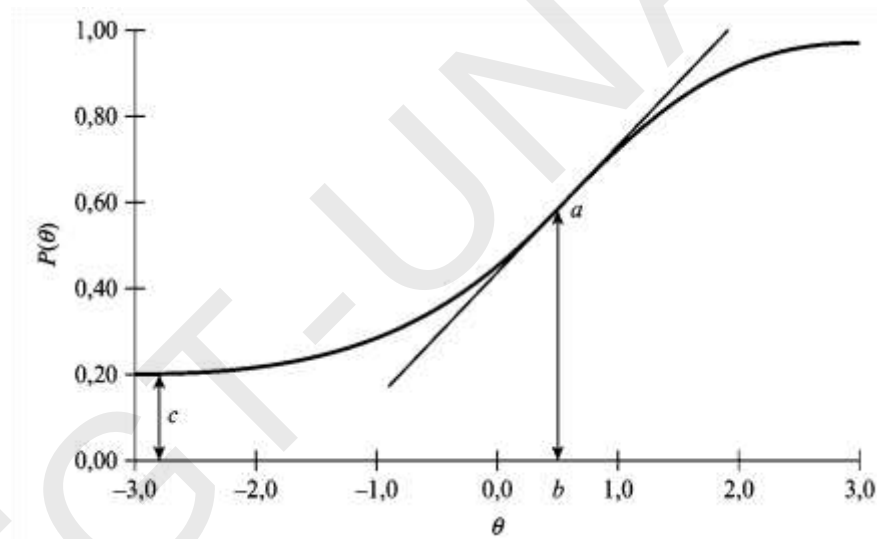
- El patrón de respuestas en un conjunto de ítems en un test es explicado por un rasgo latente, simbolizado por θ . El cual es estimado a partir del patrón de respuestas del conjunto de ítems (Attorresi et al., 2009).
- Existe una relación funcional entre los valores de la variable que miden los ítems y la probabilidad de acertar estos (Hidalgo-Montesinos y French, 2016).
- La puntuación verdadera (el nivel de aptitud, rasgo, habilidad, competencia, etc.), es independiente del ítem y del test utilizado para estimarlo. Por lo tanto, la respuesta del sujeto depende únicamente del rasgo latente (Matas, 2010).
- Invarianza de los parámetros de los ítems respecto a la muestra que se calcula. Es decir, que los parámetros del ítem no cambian, aunque las personas que contesten sean distintas (Matas, 2010).

Desde la TRI se postula que para cada nivel de rasgo existe una probabilidad asociada de contestar correctamente al ítem. Esta probabilidad es pequeña para sujetos con bajo nivel de rasgo, y alta para sujetos con altos niveles de rasgo (Matas, 2010). La relación entre la respuesta de un sujeto a un ítem y el rasgo latente puede describirse como una función monótona creciente, expresada en términos probabilísticos mediante distintos modelos matemáticos (Matas, 2010; Güler et al., 2014). Particularmente, se basa en análisis de regresión de datos multivariante que enlazan una variable respuesta con otra no observada. Normalmente, este análisis es calculado mediante funciones de probabilidad normal

acumulada y logística, y es representado gráficamente mediante la Curva Característica del Ítem (CCI) (Attorresi et al., 2009; Matas, 2010).

En la Figura 1 se observa la relación entre el rasgo latente que está siendo evaluado y la actuación observada en un ítem, indicando la probabilidad que tiene el individuo de responder correctamente al ítem según el nivel del rasgo que posee (Hidalgo-Montesinos y French, 2016). Para su representación se utiliza la ojiva logística creciente, esta curva tiene dos asíntotas una superior y otra inferior, con una inflexión en el centro de la curva, que es el punto de máxima pendiente (Navas Ara, 2001). La CCI se suele representar en una escala situada entre -3 y $+3$ puntos de habilidad (Matas, 2010).

Figura 1. Curva Característica de un ítem



Nota. Fuente: Muñiz (2018, p. 190)

El comportamiento de un individuo en un ítem es en función no solo de su nivel de aptitud, sino también de las características específicas del ítem (Martínez Arias et al., 2006). Por tanto, la forma de la CCI puede ser determinada por los siguientes parámetros del ítem:

- Localización o dificultad del ítem (parámetro b): Representa el nivel de dificultad del ítem y se define como el nivel de rasgo para el cual un individuo tiene una probabilidad del 50% de responder correctamente al ítem. Este describe dónde está situado el ítem en la escala de aptitud, por lo que se expresa en la misma

escala que el rasgo latente, correspondiente al punto en el eje x de máxima pendiente de la CCI. La escala adoptada para medir θ tiene un rango teórico de $-\infty$ a $+\infty$, con media 0 y desvío típico 1. Cuanto mayor es el valor de b más difícil es el ítem (Muñiz, 2018; Navas Ara, 2001).

- Discriminación del ítem (parámetro a): Mide la potencia discriminatoria para identificar los sujetos en función del nivel en el rasgo latente que poseen. Este es proporcional a la pendiente de la recta tangente a la CCI en el punto de máxima pendiente de esta. Tiene un rango teórico entre $-\infty$ a $+\infty$, aunque normalmente es comprendida entre -2.8 y 2.8. Cuanto mayor sea la pendiente, mayor será el poder discriminativo del ítem (Muñiz, 2018).
- Pseudoazar (parámetro c): Representa la probabilidad de que un sujeto responda correctamente, poseyendo un nivel muy bajo de rasgo o acertándolo por azar. Es definido como el valor de la asíntota inferior de la curva característica del ítem (Attorresi et al., 2009; Hidalgo-Montesinos y French, 2016).

2.6.2 Supuestos de la TRI

- Unidimensionalidad: Los ítems de un test están constituidos en función de un único rasgo (Muñiz, 2010). Este rasgo es suficiente para explicar los resultados de los sujetos (Martínez Arias et al., 2006).
- Independencia Local: La respuesta a un ítem no está condicionado a la respuesta de otro ítem cuando se considera un mismo nivel en el rasgo. Es decir, que las respuestas a distintos ítems son estadísticamente independientes (Attorresi et al., 2009; Matas, 2010).

En cuanto a estos dos últimos supuestos, se argumenta que ambos están íntimamente relacionados, ya que si se cumple la unidimensionalidad, se cumple el supuesto de independencia local (Muñiz, 2010). Esto es, si se confirma que dos ítems no son independientes significa que otro factor ajeno al que se pretende medir incide en la estimación del nivel del rasgo violando el supuesto de unidimensionalidad (Attorresi et al., 2009).

2.6.3 Modelos de la TRI

El análisis usando TRI proporciona una visión de la relación entre el nivel en el rasgo de un individuo y las características de los ítems. Tienen el objetivo de predecir la probabilidad de respuesta a un ítem basándose en diferentes parámetros de los ítems y el nivel de rasgo latente. La TRI proporciona una variedad de modelos que permite trabajar con test tanto unidimensionales como multidimensionales y con distintos formatos de respuesta (Hidalgo-Montesinos y French, 2016).

Como ya se ha mencionado, los modelos de TRI asumen que existe una relación funcional entre los valores de la variable que miden los ítems y la probabilidad de acertar a estos, dicha función se expresa mediante la CCI (Muñiz, 2018). Los dos tipos funciones más utilizadas dentro del marco de la TRI son la función logística y la curva normal acumulada. La función logística tiene muchas ventajas sobre la curva normal, principalmente por su mayor tratabilidad matemática (Muñiz, 2010). La función logística es una curva cuya fórmula es la siguiente:

$$P_i(\theta) = c_i + (Y_i - c_i) \frac{e^{Da_i(\theta - b_i)}}{1 + e^{Da_i(\theta - b_i)}}$$

Donde:

$P_i(\theta)$ = Probabilidad de acertar el ítem i a determinado nivel θ

θ = Valores de la variable medida

b_i = Índice de dificultad del ítem i

a_i = Índice de discriminación del ítem i

c_i = Índice de pseudoazar del ítem i

e = Base de los logaritmos neperianos (2.72)

D = Constante ($D = 1.7$).

El modelo logístico puede clasificarse en función del número de parámetros: modelo de un parámetro, dos parámetros y tres parámetros (Matas, 2010). Es importante señalar que los modelos logísticos asumen que la respuesta a los ítems es codificada dicotómicamente (1= acierto y 0= error), aunque involucre varias opciones de respuesta (Muñiz, 2018).

El modelo logístico de un parámetro

El modelo logístico de un parámetro o modelo de Rasch, fue formulado por Georg Rasch en 1960, de donde se le adjudica su nombre (Muñiz, 2018). Según este modelo, la respuesta a un ítem viene determinado solamente por el nivel del rasgo θ y la dificultad del ítem (parámetro b). En el modelo de Rasch la CCI adquiere la siguiente expresión:

$$P_i(\theta) = \frac{e^{D(\theta-b_i)}}{1+e^{D(\theta-b_i)}}$$

Donde:

$P_i(\theta)$ = Probabilidad de acertar el ítem i a determinado nivel θ

θ = Valores de la variable medida

b_i = Índice de dificultad del ítem i

e = Base de los logaritmos neperianos

D = Constante (cuando $D = 1.7$, los valores de la función logística apenas difieren de los de la curva normal acumulada).

Este modelo se basa en que, conocido el índice de dificultad de un ítem b , y el nivel de rasgo del individuo θ , es posible predecir la probabilidad $P(\theta)$ de acertar el ítem. El índice de dificultad b es aquel valor de θ para el cual $P(\theta) = 0.5$, en otras palabras, es el punto de inflexión de la curva (Muñiz, 2018).

El modelo de dos parámetros

El modelo de dos parámetros fue originalmente desarrollado por Birnbaum a mitad del siglo XX. Este modelo asume que la CCI es determinada por dos parámetros, el índice de dificultad b y el índice de discriminación a (Muñiz, 2018).

$$P_i(\theta) = \frac{e^{Da_i(\theta-b_i)}}{1+e^{Da_i(\theta-b_i)}}$$

Donde:

$P_i(\theta)$ = Probabilidad de acertar el ítem i a determinado nivel θ

θ = Valores de la variable medida

b_i = Índice de dificultad del ítem i
 a_i = Índice de discriminación del ítem i
 e = Base de los logaritmos neperianos
 D = Constante ($D = 1.7$)

El modelo de dos parámetros puede utilizarse si la adivinación no es posible o no se puede asumir (Hidalgo-Montesinos y French, 2016).

El modelo logístico de tres parámetros

El modelo de tres parámetros tiene su origen con Birnbaum entre los años de 1957 a 1968. El modelo asume que la CCI es determinada por tres parámetros: dificultad, discriminación y pseudoazar. Este último se define como la probabilidad de acertar el ítem al azar cuando no se conoce la respuesta $P_i(\theta) = -\infty$ (Muñiz, 2018). El modelo se expresa de la siguiente forma:

$$P_i(\theta) = c_i + \frac{1 - c_i}{1 + e^{D a_i (\theta - b_i)}}$$

Donde:

$P_i(\theta)$ = Probabilidad de acertar el ítem i a determinado nivel θ

θ = Valores de la variable medida

c_i = Índice de pseudo -azar del ítem i

b_i = Índice de dificultad del ítem i

a_i = Índice de discriminación del ítem i

e = Base de los logaritmos neperianos

D = Constante ($D = 1.7$)

Evidentemente, al aumentar la probabilidad de acertar por azar, la probabilidad de respuestas correctas aumenta (Muñiz, 2018).

Por último, la TRI permite evaluar hasta qué punto el modelo alcanzado representa bien a los datos observados. La evaluación del ajuste se basa en análisis de la discrepancia entre las CCIs teóricas, dadas por el modelo y las CCIs empíricas (Attorresi et al., 2009). Para evaluar el ajuste de cada ítem al modelo se puede utilizar el estadístico χ^2 , que evalúa

el grado de similitud entre las frecuencias de respuestas del modelo predictivo y las del modelo empírico por categoría de respuesta al ítem. Valores estadísticamente significativos indicarían que existen discrepancias entre el modelo predictivo y el observado, y por lo tanto que un ítem determinado no se ajusta al modelo (Toland, 2014).

2.7 Análisis de opciones de respuesta

Cada ítem está compuesto por varias opciones de respuesta. Una de ellas es la correcta y las otras son incorrectas, también conocidas como distractores porque distraen a los estudiantes con conocimiento parcial debido a su plausibilidad para acertar (Gierl et al., 2017). El análisis de estas opciones de respuesta permite dar a conocer la utilidad y eficacia de cada alternativa en el ítem. Existen dos tipos de análisis: El análisis tradicional basado en la Teoría Clásica o en la Teoría de Respuesta al Ítem y análisis contemporáneo que incluye análisis de diagnóstico cognitivo. En este apartado, se expondrá brevemente los conceptos y el proceso de análisis de las opciones de respuesta desde el análisis tradicional del test.

El propósito principal del análisis es eliminar los distractores no funcionales y mejorar el poder discriminador del ítem para distinguir los evaluados con habilidad alta de los de baja habilidad. Este análisis, consiste observar las frecuencias de elección de cada una de las opciones por parte de los participantes, tanto la correcta como los distractores. El objetivo de un distractor es atraer la atención de los sujetos con nivel medio o bajo en el constructo, que no conocen lo suficiente para responder correctamente al ítem (Martínez Arias et al., 2006).

Navas (2001) presenta una serie de condicionantes que garantizan la eficacia de los distractores:

1. Son elegidos por un número suficiente de sujetos. Sin embargo, si el distractor no es elegido por nadie o si es seleccionado en una proporción mayor a la respuesta correcta, esta opción está confundiendo a los evaluados y se recomienda reestructurar el ítem.
2. Son aproximadamente igual de atractivas para los sujetos. Un buen distractor desvía la atención de la respuesta correcta al presentarse como una posible alternativa si no se conoce cuál es la respuesta correcta.

3. El rendimiento medio en el test de los sujetos que escogen cada distractor debe ser inferior al de los sujetos que han elegido la respuesta correcta y a la media del test general con todos los sujetos.

Resulta oportuno detallar los análisis puntuales que se realizan, entre ellos está el análisis de la proporción de sujetos que elige cada opción, para el conjunto total de los sujetos. Este consiste en un análisis de la proporción del test para los que han elegido cada opción según su nivel de habilidad y se observa el poder discriminativo de los distractores.

En cuanto al análisis de la proporción de selección, todo distractor debe ser seleccionado por alrededor la misma cantidad de personas. Cuando un distractor es seleccionado en una baja o nula proporción, éste no está cumpliendo su función adecuadamente (Digeduca, 2015). Un distractor no es funcional cuando menos del 5% elige esa opción (Gierl et al., 2017). Todos los distractores deberían ser igualmente aceptables para los examinados que no conocen la respuesta correcta de un reactivo (Aiken, 2003).

Para estudiar el poder discriminativo de los distractores se puede calcular el índice D o un índice basado en la correlación distractor-test. El índice de discriminación D, está basado en la información resultante de la comparación de la selección de opciones de respuestas de ítem de dos grupos extremos (baja y alta) según el nivel de aptitud. Es decir, cuanto mayor cantidad de sujetos del grupo alto seleccionan la opción correcta, mayor será el poder discriminativo de la opción correcta. De esta manera, la opción correcta presentará gran poder discriminativo cuando $D \geq 0.40$, la discriminación será aceptable si $0.30 \leq D \leq 0.39$, la discriminación será poca si $0.20 \leq D \leq 0.29$, la discriminación será inadecuada y deberá ser modificado o eliminado si $0.0 < D \leq 0.19$, finalmente el ítem debe eliminarse si $D \leq 0.0$ (Navas Ara, 2001).

Por otro lado, se considera que un buen distractor debe presentar un poder discriminativo más bajo que el de la respuesta correcta y, preferiblemente, con un valor negativo. Por lo tanto, se deben revisar minuciosamente los ítems con distractores que presentan valores altos y positivos en el poder discriminativo (Navas Ara, 2001).

El análisis de distractores puede representarse mediante una gráfica de las frecuencias de elección de opciones en diferentes niveles de aptitud, permite visualizar las

distribuciones del distractor en los diferentes niveles de habilidad permite identificar los ítems que no discriminan bien, y por lo tanto no son funcionales (Martínez Arias et al., 2006).

2.8 Análisis de dimensionalidad

El estudio de dimensionalidad consiste en analizar la configuración interna del test, evaluando el nivel en que los ítems encajan con el constructo medido en el test (Navas Ara, 2001). Un supuesto importante desde la TRI es que los test son unidimensionales (aunque en la actualidad existen modelos multidimensionales), es decir, que un único rasgo latente es responsable explicar las respuestas de un sujeto ante un conjunto de ítems. Este supuesto es muy importante para obtener buenas mediciones y obtener evidencias de validez de estructura interna del test (Martínez Arias et al., 2006).

Sin embargo, en la práctica, la unidimensional no podrá cumplirse totalmente, ya que la medición psicológica siempre involucra un conjunto de elementos interrelacionados que influyen en el comportamiento de respuesta a un conjunto de ítems (Burga León, 2006). No obstante, el supuesto se considera satisfecho cuando un solo factor dominante es capaz de explicar las respuestas del test y las relaciones entre ítems (Auné et al., 2019).

El análisis factorial es la herramienta más utilizada para evaluar la dimensionalidad de un test. Matemáticamente está basado en los principios de la regresión lineal (Ferrando y Anguiano-Carrasco, 2010). Éste examina las relaciones entre indicadores observables o ítems y variables latentes o inobservables denominadas factores, en concordancia con el constructo teórico (Rios y Wells, 2014).

El análisis factorial tiene sus inicios en los trabajos de Galton en 1889, quien planteó la idea de un rasgo latente. Galton propuso que, si dos variables están correlacionadas entre sí, es debido a que poseen un rasgo latente en común, que explica parte de la varianza de las variables. En 1901 Pearson desarrolló el coeficiente de correlación y aportó principios fundamentales al análisis. Sin embargo, el creador por excelencia del análisis factorial fue Spearman en 1904. Él estudió las correlaciones en diferentes pruebas de aptitudes, las cuales reflejaban un factor subyacente de inteligencia. Más adelante, a mitad de siglo XX, Thurstone consolidó las bases metodológicas y reformuló el análisis (Pérez y Medrano, 2010).

El análisis factorial se divide en el Análisis Factorial Exploratorio (AFE) y el Análisis Factorial Confirmatorio (AFC) (Burga León, 2006). El análisis factorial exploratorio es un método estadístico que encuentra patrones de asociación entre ítems y variables para establecer una estructura de dimensiones en base a las correlaciones entre las variables observadas (Martínez Arias et al., 2006; Rojas Torres, 2014). Este se utiliza principalmente en las etapas iniciales de desarrollo de test ya que no se imponen restricciones al modelo (Ledesma et al., 2019). En cambio, el análisis factorial confirmatorio comprueba el ajuste de los ítems conforme a un modelo teórico previamente establecido (Rojas Torres, 2014). En este, las restricciones son más rigurosas, definiendo el número de factores y su relación entre ellas (Ferrando y Anguiano-Carrasco, 2010).

2.8.1 Análisis factorial exploratorio

El análisis factorial exploratorio es un conjunto de métodos multivariados de interdependencia, cuya finalidad es identificar una estructura de variables subyacentes inobservables a un conjunto amplio de datos observables (Pérez y Medrano, 2010).

Este modelo de análisis factorial normalmente se utiliza para explorar la estructura interna y las dimensiones subyacentes de los ítems que conforman un test. También se usa para reducir gran número de ítems a un número inferior de variables conceptuales. Además, se utiliza para seleccionar y otorgar significado teórico a un conjunto inicial de ítems de un test (Pérez y Medrano, 2010).

Mediante el AFE se busca determinar la dimensión o factores responsables de la variabilidad en las puntuaciones de un test en línea con su significación teórica (Pérez y Medrano, 2010). En esta línea, la varianza total de una variable se debe a: factores compartidos con otras variables (comunalidad), factores específicos de la variable (especificidad) y varianza residual. El análisis factorial busca reproducir únicamente la varianza común del ítem, no la varianza residual, para así agrupar los ítems conforme a las relaciones encontradas (Ferrando y Joan, 1996).

2.8.2 Supuestos del análisis factorial

El análisis factorial trabaja en base a ciertos supuestos que al seguirlos se podrá obtener mediciones significativas:

- Normalidad: los datos siguen una distribución normal.
- Linealidad: la relación entre las variables debe ser lineal.
- Multicolinealidad: Los ítems no deben tener correlaciones elevadas o redundantes (Pérez y Medrano, 2010).

2.8.3 Fases del análisis factorial

El análisis factorial exploratorio consiste en cuatro fases principales 1. Preparación inicial de datos, 2. extracción de factores, 3. rotación de factores e 4. interpretación de factores.

1. Preparación inicial de datos:

Es preciso realizar un análisis previo para depuración de errores, verificación de existencia de puntuaciones marginales, comprobación de supuestos y determinación de la matriz de correlaciones a utilizar (Martínez Arias et al., 2006).

El análisis puede verse afectado por la existencia de outliers o puntuaciones marginales, ya que el AFE se basa en las correlaciones entre variables, las cuales son estimadas mediante la media, que es una medida de tendencia central muy sensible a los valores extremos. Por tal razón, conviene en primer lugar realizar un análisis inicial de datos para detectar dichos valores extremos. Existen diferentes formas de realizar el análisis. Comúnmente, se calculan las puntuaciones típicas de cada variable y se identifican las puntuaciones que se encuentren fuera del rango ± 3 . También se puede visualizar los gráficos de caja y bigotes, en los cuales se muestran claramente los valores atípicos como puntos aislados en los extremos de los bigotes. Otra alternativa es utilizar el procedimiento de distancia Mahalanobis (D^2) para identificar los casos atípicos que superen la significación de $p < .001$ (Pérez y Medrano, 2010).

Otro análisis previo esencial es evaluar si los ítems están suficientemente interrelacionados para ser analizados por medio del AFE. Para esto, se puede realizar algunas

pruebas estadísticas llamadas medidas de adecuación muestral, como el test de esfericidad de Bartlett, el cual pone a prueba la hipótesis nula de que las variables no están correlacionadas en la población. Si los resultados son significativos ($p < .05$) se rechaza la hipótesis nula y se considera que las variables están suficientemente correlacionadas para realizar el AFE (Ferrando y Anguiano-Carrasco, 2010). También se puede realizar la medida de adecuación muestral de Kaiser-Meyer-Olkin (KMO), este evalúa hasta qué punto las puntuaciones en cada una de las variables son predecibles desde las demás. El rango de valores es de 0 a 1, si $KMO > .90$ el test es muy bueno, es notable si $KMO > 0.80$ y medianamente aceptable si $KMO > 0.70$ valores menores que 0.60 sugerirán interrelaciones bajas entre los ítems (Pérez y Medrano, 2010).

Habiendo realizado estos análisis iniciales, se procede a verificar los supuestos de normalidad, linealidad y multicolinealidad. Usualmente se utiliza pruebas de contraste de bondad de ajuste como Shapiro-Wilk y Kolmogorov-Smirnov para evaluar la distribución normal. Sin embargo, también es recomendable estimar los índices de asimetría y curtosis, considerando que los valores dentro del rango ± 1.5 indican variaciones leves de la normal y, por lo tanto, los datos son adecuados para realizar el AFE. También se puede visualizar los gráficos Q-q plot, para determinar si los datos se ajustan a una distribución normal (Pérez y Medrano, 2010).

El supuesto de linealidad es evaluado mediante diagramas matriciales de dispersión. Los puntos deben organizarse a lo largo de una línea recta para que exista linealidad. Además, se puede realizar una Estimación Curvilínea por medio de análisis de regresión múltiple (Pérez y Medrano, 2010).

Por último, el supuesto de colinealidad se analiza por medio de las correlaciones entre variables, si estas son superiores a 0.90 existe multicolinealidad y terminará afectando la estabilidad de la solución factorial. También puede estimarse los índices de tolerancia y la inflación de la varianza (VIF). Valores pequeños de tolerancia (menores que 0.10) y elevados de VIF (mayores que 10) demuestran alta colinealidad (Pérez y Medrano, 2010).

Ahora bien, se debe decidir el tipo de matriz de correlaciones a utilizar, esta decisión debe ir en concordancia con el tipo de datos. Usualmente se utiliza la matriz de correlaciones de Pearson, sin embargo, surge un problema al analizar ítems dicotómicos, ya

que este tipo de correlación asume que las variables están medidas a nivel de intervalo y tienen distribución normal. Para una estimación más precisa, la solución propuesta es utilizar una matriz de correlaciones tetracóricas, estas son estimaciones máximo-verosímil de la correlación de Pearson que había en las variables antes de ser dicotomizadas, asumiendo que tras esta dicotomía se encuentran dos variables distribuidas normalmente (Burga León, 2006; Ferrando y Joan, 1996).

2. Extracción de factores: En esta fase se determina el número de factores y el método de extracción.

Tras haberse verificado los supuestos, se procede a seleccionar un método de extracción de factores para estimar una solución factorial inicial. Los más utilizados en el AFE son el de Componentes Principales, Ejes Principales y Máxima Verosimilitud (Mavrou, 2015; Pérez y Medrano, 2010).

El método de Componentes Principales explica la mayor cantidad de varianza en los datos observados (Mavrou, 2015). Este es un método de extracción de factores, pero en realidad, no es un procedimiento para estimar un modelo factorial, ya que funciona en base a la varianza total, por lo que no siempre produce los mismos resultados que un modelo factorial (que sí distingue entre la varianza común y la varianza de error). Por tal razón no es un método apropiado a considerar, sin embargo, es mencionado aquí debido a que existe gran confusión entre el análisis factorial y componentes principales y muchas veces es utilizado injustificadamente llevando a mediciones erróneas (Pérez y Medrano, 2010).

Otros métodos de estimación son los basados en Mínimos Cuadrados Ordinarios y Máxima Verosimilitud. Los estimadores por Mínimos Cuadrados buscan determinar una solución factorial que minimice la suma de cuadrados de las diferencias entre las correlaciones observadas y las reproducidas desde el modelo, logrando que los residuales sean lo más cercanos a 0 posible (Ferrando y Anguiano-Carrasco, 2010). Un método relevante dentro de este criterio es el de Ejes Principales, el cual es un método iterativo basado en la extracción de comunales de la varianza, este suele ser recomendado cuando los datos violan el supuesto de normalidad. Tiene la ventaja de ofrecer soluciones estables y recuperar valores de correlaciones débiles en muestras pequeñas. Sin embargo, no proporciona índices de ajuste ni intervalos de confianza (Pérez y Medrano, 2010).

También existe otro método de Máxima Verosimilitud, que permite estimar la significación estadística de los pesos factoriales y su ajuste al modelo (Ferrando y Anguiano-Carrasco, 2010). Sin embargo, este es un método estadístico inferencial que presupone que las variables son continuas y distribuidas con normalidad, en ausencia de casos atípicos (Mavrou, 2015), por lo cual es una buena elección cuando los datos se distribuyen con normalidad (Pérez y Medrano, 2010). En contraste, el incumplimiento de los supuestos conllevará a resultados problemáticos y no fiables.

El siguiente aspecto para considerar es el número de factores a extraer, lo cual debe fundamentarse tanto en la teoría como en la evidencia empírica (Mavrou, 2015). En la literatura científica no existe un consenso claro sobre un criterio para verificar el supuesto de unidimensionalidad. El criterio más utilizado es la regla de Gutman-Kaiser, en el cual se debe conservar aquellos factores con valores mayores que 1 (Martínez Arias et al., 2006). Según el criterio de Gutman-Kaiser, para que el test sea aceptado como unidimensional debe existir un solo autovalor superior a uno (Auné et al., 2019). Esta regla implica que un factor para que sea útil debería explicar tanta varianza como una variable individual (Pérez y Medrano, 2010). Sin embargo, presenta algunos problemas ya que se basa en los valores propios de la matriz sin reducir que son equivalentes a la proporción de la varianza total, cuando lo que realmente interesa es la varianza común (Ferrando y Anguiano-Carrasco, 2010).

Un criterio clave para la extracción de factores analizar es el porcentaje de varianza explicada por la solución factorial obtenida (Pérez y Medrano, 2010). Un test se considera unidimensional si explica una buena proporción de la varianza total y si los sucesivos factores explican en mucho menor cantidad entonces se considera unidimensional. Existen diversas opiniones sobre la cantidad de varianza explicada del primer factor. En general, el primer factor debería explicar entre el 30 a 40%, de la varianza (Burga León, 2006; Ferrando y Joan, 1996).

Además, se utiliza el test de sedimentación, un gráfico grafico bivariado de los autovalores de la matriz de correlación original sin rotar (Pérez y Medrano, 2010). Cuando entre el primer autovalor se identifica un salto brusco en la pendiente y los restantes comienzan a generar una pendiente de poca inclinación se acepta la unidimensionalidad (Auné et al., 2019). Sin embargo, este método presenta dos problemas, el primero es que se

apoya en la subjetividad de la inspección visual y aunado a esto, se basa en la varianza total, sin distinguir la varianza común del error. Una solución es utilizar las comunalidades en lugar de los valores propios de la matriz sin reducir (Ferrando y Anguiano-Carrasco, 2010).

También puede utilizarse un análisis paralelo, en el cual se comparan los autovalores en los valores propios de la matriz empírica con los autovalores simulados (Pérez y Medrano, 2010). Gráficamente es como un doble test de sedimentación, para afirmar unidimensionalidad el punto de intersección entre ambas líneas se deberá encontrar en el primer autovalor (Ferrando y Anguiano-Carrasco, 2010).

Alternativamente, Ferrando y Anguiano-Carrasco (2010) proponen evaluar las correlaciones residuales como un método más acertado para determinar el número de factores. Si la distribución de frecuencias residuales es normal y centrada en torno a una media de cero indicará que el número de factores es adecuado.

Existen otras evidencias para evaluar la unidimensionalidad, mediante el Análisis Factorial Confirmatorio, en el cual se impone una estructura unifactorial. También mediante el Test Estadístico de Unidimensionalidad Esencial de Stout, sin embargo, solo puede ser aplicado en muestras de por lo menos 500 participantes (Auné et al., 2019).

3. Rotación de factores: En esta fase se elige el procedimiento de rotación de los factores.

El resultado inicial del análisis factorial requiere un procedimiento de rotación para facilitar su interpretación. Sin embargo, para test unidimensionales la rotación no es tan imprescindible, ya que se puede interpretar directamente (Ferrando y Anguiano-Carrasco, 2010). Existen varios métodos de rotación de factores, comúnmente se clasifican en rotaciones ortogonales (varimax, quartimax y equimax) u oblicuas (oblimin direct, promax, orthoblique, procrustes) (Mavrou, 2015). Los métodos más comunes son Varimax y Promax. Las rotaciones oblicuas como la Promax son más apropiadas para variables psicológicas, ya que los factores normalmente están correlacionados, estas permiten obtener soluciones precisas, simples y realistas. Las rotaciones ortogonales son consideradas más simples y fáciles de interpretar, pero asumen que los factores no están correlacionados, normalmente se utilizan cuando las correlaciones son menores que 0.32 (Pérez y Medrano, 2010).

4. Interpretación de factores: La tarea final del análisis es interpretar y denominar los factores. Para esto, se analiza en la matriz rotada, los patrones de correlaciones altas y bajas en los distintos factores posibles (Pérez y Medrano, 2010).

Se asume que los ítems pertenecientes a una misma dimensión teórica deben presentar altas correlaciones entre sí (Rojas Torres, 2014). En las Ciencias Sociales, particularmente en la medición de constructos psicológicos las saturaciones generalmente son moderadas (menores que 0.32), por tanto, saturaciones por encima de .50 son consideradas saturaciones fuertes. Se recomienda que la dimensión debe poseer al menos 4 ítems con correlaciones superiores a 0.40 (Mavrou, 2015). En la aplicación del AFE, una exigencia muy importante es que el tamaño de la muestra debe ser grande para asegurar un menor error de muestreo. Para ítems dicotómicos, se recomienda un numero alrededor de 300 sujetos en las muestras.

Por último, es muy importante la estimación de índices de bondad de ajuste, como el GFI (*Goodness of fit index*), el cual indica la proporción de covarianza explicada por el modelo. Valores superiores a 0.95 indican un buen ajuste (Lloret-Segura et al., 2014). También está el SRMR (*Standardized Root Mean Squared Residual*) el cual informa sobre la diferencia estandarizada entre las correlaciones observadas y las predictivas. Valores inferiores a 0.08 se consideran un buen ajuste (Hu y Bentler, 1998). En general, valores menores que 0.10 son considerados aceptables (Cangur y Ercan, 2015).

2.9 Independencia local

La independencia local se refiere a que, para un nivel de rasgo determinado, las respuestas a los ítems son estadísticamente independiente (Auné et al., 2019). Es decir, no existe relación entre las respuestas a diferentes ítems, dentro del mismo nivel de aptitud. La lógica subyacente es que, la probabilidad de dar respuesta a un primer ítem no condiciona acertar el segundo, sino que es influida únicamente por su nivel de rasgo latente (Muñiz, 2018).

Como ya se ha mencionado, la independencia local se deriva de la unidimensionalidad ya que la respuesta a un ítem solo depende del nivel del único rasgo latente que mide el test. La violación de este supuesto se conoce como dependencia local, y

ocurre cuando la respuesta de un sujeto no depende solamente de su nivel de rasgo latente, sino también de su respuesta a otro ítem o factores comunes a ellos (Martínez Arias et al., 2006). Ítems con dependencia local contienen menos información de lo que el modelo puede predecir y terminarán afectando la precisión de la estimación de los parámetros, particularmente la discriminación, debido a que este parámetro mide la relación de un ítem con el resto de los ítems del test (Chen y Thissen, 1997).

La independencia local puede verse afectada ante la presencia de ítems muy similares, paralelos o basados en el mismo contenido (Auné et al., 2019). También puede observarse dependencia local en tests de velocidad, cuando algunos ítems ubicados al final del test son omitidos por falta de tiempo (Chen y Thissen, 1997).

En la actualidad existen diversos métodos para evaluar la independencia local. La forma más común es mediante el índice χ^2_{LD} propuesta por Chen y Thissen (1997). El cual está basado en la comparación de las frecuencias observadas y las esperadas para cada par de ítems, mediante tablas de contingencia. Si los datos tienen un buen ajuste al modelo, la covarianza de las frecuencias observadas y las esperadas debería ser igual. Caso contrario, será evidencia de que el modelo presenta dependencia. Si se presentan valores mayores que 10 indican la existencia de dependencia local entre determinados pares de ítems. Dentro de este análisis si el estadístico ϕ y la diferencia de odds ratio es positiva significa que las frecuencias observadas del par de ítems tienen más dependencia local que la que predice el modelo, mientras que el signo negativo indica lo contrario (Auné et al., 2019).

Asimismo, Yen (1984) analizó diferentes métodos para determinar existencia de dependencia local, estos son útiles para identificar dependencia local en ítems influidos por el mismo factor o basados en contenido similar. El primero, Q_1 mide el ajuste de ítems al modelo, comparando las curvas características de los ítems observados y predictivos mediante una prueba de chi-cuadrado. Este índice evalúa la dependencia local, pero solo de manera indirecta. El segundo, Q_2 para medir el ajuste para modelo de Rasch, mediante tabla de contingencia de Chi-cuadrado (X^2) de Pearson para cada par de ítems (Yen, 1984). Por último, Q_3 basado en las correlaciones de los residuos para cada par de ítems (Chen y Thissen, 1997).

2.10 Fiabilidad

La fiabilidad es un criterio de la calidad métrica que evalúa el grado de precisión de la medida obtenida en un test (Navas Ara, 2001). Las interpretaciones de las puntuaciones en un test son fiables cuando las medidas son consistentes y carecen de errores de medida (Muñiz, 2018). El análisis de fiabilidad proporciona información sobre la cantidad de error aleatorio que contiene la medida. Es decir, factores no relacionados al constructo medido, pero que influyen en la puntuación del individuo. Estos errores pueden emerger de dos fuentes principales: a) fluctuaciones o inconsistencias de las puntuaciones a lo largo del tiempo y b) diferencias en el contenido entre los ítems o incongruencias entre las respuestas del test. A partir de estas fuentes de error, se conceptualiza la fiabilidad en primer lugar, como estabilidad temporal y, en segundo lugar, como consistencia interna (Navas Ara, 2001).

2.10.1 Fiabilidad desde la TCT

Desde la TCT la fiabilidad es definida como la correlación al cuadrado entre las puntuaciones verdaderas y las puntuaciones empíricas. En otras palabras, es el cociente entre la varianza de las puntuaciones verdaderas y las empíricas (Navas Ara, 2001).

$$\rho_{xx} = \rho_{xw}^2 = \frac{\sigma_v^2}{\sigma_x^2}$$

Donde:

ρ_{xx} = coeficiente de fiabilidad del test;

ρ_{xv} = índice de fiabilidad del test.

Fiabilidad como estabilidad temporal. Una forma de verificar la precisión de un test es medir lo mismo varias veces y evaluar si se obtienen los mismos resultados. En este sentido, un instrumento es fiable si proporciona el mismo tipo de respuestas para diferentes muestras o en repetidas aplicaciones aun mismo grupo, bajo el supuesto de que el constructo medido permanece constante (Frias-Navarro, 2021). Este procedimiento usualmente se realiza mediante el método test-retest en el cual se aplica un test en dos ocasiones a una muestra de sujetos. Para calcular el coeficiente de fiabilidad o estabilidad se calcula la correlación entre las dos puntuaciones obtenidas (Navas Ara, 2001). Una alta correlación

indicará que las dos veces los individuos han quedado ordenados de forma parecida. Cabe señalar que no se sugiere utilizar este método cuando se prevé un cambio en el aprendizaje, o proceso de maduración en el dominio del constructo en los sujetos, como en los instrumentos de evaluación académica, pero se puede utilizar satisfactoriamente en la medición de rasgos y actitudes psicológicas (Morales, 2007).

Otro método desde esta perspectiva es el de formas paralelas, en el cual se aplica dos formas paralelas de un test en dos ocasiones. Para estimar el coeficiente de equivalencia se calcula la correlación entre las puntuaciones obtenidas en las dos formas del test (Navas Ara, 2001). Una correlación alta indicará que ambas formas son equivalentes ya que ordenan los sujetos de manera parecida (Morales, 2007).

Fiabilidad como consistencia interna. Este enfoque busca evaluar la congruencia o consistencia en la actuación de respuestas a los ítems en un test (Navas Ara, 2001). Los métodos para estimar la fiabilidad desde esta perspectiva se basan en la división del test en dos mitades o en un análisis de la covarianza de los ítems. El primero, consiste en dividir el test en dos mitades y ver si la actuación del sujeto en las dos partes es consistente. La correlación entre las dos mitades daría cuenta de la fiabilidad de la equivalencia y consistencia interna del test. El segundo, consiste en el análisis de covarianza de los ítems (Navas Ara, 2001). Es importante señalar que ambos tipos de análisis requieren una única aplicación del test, por lo que son más utilizados en la práctica (Martínez Arias et al., 2006).

El análisis de covarianza, evalúa la proporción de varianza que los ítems tienen en común, en otras palabras, la proporción de la variabilidad de las respuestas que se atribuye al rasgo que está siendo medido, respecto a la variabilidad total de los puntajes (Frias-Navarro, 2021).

El coeficiente más utilizado para estimar la fiabilidad es el coeficiente alfa propuesto por Cronbach en 1951.

$$\alpha = \frac{n}{n-1} \left(\frac{\sum_{j=1}^n \sigma_j^2}{\sigma_x^2} \right) = \frac{n}{n-1} \left(\frac{\sum_{j=1}^n \sum_{k \neq j} \sigma_{jk}}{\sigma_x^2} \right)$$

Donde:

n = número de ítems del test

σ_j^2 = varianza de las puntuaciones de los sujetos en el ítem j

σ^2_x = varianza de las puntuaciones de los sujetos en el test

σ^2_{jk} = covarianza entre las puntuaciones de los sujetos en los ítems j y k

También destaca la fórmula de Kurder-Richardson, derivada del alfa de Cronbach para test con ítems dicotómicos.

$$\alpha = \frac{n}{n-1} \left(\frac{\sum_{j=1}^n p_j q_j}{\sigma_x^2} \right)$$

Donde:

p_j = proporción de respuestas correctas observadas en el ítem j

q_j = proporción de respuestas incorrectas observadas en el ítem j

Existen algunos factores que influyen en la estimación del α de Cronbach, como la longitud del test, de forma que cuantos más ítems tenga el test, el valor del coeficiente sería más alto. Además, la fiabilidad depende de las características de la muestra a la que se aplicó el test, esta puede aumentar cuando la muestra es heterogénea, ya que, a mayor variabilidad, existirá mayor correlación. La muestra también debería ser numerosa, por lo que desde este enfoque la fiabilidad es una propiedad de las puntuaciones del test en una muestra determinada (Navas Ara, 2001).

Es importante señalar que los coeficientes de consistencia interna apoyan el supuesto de unidimensionalidad (Morales, 2007). Debido a que se espera que los ítems que conforman un test midan el mismo constructo, cuando los datos son multidimensionales el valor de consistencia interna será bajo (Frias-Navarro, 2021). Por tanto, se recomienda utilizar estos análisis cuando el test es unidimensional o con contenido homogéneo, sin embargo, no es apropiado en test de velocidad (Navas Ara, 2001).

En la actualidad se acumula evidencia sobre algunas limitaciones del alfa de Cronbach. Primeramente, el alfa de Cronbach asume que los ítems miden un mismo constructo (tau-equivalencia) y que están correlacionados entre sí. También supone que las respuestas son politómicas en escala ordinal o continuas. Además, la independencia o no correlación de los errores. Si el supuesto de la tau-equivalencia no se cumple, el valor del alfa es infraestimado, por el contrario si el supuesto de no correlación entre los errores se

viola, el valor del alfa se sobreestima (Frias-Navarro, 2021). Aunado a esto, el alfa de Cronbach también depende del número de alternativas de respuesta (Ventura-León y Caycho-Rodríguez, 2017).

El omega de McDonald Ω surgió como un método alternativo debido al uso inadecuado del alfa de Cronbach cuando no se cumplen sus supuestos. Este, al igual que el alfa de Cronbach, está basado en el promedio de las correlaciones entre los ítems. El omega de McDonald no exige que los errores estén correlacionados ni la tau equivalencia. La principal diferencia entre el alfa de Cronbach y el omega de McDonald es que éste se fundamenta en la matriz de correlación policórica y entre los ítems, en lugar de la matriz de covarianza o correlación de Pearson, y por ello es más apropiado cuando los datos son ordinales o dicotómicos (Frias-Navarro, 2021). Además, el coeficiente omega de McDonald trabaja con las cargas factoriales estandarizadas, por lo que no depende del número de ítems (Ventura-León y Caycho-Rodríguez, 2017).

$$\Omega = \frac{(\sum_{i=1}^I \lambda_i)^2}{(\sum_{i=1}^I \lambda_i)^2 + (\sum_{i=1}^I 1 - \lambda_i^2)}$$

Donde:

Ω = símbolo de coeficiente omega

λ_i = carga factorial estandarizada

Los valores del coeficiente de fiabilidad oscilan entre 0 a 1, su interpretación depende del uso previsto para el instrumento. Si es para la toma de decisiones debe ser mayor que 0.85, si es para investigación puede ser mayor que 0.60, si es para descripción de grupos puede ser mayor que 0.60 (Morales, 2007). En general, valores menores que 0.70 indicarán baja fiabilidad y valores mayores que 0.95 serían considerados como indicadores de redundancia o duplicación de ítems (Frias-Navarro, 2021).

La fiabilidad juega un rol trascendental en el uso de test con fines de investigación, una baja fiabilidad limitaría el valor de los análisis estadísticos, asimismo, en la utilización de test para la toma de decisiones, una baja fiabilidad afectaría una toma de decisión acertada (Frias-Navarro, 2021).

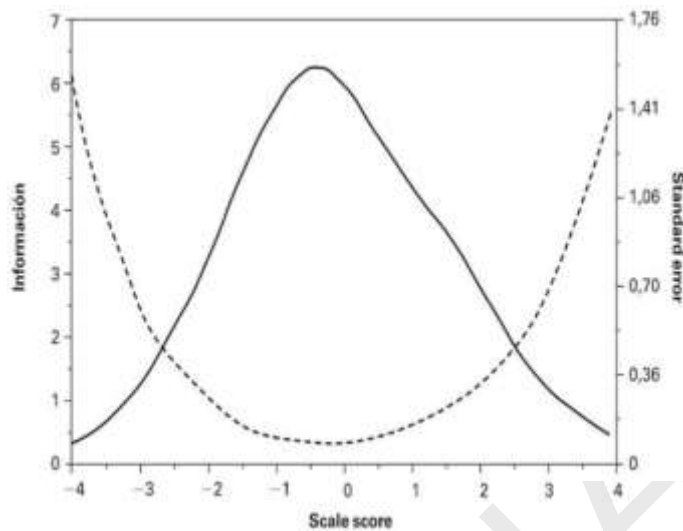
Un coeficiente de fiabilidad bajo indica que hay errores de medición, puede ser producto de una inadecuada formulación de los ítems, causando confusión en los individuos, también puede deberse a que la muestra es muy homogénea (Morales, 2007).

2.10.2 Fiabilidad desde la TRI

Desde la TRI, el indicador de la precisión o fiabilidad del test es definida como información, pues indica la magnitud de la información que proporciona el test para medir el constructo en cada nivel del rasgo (Attorresi et al., 2009; Navas Ara, 2001). Cuanta más información tiene un ítem o un test, más precisa es la estimación de la habilidad para un sujeto (Hidalgo-Montesinos y French, 2016). Esta es representada gráficamente mediante la curva de información del test (Figura 2) (Navas Ara, 2001).

La función de información del ítem se obtiene de una transformación de las curvas características de los ítems (Hidalgo-Montesinos y French, 2016), por lo que está estrechamente relacionado con el parámetro de discriminación. Cuanto mayor sea el poder discriminativo del ítem, mayor será la información que proporcione acerca del constructo medido (Navas Ara, 2001).

Además de la función de información, el error de medición también da cuenta de la precisión de la medida, cuanto mayor sea la información, menor el error típico de estimación. En la TRI, el error se grafica como la inversa de la función de información (Attorresi et al., 2009). Los errores típicos de estimación se establecen en según el nivel de la escala de la habilidad, por lo que un test podrá tener más errores en ciertos niveles que en otros. Esto permite determinar los intervalos de confianza para las puntuaciones (Martínez Arias et al., 2006).

Figura 2. Función de información

Nota. Fuente: Martínez Arias et al. (2006)

La información de los ítems tiene propiedad aditiva, por lo que se pueden sumar las informaciones de cada uno de los ítem y de esta forma, determinar la precisión global de la medición según los niveles de habilidad (Hidalgo-Montesinos y French, 2016). De esto se infiere que el test será más fiable en la medida en que los ítems que lo componen son informativos para estimar con mayor precisión el nivel de habilidad (Martínez Arias et al., 2006).

Evaluar la fiabilidad desde la TRI tiene la ventaja de que indica no solo la cantidad de información que el ítem aporta a la medida del rasgo, sino que la información varía a lo largo de la escala θ , lo que implica que el error de estimación también varía a lo largo de la habilidad. Esto a su vez, supone que el test es fiable para evaluar a los sujetos en cierto nivel de habilidad (Navas Ara, 2001). En este sentido, a diferencia de la TCT, la precisión de las medidas obtenidas no es un valor único, sino que cambia en los diferentes niveles de la habilidad, por lo que permite identificar en cuál nivel, es más preciso el test (Martínez Arias et al., 2006).

Para la interpretación de la función de información no existen reglas generales ya que dependerá del uso previsto y la población para la que se ha diseñado el test (Martínez Arias et al., 2006).

Capítulo 3.

Marco contextual

Esta sección permite obtener una visión panorámica del contexto en que se desarrolla la teoría que fundamenta el estudio, además señala los límites del entorno en que éste se lleva a cabo. Primeramente, se presentan de manera cronológica, tanto internacional como local, los instrumentos que miden la aptitud en un contexto académico. Por otro lado, se muestran los diferentes instrumentos de selección para ingreso a la universidad en el continente americano y en Honduras. También se trata específicamente el ingreso a la formación superior para docentes.

3.1 Instrumentos de aptitud académica

La medición de aptitudes diferenciales surgió de la necesidad de obtener una visión completa de las diversas capacidades de los individuos, en lugar de medidas limitadas como lo eran los test de inteligencia, los cuales proporcionaban una sola calificación de la inteligencia global. Por esta razón, a partir del año 1945 se desarrollaron los test de aptitudes múltiples (Anastasi y Urbina, 1998) con el propósito de medir las capacidades necesarias para aprender ciertas habilidades, además de predecir el éxito en alguna área de trabajo futuro (Brown, 1999). Actualmente existe una amplia cantidad de test que evalúan las aptitudes intelectuales (Lázaro Tortosa, 2015).

Una de las baterías de test más reconocida a nivel global para evaluar aptitudes diferenciales es el Test de Aptitudes Diferenciales (DAT por sus siglas en inglés), ésta fue diseñada en Estados Unidos por G. Bennett, H. Seashore y A. Wesman en 1947. Diseñada para medir las aptitudes múltiples de los alumnos de secundaria y preparatoria. Mide razonamiento, habilidad numérica, exactitud verbal, aptitud mecánica, y secretariales, razonamiento abstracto, razonamiento relaciones espaciales, uso del lenguaje y ortografía (Costa Neiva, 1996). Numerosos estudios reportan la utilidad de la batería DAT para discriminar perfiles de aptitudes (González et al., 2008). Y por esas razones, resulta una gran contribución a la medición de aptitudes (Costa Neiva, 1996).

En suma, los instrumentos de aptitud académica surgieron en respuesta al interés de diferenciar las potencialidades de los individuos; destacando entre ellos el DAT. Como será evidente en los párrafos posteriores, los instrumentos de aptitud han sido utilizados como instrumentos de selección universitaria, por lo que guardan una estrecha relación entre ellos.

3.2 Sistemas de selección universitaria

Históricamente, las instituciones educativas del nivel superior han empleado diversos criterios de selección para elegir su estudiantado (Mendiola et al., 2020). Estos procesos responden a la naturaleza del sistema educativo y de los organismos que administran los procesos de admisión. Algunos de los elementos considerados por los distintos países son: exámenes de egreso de secundaria y de ingreso a universidad, pruebas de aptitud, etc. Suplementariamente, también están los sistemas que otorgan mucho valor al desempeño previo del aspirante durante la secundaria y utilizan criterios como el promedio o ranking de notas de secundaria. Por otro lado, existen otros sistemas que exigen documentos para postulación como ensayos y cartas de recomendación (MINEDUC, 2019). Para efectos del presente estudio, se focalizará únicamente en los sistemas de selección que utilizan pruebas de aptitud o conocimientos.

Estados Unidos. Uno de los sistemas pioneros en la evaluación estandarizada para ingreso a la universidad, fue la Prueba de Aptitud Académica (*Scholastic Aptitude Test*) diseñada y ejecutada en Estados Unidos a partir del año 1926. Esta prueba pretendía medir las capacidades generales para aprender de los estudiantes, y a su vez, predecir su futuro desempeño académico. Más adelante se diseñó el *American College Testing* (ACT), basada en una alineación curricular con el programa de secundaria, el cual se enfocaba en el dominio conceptual de disciplinas particulares. Ambas pruebas obedecían a una metodología normativa, es decir, basados en la comparación entre los estudiantes evaluados. Sin embargo, pruebas basadas en criterios, como *Subject Tests* (SAT) en 1930 y los *Advanced Placement Exams* en 1955 que evalúan el dominio en cada área de aprendizaje han cobrado mayor relevancia, siendo en la actualidad, las más utilizadas en la selección e ingreso a las universidades más prestigiosas del país (Geiser, 2016).

Actualmente el sistema de selección para ingreso a la educación superior opera de manera descentralizada, puesto que cada universidad puede decidir los requisitos que demandará a los postulantes, como la aprobación de prueba de aptitud como la SAT o la ACT, asimismo algunas universidades podrían requerir referentes académicos de secundaria como calificaciones o ranking de egreso; y/o documentos de postulación como ensayos o cartas de recomendación. Aunado a esto, la plataforma *Common Application* funciona como un medio donde los postulantes envían su perfil académico a distintas universidades y son ellas quienes deciden aceptar o rechazar al estudiante, en concordancia con los requerimientos específicos de cada institución (AcciónEducar, 2019). En el presente, *Educational Testing Service* (ETS) es la organización encargada del diseño, estandarización y ejecución de instrumentos de medición con especialistas en las áreas de psicometría y evaluación educativa (Mendiola et al., 2020).

Chile. En este país, han existido tres sistemas de selección para ingresar a la universidad:

- El Bachillerato (1842 a 1966), consistía en un examen oral no estandarizado que evaluaba el dominio de lenguaje, historia, religión y latín, el cual era aplicado a los egresados de la secundaria. Su objetividad y uniformidad eran cuestionables (MINEDUC, 2019), además, la prueba no poseía capacidad predictiva, por lo que perdió credibilidad social (Caiceo Escudero, 2016).
- La Prueba de Aptitud Académica (1967 y 2002), evaluaba áreas comunes: aptitud académica verbal y matemática, e historia de Chile; y pruebas específicas de biología, física, química, matemática y ciencias sociales (Caiceo Escudero, 2013).
- La Prueba de Selección Universitaria (2003 a la actualidad), es una prueba con alineación curricular a los contenidos mínimos obligatorios contenidos en el Plan de Formación general para la Enseñanza Media (Caiceo Escudero, 2013). Está compuesta por dos pruebas obligatorias: Lenguaje y comunicación, y Matemáticas; y pruebas electivas: una prueba de humanidades que agrupa historia, geografía y ciencias sociales y otra prueba se enfoca en el área científica incluyendo química, física y biología (MINEDUC, 2019). Esta prueba es analizada periódicamente para cumplir altos criterios psicométricos analizados desde la TCT y la TRI. Actualmente, es coordinada por el Departamento de Evaluación, Medición y Registro Educacional (DEMRE) (Pérez et al., 2011).

Es importante señalar que el actual sistema de ingreso en Chile es centralizado y toma en cuenta tres factores de selección: las notas de enseñanza media, ranking de notas y la prueba de selección universitaria (MINEDUC, 2019). Evidentemente, este sistema mixto, es muy efectivo ya que, valora tanto los antecedentes académicos como las aptitudes intelectuales de los estudiantes, con la finalidad de escoger los prospectos más capaces (Pérez et al., 2011).

México. De forma contraria a los sistemas mencionados anteriormente, en México, el sistema de selección para ingreso a la educación superior es descentralizado, es decir, cada institución define sus propios requisitos de ingreso. En el caso particular de la prestigiosa Universidad Nacional Autónoma de México (UNAM), el proceso de selección se lleva a cabo desde 1997, mediante un examen escrito con alineación curricular a los planes de estudio de educación media, está compuesto por reactivos de selección única. El examen cumple satisfactoriamente con los criterios psicométricos como evidencias de validez, grado de dificultad, discriminación, confiabilidad, bajo la TCT y la TRI, actualmente es dirigido por la Dirección de Evaluación Educativa del departamento de Coordinación de Desarrollo Educativo e Innovación Curricular (CODEIC) (Mendiola et al., 2020).

Costa Rica. Otro país que destaca por poseer un robusto sistema de selección para ingreso a la universidad es Costa Rica, la trayectoria de la Prueba de Aptitud Académica (PAA) está plasmada detalladamente en la obra de Mainieri (2008), Reseña Histórica de la Prueba de Aptitud Académica que fue presentada en el 9° Congreso Centroamericano de Historia de la Universidad de Costa Rica. Dicha PAA surgió a mediados del siglo 20 como respuesta al ingreso masivo de estudiantes, aumento de la deserción, incapacidad infraestructural para atender la cantidad de estudiantes y necesidad de asegurar la calidad educativa ofrecida. Basada en el SAT de Estados Unidos, la PAA de Costa Rica, mide la aptitud general del aspirante universitario. Actualmente, es un instrumento psicométrico con altos índices de validez y confiabilidad, cuya calidad técnica es evaluada anualmente bajo la Teoría de Respuesta al Ítem con el fin de seleccionar los aspirantes más aptos y predecir su potencial desempeño en los programas ofertados, evitando la deserción por bajos rendimientos.

Venezuela. El Sistema Nacional de Admisión en Venezuela fue constituido en 1984, para ser el organismo regulador de un sistema de admisión centralizado para el ingreso a la Educación Superior, este surge como respuesta a la creciente demanda. La Prueba de Aptitud Académica es, desde entonces, el instrumento que permite seleccionar los candidatos más capaces para desempeñarse en las Instituciones de Educación Superior. La PAA evalúa comprensión de lectura y razonamiento matemático. Sin embargo, existen cuestionamientos sobre la capacidad predictiva de la prueba, además, es considerado como una prueba elitista y excluyente (Pérez y Díaz, 2006).

Honduras. El sistema de admisión a la educación superior en Honduras opera de manera descentralizada en las distintas universidades, siendo estas quienes deciden los requerimientos de ingreso de sus aspirantes. Entre ellas, destaca la Universidad Nacional Autónoma de Honduras, que desde el año 2007 aplica periódicamente como requisito obligatorio de ingreso, la Prueba de Aptitud Académica (PAA) (Duriez et al., 2016; UNAH, 2018). Esta prueba permite predecir el potencial desempeño académico de los aspirantes, evalúa el razonamiento verbal y lectura crítica, razonamiento lógico-matemático y la redacción en español. Cabe mencionar que la PAA, es diseñada, administrada y evaluada por *College Board* en Puerto Rico, cumpliendo con altos estándares psicométricos. Además de la PAA, se realizan pruebas de conocimientos específicos para carreras de las ciencias de la salud y una Prueba de Aprovechamiento Matemático para las ingenierías.

La UPNFM, es la única universidad en el país que diseña su propio test de admisión fundamentado en las teorías psicométricas Clásica y de Respuesta al Ítem, administra y evalúa el test de admisión para sus aspirantes (CEA, 2019b). Este evalúa comprensión lectora, razonamiento verbal, redacción, habilidad numérica, resolución de problemas, razonamiento abstracto y vocación profesional.

En suma, las instituciones educativas de nivel superior tienen sistemas de selección (centralizados o descentralizados) con una variedad de criterios que buscan medir su potencialidad académica de una forma retrospectiva y prospectiva. Uno de los criterios más utilizados son los test de aptitud o conocimientos. Estos se basan en las teorías psicométricas para garantizar la selección de los mejores postulantes. En el siguiente apartado se describirá los instrumentos de selección utilizados para admisión a la formación docente.

3.3 Instrumentos de selección para formación docente

Los sistemas de selección son filtros que reflejan el perfil de profesor que el país demanda (Egido Gálvez, 2010). Sin embargo, en muchas ocasiones los mejores perfiles académicos normalmente no son atraídos por la profesión docente (Katz y Frish, 2016). Sistemas educativos con alto desempeño priorizan la selección de los mejores postulantes para la capacitación docente (McKinsey, 2007; Aguirre-Gómez et al., 2015). Ya que, una mayor selectividad contribuye a atraer candidatos más calificados que puedan promover el aprendizaje de los estudiantes (Aguirre-Gómez et al., 2015; Elacqua et al., 2017).

En cuanto a los criterios de selección tomados en cuenta en los sistemas de admisión, normalmente se otorga mucha relevancia a aspectos cuantitativos como, la calificación promedio mínima de educación media, preparación académica, pruebas de conocimiento o aptitud (Egido Gálvez, 2010). Además, otros se valen del análisis de atributos personales, actitudinales, aptitudinales y de comportamiento, mediante entrevistas individuales, con el fin de indagar factores como la motivación, comunicación oral, liderazgo, relaciones interpersonales y personalidad de los aspirantes (Aguirre-Gómez et al., 2015; Katz y Frish, 2016).

Países como Singapur, Finlandia y Corea de Sur han elevado el estatus de la profesión docente, ya que cuentan con mecanismos eficientes y altamente selectivos de aspirantes a la capacitación docente, atrayendo los futuros docentes de entre los mejores graduados de la cohorte de educación media. Dando especial énfasis en el logro académico de los postulantes, habilidades de comunicación y motivación para la docencia. Todos estos mecanismos elevan la profesión docente, promueven la selectividad y atraen a los postulantes con más alto potencial.

Singapur. El sistema de selección para la capacitación docente en Singapur es centralizado, dirigido por el Ministerio de Educación. El primer requerimiento para todo aspirante es tener una calificación de estudios previos de secundaria dentro del 30% superior. Asimismo, se aplican pruebas de conocimiento y entrevistas donde se analiza la actitud, aptitud y personalidad del estudiante. Además, los futuros docentes son monitoreados durante su formación y reciben un salario, expulsando aquellos que no muestran un óptimo

desempeño. Finalmente, al terminar la formación, son empleados por el Ministerio de Educación (Katz y Frish, 2016; McKinsey, 2007).

Finlandia. El sistema de selección para aspirantes a la docencia en Finlandia es centralizado, siendo el estado quien controla el proceso. Los cupos para ingreso a la formación docente son limitados, esto obedece a la demanda laboral de profesores en el sistema. Involucra primeramente una evaluación nacional de conocimientos sobre aritmética, lengua y resolución de problemas. Además de Prueba de Capacidad Académica General coordinada por las distintas universidades que ofertan el profesorado, esta evalúa procesamiento de datos, razonamiento y habilidades de síntesis. También se aplican entrevistas donde se trata la motivación para enseñar, habilidades comunicativas e inteligencia emocional. Asimismo, se realizan pruebas prácticas grupales donde se evalúa las habilidades interpersonales y de comunicación (McKinsey, 2007).

Corea del Sur. El ingreso a la formación docente primario en Corea del Sur tiene vacantes limitadas, conforme a la demanda y se rige por orden de mérito. Los postulantes a docente primario deben someterse a un Examen Universitario de Ingreso Nacional y solo los que se encuentren dentro del 5% superior son aceptados (McKinsey, 2007). Además, toma en cuenta el desempeño en la secundaria y realiza entrevistas en relación a las actitudes para la profesión docente (Katz y Frish, 2016).

Inglaterra. En este país, el gobierno limita la cantidad de aspirantes mediante el control del financiamiento, con un riguroso control de calidad para las universidades que ofertan el profesorado. Por tal razón, las universidades buscan llenar su cohorte con los mejores estudiantes y así cumplir con el perfil de egreso definido por el gobierno. Los mejores sistemas educativos seleccionan sus aspirantes antes de comenzar su capacitación docente, esto evita la sobreoferta de los candidatos y se garantiza que los futuros docentes tendrán una oportunidad segura de empleo (McKinsey, 2007).

Israel. El sistema de admisión para docentes en Israel involucra el Examen de matriculación (*Matriculation Exam*) y el Examen Psicométrico de Ingreso (*Psychometric Entrance Test*), que son aplicados al terminar la educación secundaria. El primero evalúa el dominio de los contenidos de asignaturas cursadas durante la educación media, el segundo, similar al SAT del Estados Unidos, cubre las áreas de matemáticas, razonamiento verbal e

inglés. Administrados por el Ministerio de educación y el Instituto Nacional de Evaluación. Ambas pruebas poseen un alto grado de validez predictiva en cuanto al desempeño académico de los estudiantes (Katz y Frish, 2016).

Estados Unidos. La selección de docentes se realiza de forma variada en los distintos Estados. El programa de formación en Boston, Chicago y Nueva York, cuenta con criterios de selección que los candidatos seleccionados tienen garantizado su puesto docente.

América Latina. Algunos países permiten que cualquier egresado de secundaria pueda ingresar a la formación docente; sin embargo, otros limitan su ingreso al cumplimiento de ciertos requerimientos, como examen de admisión, puntaje mínimo en pruebas nacionales de habilidades verbales y matemáticas (Puryear, 2015). Por ejemplo, Chile toma en cuenta los puntajes en la prueba de admisión a la universidad y calificaciones de la educación secundaria. De igual forma, Ecuador tiene un puntaje mínimo en el examen de ingreso a las carreras de educación.

México. El proceso de selección para ingreso a la formación docente en México se caracteriza por tener un bajo nivel de exigencia donde el criterio de ingreso es la aprobación del Examen Nacional de Ingreso a la Educación Superior, que evalúa competencias genéricas en áreas de pensamiento matemático, pensamiento analítico, estructura de la lengua y comprensión lectora. Además de las notas promedio de secundaria, sin embargo, no se incluyen aspectos en relación con aptitudes, habilidades o vocación docente (Aguirre-Gómez et al., 2015; Elacqua et al., 2017).

Honduras. En este país, la UPNFM es la única universidad formadora de docentes. Para poder cursar las diferentes carreras ofertadas, los aspirantes son sometidos a un test de admisión que evalúa comprensión lectora, razonamiento verbal, redacción, habilidad numérica, resolución de problemas, razonamiento abstracto y vocación profesional (CEA, 2019b).

De manera general, la selección de futuros docentes es un proceso complejo que involucra aspectos tanto académicos como actitudinales. Se identifica un perfil común tomado en consideración en los diferentes mecanismos de selección para ingreso a la formación docente a lo largo del mundo: primeramente, el docente debe poseer un alto nivel de desempeño en literatura y aritmética. Además, se valora mucho las habilidades

interpersonales y comunicativas, el deseo de aprender, la motivación para enseñar, la adaptabilidad, el pensamiento crítico y reflexivo, la creatividad y principios morales (Katz y Frish, 2016).

DEGT-UNAH

Capítulo 4.

Metodología

En esta sección se detallan los procedimientos metodológicos realizados en el estudio de validación del Test de Razonamiento Mecánico. Este apartado incluye el tipo de investigación, caracterización de la muestra, operacionalización de variables, una descripción del instrumento a validar, además del procedimiento que se ejecutó en el proceso del estudio.

4.1 Tipo de investigación

El estudio de validación del Test de Razonamiento Mecánico es de tipo cuantitativo, ya que se analiza la base de datos de los puntajes obtenidos en la aplicación de la prueba piloto (Hernández Sampieri et al., 2010). Consiste en un diseño Instrumental, ya que se realiza un análisis psicométrico para reunir evidencias de validez del test. La estrategia utilizada es descriptiva, puesto que pretende describir sin manipular las variables, comparar grupos o predecir comportamientos (Ato et al., 2013), de corte transversal, ya que se tomarán en cuenta los datos obtenidos del pilotaje del test, aplicado el 3 de noviembre del año 2019.

4.2 Población y muestra

La población objeto de estudio está compuesta por los aspirantes universitarios a la oferta académica de la UPNFM. Para el pilotaje, se utilizó una muestra no aleatoria, escogida intencionalmente. La muestra utilizada fue de 287 estudiantes de 10° grado de Bachillerato en Ciencias y Humanidades y de Bachilleratos Técnicos Profesionales, en el nivel de educación media de institutos públicos, con cierto nivel de representación en el departamento de Francisco Morazán. La razón de no utilizar estudiantes de último año fue debido a que fue difícil convocarlos a la institución ya ellos se encontraban realizando su práctica profesional. Se presupuso que, utilizar una muestra compuesta por estudiantes de un año inferior, no afecta los resultados, ya que es difícil que al evaluar una variable psicológica en sujetos cuya edad promedio difiera en menos de un año se observe una diferencia importante en la variable latente.

4.3 Variables

Tabla 2. Matriz de Operacionalización de Variables

Definición conceptual	Definición operacional	Dimensiones	Indicadores
Habilidad para comprender los principios mecánicos básicos relacionados con el movimiento, la operación de máquinas y de herramientas.	Puntajes obtenidos en el Test de Razonamiento Mecánico de los aspirantes a ingresar a la UPNFM.	Una sola dimensión: Razonamiento Mecánico	Dado un sistema, determina la fuerza aplicada para mover o sostener un objeto.
			Dado un sistema, determina el estado de un objeto respecto a su velocidad.
			Dado un sistema, determina la dirección de movimiento de un objeto.
			Dado sistema determina la resistencia de un objeto.
			Dado un sistema, determina propiedades de un objeto respecto a la densidad.

Nota. Fuente: Elaboración propia.

4.4 Instrumento de medición

El Test de Razonamiento Mecánico creado por la UPNFM en 2019, evalúa el nivel de aptitud en razonamiento mecánico. El Test responde a un diseño unidimensional, compuesto por 40 ítems. El formato de respuesta es de selección única, con tres opciones de respuesta, recodificadas dicotómicamente (0=correcta y 1=incorrecta).

Para la medición de esta habilidad se consideran sistemas simples los cuales podrían incluir máquinas o herramientas. Se toman en cuenta conceptos como fuerza, movimiento, reposo, energía, densidad y rotación. Asimismo, se considera sistemas como:

1. Máquinas simples (rueda, plano inclinado, cuña, palanca y torno)
2. Mecanismos de transmisión (palancas —de primero, segundo y tercer grado— poleas —simples y de transmisión— y engranajes).

3. Mecanismos de transformación de movimientos (circular a rectilíneo: torno, piñón, cremallera, tornillo y tuerca, circular en alternativo: leva, seguidor, excéntrica, Biela y manivela).
4. Mecanismos de acoplamiento (rígidos, móviles, cojinetes y embragues).
5. Mecanismos disparadores de energía y acumuladores de energía.

4.5 Plan de análisis de datos

Básicamente las fases de la validación del test fueron las siguientes:

1. Análisis de opciones de respuestas: Se analizó las frecuencias de las opciones de respuesta según niveles de habilidad y análisis del poder discriminativo de las opciones de respuesta, sustentado en la TCT.
2. Estimación de parámetros desde la TCT: Se calculó el índice p de dificultad, y la discriminación mediante el punto biserial.
3. Dimensionalidad del test: Se realizó un análisis factorial exploratorio, para determinar si el test cumplía con el supuesto de unidimensionalidad y así poder aplicar modelos de la TRI.
4. Independencia local de los ítems: Se calcularon los estadísticos estandarizados de dependencia local LD χ^2 , para apoyar el supuesto de independencia local de los ítems.
5. Estimación de parámetros desde la TRI: Se aplicó el modelo logístico de dos parámetros.
6. Fiabilidad del Test de Razonamiento Mecánico: Se calculó el Alfa de Cronbach y el Omega de McDonald desde la TCT; además, desde la TRI se estimaron las funciones de información de los ítems y del test.

Para los análisis se utilizaron los programas SPSS, R Studio, IRT pro, Factor Analysis y Jamovi.

4.6 Procedimientos

Los procedimientos logísticos, metodológicos y éticos para conducir la totalidad de la investigación fueron los siguientes:

1. Logísticos: Para la recolección de los datos, se seleccionó una muestra de estudiantes de decimo grado de Bachillerato en Ciencias y Humanidades y de Bachilleratos Técnicos Profesionales de Centros Educativos que tuvieran características similares a la población objeto de estudio. El pilotaje fue realizado al finalizar el año escolar, el día 3 de noviembre del año 2019. Teniendo una duración de 40 minutos su aplicación.
2. Éticos: Se contactó con los centros educativos vía telefónica, y luego de ser confirmada la participación voluntaria, se remitió un correo electrónico informativo con los datos generales del procedimiento de la prueba piloto y los directivos firmaron un consentimiento informado. Antes de la aplicación se notificó a los estudiantes que no estaban obligados a tomar el test, que los resultados obtenidos en el test no tendrían ningún impacto en sus calificaciones y que los datos serían utilizados únicamente para la validación del test. El test solamente fue aplicado a aquellos estudiantes que dieron su consentimiento de forma verbal.
3. Metodológicos: El Test de Razonamiento Mecánico se aplicó a todas las secciones de decimo grado de bachillerato de los centros educativos seleccionados, para lograr una muestra de entre 250 a 300 estudiantes.

Capítulo 5.

Resultados de la investigación

En este apartado se pretende describir e interpretar los resultados del estudio de validación del Test de Razonamiento Mecánico. Está estructurado en función de los objetivos propuestos, se comenzará por el análisis de alternativas de respuestas, seguido por los parámetros psicométricos de los ítems desde la TCT y su confiabilidad, luego la comprobación de supuestos de la TRI por medio del análisis factorial y el análisis de independencia local, por último, se estiman los parámetros psicométricos desde la TRI y la función de información.

5.1 Análisis de alternativas de respuestas

El primer objetivo del presente estudio fue analizar el funcionamiento de las opciones de respuesta de los ítems del Test de Razonamiento Mecánico, para esto se realizó un análisis de la proporción de respuestas por opciones de respuesta y el poder discriminativo de las mismas.

5.1.1 Análisis de proporción de respuestas

Un total de 39 ítems fueron analizados, los cuales consistían en 3 opciones de respuesta: una clave y dos distractores. En suma, se analizaron 39 claves y 78 distractores. En el Apéndice 1 se presenta la tabla completa del funcionamiento de las opciones de respuesta de los ítems. En la Tabla 3 se muestra un resumen de la cantidad de ítems que cumplen con los criterios de calidad de las opciones de respuesta según su proporción de respuesta.

Al inspeccionar la cantidad de estudiantes que seleccionaron cada opción de respuesta, se observa que, en general que la opción correcta es la que tiene una mayor proporción ($M=0.50$). No obstante, de los 39 ítems, 13 parecen ser confusos a los evaluados (ítems 3,10, 12, 13, 14, 15, 19, 25, 28, 33, 34, 37 y 39) ya que los distractores están siendo seleccionados en una proporción mayor a la respuesta correcta, por lo que, se recomienda a los diseñadores del test su revisión o reestructuración (Navas, 2001).

Al analizar la eficiencia de los distractores en cuanto a su proporción de selección, se observa que los ítems 2, 6, 18 y 40, contienen distractores que tuvieron una muy baja proporción de respuesta, siendo menor al 5%, por lo que no están cumpliendo adecuadamente su función como distractores, ya que éstos deberían ser igualmente atractivos para los que no conocen la respuesta correcta.

Al analizar la proporción de la selección de la clave según el nivel de aptitud, se identifica que una proporción mayor de los estudiantes de alta aptitud seleccionaron la respuesta correcta. Solamente los ítems 3, 12, 14, 15, 19, 34, 37 y 39 presentan funcionamiento impropio, donde la proporción de estudiantes con alta aptitud es mayor en los distractores que en la respuesta correcta.

Tabla 3. Análisis de proporción de respuesta

Criterios	Ítems ($N=39$)
La clave posee mayor proporción de respuesta	26 ítems
Ítems con distractores eficientes	35 ítems
Selección de la clave por grupo de alta aptitud	31 ítems

Nota. Fuente: Elaboración propia

5.1.2 Análisis del poder discriminativo de las opciones de respuesta

En la Tabla 4 se presenta un resumen del análisis del poder discriminativo de las opciones de respuesta. En cuanto al índice de discriminación de las opciones de respuesta, se observa que la respuesta correcta presenta un índice mayor y positivo ($M=0.26$) en contraste con los distractores, que en su mayoría son negativos, únicamente el ítem 37 presenta una discriminación negativa en la respuesta correcta. Por otro lado, solamente los ítems 1, 4, 5, 13, 16, 24 y 33, presentan un gran poder discriminativo donde $D \geq 0.40$; los ítems 7, 10, 18, 25, 27, 28, 29, 30, 31, 32, 38 y 40 poseen un índice de discriminación aceptable ($0.30 \leq D \leq 0.39$) entre los estudiantes de bajo desempeño y alto desempeño.

Los siguientes ítems pueden ser sujetos a revisión 3, 8, 9 y 36 ya que poseen poco poder discriminativo ($0.20 \leq D \leq 0.29$). Los ítems para eliminar o modificar sustancialmente

serían 2, 6, 11, 12, 14, 15, 17, 19, 20, 21, 22, 23, 26, 34, 37 y 39 considerando que presentan un inadecuado poder discriminativo ($0.0 < D \leq 0.19$).

Tabla 4. Análisis del poder discriminativo

Criterio	Ítem	Interpretación	Acción
$D \geq 0.40$	7 ítems	Alto poder discriminativo	Conservar
$0.30 \leq D \leq 0.39$	12 ítems	Aceptable poder discriminativo	Conservar
$0.20 \leq D \leq 0.29$	4 ítems	Poco poder discriminativo	Eliminar o modificar
$0.0 < D \leq 0.19$	16 ítems	Inadecuado poder discriminativo	Eliminar

Nota. Fuente: Elaboración propia

En resumen, el funcionamiento de las alternativas de respuesta fue evaluado mediante el análisis de proporción de respuestas seleccionadas en función de alto y bajo nivel de aptitud. Se encontraron 18 ítems que presentan problemas como una mayor proporción de selección a los distractores que la opción correcta en función del nivel de aptitud y distractores con funcionamiento impropio. Además, el análisis del poder discriminativo mediante el índice de discriminación D evidenció que 20 ítems presentan índices con poco e inadecuado poder discriminativo.

Tomando en cuenta ambos análisis solo 19 ítems satisfacen apropiadamente los criterios, por lo que serán empleados en los análisis posteriores ya que son capaces de identificar los distintos niveles de aptitud de razonamiento mecánico.

5.2 Cálculo de parámetros psicométricos desde la TCT y su confiabilidad

En la Tabla 5 se presentan los parámetros de los ítems e índices de confiabilidad del Test de Razonamiento Mecánico calculados desde la TCT. Cabe señalar que solo se tomaron en cuenta los 19 ítems que cumplieron con los criterios de calidad del análisis de opciones de respuesta.

Para determinar el nivel de dificultad de los ítems se calculó la media de aciertos. La dificultad promedio del test fue de 0.519. Se identifica que los ítems la dificultad oscila entre 0.328 y 0.742. Aplicando el criterio de calidad de índice de dificultad, todos los ítems

de la escala se encuentran dentro del rango 0.30-0.80, donde no existe la presencia de valores extremos (Navas, 2001).

De los 19 ítems explorados, 4 ítems (21%) se consideran como difíciles ($p < 0.40$), 5 ítems (26%) clasifican como medianamente difíciles ($0.40 \leq p \leq 0.50$), 10 ítems (52%) clasifican como dificultad media ($0.51 \leq p \leq 0.80$) (Ortiz Romero et al., 2015). Dada esta distribución, se observa que la mayor cantidad de ítems son de dificultad media, medianamente difíciles y difíciles, lo cual es apropiado para el subtest de admisión de razonamiento mecánico, donde se pretende determinar aquellos estudiantes que poseen el rasgo.

Tabla 5. Parámetros TCT

Item	Media de aciertos	Punto Biserial	Alfa de Cronbach si se elimina el ítem	Omega de MacDonald si se elimina el ítem
rm_01	0.564	0.290	0.621	0.624
rm_04	0.463	0.305	0.619	0.623
rm_05	0.714	0.276	0.624	0.627
rm_07	0.523	0.163	0.637	0.640
rm_10	0.425	0.277	0.623	0.627
rm_13	0.390	0.263	0.625	0.629
rm_16	0.547	0.198	0.633	0.637
rm_18	0.742	0.341	0.616	0.618
rm_24	0.481	0.226	0.629	0.633
rm_25	0.328	0.207	0.632	0.635
rm_27	0.436	0.199	0.633	0.636
rm_28	0.334	0.175	0.635	0.639
rm_29	0.620	0.193	0.633	0.637
rm_30	0.638	0.232	0.629	0.632
rm_31	0.557	0.293	0.621	0.624
rm_32	0.599	0.222	0.630	0.633
rm_33	0.338	0.254	0.626	0.630
rm_38	0.456	0.145	0.640	0.642
rm_40	0.700	0.179	0.635	0.639
Escala	0.519		$\alpha = 0.641$	$\Omega = 0.644$

Nota. Fuente: Adaptado de Rstudio y Jamovi

Los resultados obtenidos al determinar la discriminación de los ítems a través de punto biserial muestran que los ítems oscilan entre 0.145 a 0.341. En general, 7 ítems (36%) ítems presentan una discriminación pobre $r_{pb} < 0.20$; 10 ítems (52%) discriminan de forma

regular $0.20 \leq r_{pb} \leq 0.30$ y 2 (10%) ítems tienen un poder discriminativo aceptable $r_{pb} > 0.30$. Para los análisis posteriores, solo se tomaron en cuenta ítems con un punto biserial mayor que 0.20; al hacerlo, 7 ítems fueron descartados. En general, estos resultados informan que la escala tiene, a nivel general, parámetros aceptables dentro del marco de la Teoría Clásica.

Sin embargo, en cuanto a la confiabilidad de la escala, se calculó el alfa de Cronbach (0.641) y el Omega de MacDonald (0.644). Siendo ambos índices menores que 0.70; con lo cual pone en evidencia la baja fiabilidad que posee el test.

5.3 Análisis de dimensionalidad

Antes de efectuar la estimación de los parámetros desde la TRI, se corroboró el supuesto de unidimensionalidad de los ítems requerido por este modelo, mediante un análisis factorial exploratorio.

Un análisis previo para detectar la existencia de puntuaciones marginales demostró que las puntuaciones típicas de las variables presentaban valores fuera del rango ± 3 . Posteriormente se procedió a verificar los supuestos del análisis factorial. Se realizó un test de Kolmogorov-Smirnov para evaluar la normalidad de los datos y se encontró que las variables no seguían una distribución normal ($p < 0.05$). Por otro lado, los valores absolutos de simetría por cada ítem estuvieron en el rango de -0.430 a 1.175, lo que indica variaciones leves de la normal (Pérez y Medrano, 2010). De igual forma, los valores absolutos para curtosis fueron menores que 4, por lo que los ítems tampoco presentan curtosis extrema. Al estimar los índices de asimetría y curtosis multivariada, se encontró que los datos no presentaron valores significativos ($p > 0.05$). Por lo anterior, se decidió realizar el análisis factorial mediante el método de extracción de Ejes Principales, ya que este método no demanda el cumplimiento del supuesto de normalidad.

Se evaluó la linealidad de los resultados mediante la visualización del diagrama de dispersión. Sin embargo, los puntos no se alinearon en torno a una línea recta. Además, para probar el supuesto de colinealidad se examinaron las correlaciones tetracóricas inter-ítem presentados en la Tabla 6, donde ninguno tuvo valores mayores que 0.95, por lo que se asume la existencia de correlación entre las variables sin multicolinealidad. Cabe señalar que se decidió utilizar este tipo de correlación debido a la naturaleza dicotómica de los ítems.

Al revisar la adecuación muestral de los datos se aplicó el test de esfericidad de Barlett (306.5; $gl=45$) cuyo resultado ($p<0.05$) permitió rechazar la hipótesis nula de no correlación entre las variables. Además, el índice de medida adecuación muestral KMO obtenido fue de 0.7 considerado como medianamente aceptable, lo que indica que las correlaciones entre pares de variables son predecibles desde las demás variables, por lo que se considera apropiado realizar el análisis factorial.

El análisis factorial realizado sugiere la presencia de un factor incluyendo únicamente 10 ítems, de los 12 ítems resultantes del análisis de los parámetros desde la TCT. Como se observa en la Tabla 7, el porcentaje de varianza explicada según autovalores del primer factor es de 24%, el cual es superior al porcentaje de varianza explicada del resto de los factores (12% para el segundo factor). Además, el análisis paralelo, presentado en la Tabla 8, sugirió una dimensión con los 10 ítems analizados. En la Tabla 9 se muestra la matriz de estructura. Por último, los índices de ajuste ($GFI=0.951$; $SRMR=0.0830$) indican un ajuste adecuado. Además, el porcentaje real de varianza explicada por el análisis paralelo (32%) es aceptable.

Tabla 6. Matriz de correlaciones tetracóricas

Variable	4	10	13	16	18	24	28	30	31	33
rm_4	1.000									
rm_10	0.242	1.000								
rm_13	0.249	0.273	1.000							
rm_16	0.263	0.148	0.114	1.000						
rm_18	0.113	0.295	0.120	-0.059	1.000					
rm_24	0.135	0.009	0.146	0.054	0.146	1.000				
rm_28	0.145	-0.019	0.223	0.041	0.183	0.339	1.000			
rm_30	-0.021	0.045	0.134	0.115	-0.096	-0.010	0.088	1.000		
rm_31	0.197	0.168	0.069	0.163	0.192	0.000	0.234	0.117	1.000	
rm_33	0.266	0.172	0.182	0.270	0.077	0.268	0.264	0.118	0.213	1.000

Nota. Fuente: Tomado de Factor Analysis

Tabla 7. Proporción de varianza explicada

Variable	Autovalores	Proporción de varianza	Proporción de varianza acumulada
1	2.36491	0.23649	0.23649
2	1.26903	0.12690	
3	1.23548	0.12355	
4	1.02487	0.10249	
5	0.98348	0.09835	
6	0.77480	0.07748	
7	0.65091	0.06509	
8	0.62956	0.06296	
9	0.55734	0.05573	
10	0.50961	0.05096	

Nota. Fuente: Tomado de Factor Analysis

Tabla 8. Análisis paralelo

Factor	% real de varianza explicada	Media de varianza aleatoria	Percentil 95 de % de varianza aleatoria
1	32.1609*	20.3067	23.7439
2	16.3340	17.6902	20.2682
3	12.2628	15.3799	17.4603
4	11.8891	13.1080	14.8491
5	9.9413	10.9462	12.5532
6	7.1605	8.8182	10.5112
7	5.1680	6.6976	8.6215
8	3.6990	4.5992	6.7849
9	1.3842	2.4541	4.4732

Nota. Fuente: Tomado de Factor Analysis

Tabla 9. Matriz de estructura

Ítem	F1
rm_04	0.497
rm_10	0.379
rm_13	0.434
rm_16	0.335
rm_18	0.294
rm_24	0.339
rm_28	0.434
rm_30	0.144
rm_31	0.382
rm_33	0.556

Nota. Fuente: Tomado de Factor Analysis

5.4 Análisis de independencia local

Para estudiar la independencia local, se utilizaron los estadísticos estandarizados de dependencia local (Local Dependency, LD χ^2). Los datos reflejan un buen ajuste al modelo ya que no existe dependencia local, puesto que ninguno de los valores absolutos es mayor o igual a 10. Por tal razón, se puede aceptar el supuesto de independencia local.

Tabla 10. Estudio de independencia local

ítem	Marginal	LD- χ^2 estandarizado								
	χ^2	1	2	3	4	5	6	7	8	9
rm_04	0.0									
rm_10	0.0	-0.7								
rm_13	0.0	-0.5	-0.1							
rm_16	0.0	2.3	-0.6	-0.6						
rm_18	0.0	1.2	0.3	-0.6	2.9					
rm_24	0.0	0.3	-0.6	-0.7	-0.2	1.9				
rm_28	0.0	-0.7	1.0	-0.6	0.5	0.3	0.4			
rm_30	0.0	-0.7	-0.6	-0.3	-0.4	-0.6	-0.1	-0.7		
rm_31	0.0	0.2	-0.6	-0.7	-0.4	-0.7	1.4	-0.7	1.1	
rm_33	0.0	-0.2	-0.5	-0.5	-0.3	-0.4	1.1	-0.6	-0.1	0.9

Nota. Fuente: Tomado de IRT PRO

5.5 Estimación de parámetros psicométricos desde la TRI y su función de información

Para la estimación de parámetros desde la TRI, se aplicó el modelo logístico de dos parámetros. Esta decisión responde en primer lugar a la naturaleza dicotómica de los ítems que componen el test. Además, los análisis previos desde la TCT evidenciaron que los ítems analizados presentan índices de discriminación heterogéneos, por lo que es desaconsejable el modelo de un parámetro que considera la discriminación constante para todos los ítems (Muñiz, 2018).

Dicho lo anterior, para la estimación de los parámetros se tomó en cuenta únicamente los 10 ítems que presentaron correlación punto biserial mayor que 0.20 y que

demonstraron un comportamiento unidimensional. La estimación de los parámetros se efectuó por el método de máxima verosimilitud. En la Tabla 11 se resumen los resultados obtenidos en cuanto a la estimación de los parámetros de discriminación (a) y localización (b) con sus respectivos errores estándar (s.e.) entre paréntesis y los valores de los estadísticos de ajuste para cada uno de los ítems.

El valor promedio de los parámetros de localización de los ítems fue de 0.03 ($SD=0.79$), oscilando en el intervalo de $[-1.39, 0.98]$. En cuanto a los parámetros de discriminación el rango de los ítems osciló en el intervalo de $[0.61, 0.95]$, siendo el promedio 0.80 ($SD=0.13$). Por tanto, ambos parámetros se encuentran dentro de los rangos aceptables ± 3 . En el Apéndice 2 se presentan las curvas características y su función de información para cada ítem. Seguidamente, en el Apéndice 3 se muestra la Curva Característica del Test.

La bondad de ajuste se examinó ítem a ítem mediante el estadístico χ^2 , que contrasta las frecuencias observadas con las esperadas por el modelo, analizando las discrepancias en distintos niveles del rasgo latente. Al considerar un nivel de significancia de 0.01, los 10 ítems presentan un ajuste adecuado al modelo.

Asimismo, se obtuvo la Función de Información del test, la cual es presentada en el Apéndice 4. El valor máximo de Función de Información del Test fue de 2.5, se observa que el test brinda más información en el intervalo de -2 y 2, siendo el punto de inflexión de la curva cercano a 0.5. El error típico de medida alcanzó un valor mínimo de 1.7.

Tabla 11. Parámetros psicométricos desde la TRI

Item	<i>a</i> (s.e.)	<i>b</i> (s.e.)	Ajuste del modelo		
			χ^2	g.l.	<i>p</i>
rm_04	0.92 (0.23)	0.19 (0.15)	8.40	8	0.3969
rm_10	0.67 (0.20)	0.50 (0.23)	3.23	8	0.9190
rm_13	0.88 (0.24)	0.59 (0.20)	6.54	7	0.4795
rm_16	0.59 (0.19)	-0.34 (0.24)	7.06	7	0.4241
rm_18	0.88 (0.25)	-1.39 (0.34)	8.41	7	0.2999
rm_24	0.83 (0.22)	0.11 (0.17)	8.28	7	0.3103
rm_28	0.87 (0.23)	0.91 (0.25)	6.22	7	0.5150
rm_30	0.61 (0.19)	-1.00 (0.35)	4.73	7	0.6940
rm_31	0.95 (0.24)	-0.29 (0.16)	3.53	7	0.8322
rm_33	0.78 (0.23)	0.98 (0.29)	2.22	7	0.9467

Nota. Fuente: Tomado de IRT PRO

Capítulo 6.

Discusión de los resultados

Esta sección tiene el propósito de explicar los resultados obtenidos, haciendo una evaluación crítica de los mismos. Las implicaciones del presente estudio serán informadas tomando como marco de referencia la literatura existente e investigaciones previas realizadas en el campo de la psicometría y validación de tests. Por último, se enumerarán las limitaciones del estudio y recomendaciones para futuras investigaciones.

Este estudio tuvo como propósito validar el Test de Razonamiento Mecánico en etapa de pilotaje de la Universidad Pedagógica Nacional de Honduras. Este test, compuesto por 40 ítems de selección única con 3 opciones de respuesta fue aplicado en noviembre de 2019 a una muestra de 287 estudiantes de decimo grado de Bachillerato en Ciencias y Humanidades y Bachilleratos Técnicos Profesionales, en el nivel de educación media de institutos públicos en Tegucigalpa, Honduras.

En la actualidad, existe una variedad de métodos para el análisis de validez, desde la Teoría Clásica de Test y Teoría de Respuesta al Ítem (Elousa, 2003). El presente estudio de validación consistió acumular evidencias de validez para respaldar la interpretación prevista de los puntajes del test para el uso propuesto, en este caso, la admisión a la UPNFM (APA, AERA, NCME, 2018, p. 14).

Dicho lo anterior, es importante señalar que, en el proceso de validación lo que se valida no es el test en sí mismo, sino el significado o la interpretación que se hace de sus puntuaciones con relación al objetivo al que responde el test. Por lo que la validez es un criterio métrico de la calidad de las interpretaciones del test (Navas Ara, 2001). Dichas interpretaciones o inferencias extraídas de las respuestas de la prueba no son medidas universales, están condicionadas por el uso, contexto y población determinada (Muñiz y Fonseca-Pedrero, 2019). A pesar de que esta nueva noción de la validez como concepto unitario fue introducida desde 1998, aun en la actualidad sigue imperando la idea tradicional de la validez segmentada y como propiedad del test. Lo cual tiene implicaciones metodológicas en todo el proceso de validación (Montero Rojas, 2013).

En esta línea, lo que el presente estudio valida es la calidad métrica del test para hacer inferencias apropiadas sobre la admisión del aspirante a la UPNFM en base al nivel de aptitud de razonamiento mecánico. Los análisis fueron realizados en torno al funcionamiento de las opciones de respuesta de los ítems, determinación de los parámetros psicométricos desde la TCT y su confiabilidad, además se evaluaron los supuestos de unidimensionalidad e independencia local para la estimación de los parámetros psicométricos desde la TRI mediante el modelo logístico de dos parámetros.

De manera general, los resultados a partir de los datos recopilados del estudio piloto indican que solamente 10 ítems cumplen con los criterios métricos de calidad desde la TCT y se ajustan al modelo logístico de dos parámetros, los cuales permiten discriminar niveles medios de razonamiento mecánico.

6.1 Análisis desde la TCT

El primer análisis realizado desde la TCT fue en torno a las alternativas de respuesta, el cual es muy útil e informativo sobre la calidad métrica del test. Particularmente, la calidad de los ítems de opción múltiple dependen del análisis de la proporción de selección de la clave y distractores según niveles de aptitud, y del índice de discriminación (D'Sa y Visbal-Dionaldo, 2017). De forma similar a otros estudios (D'Sa y Visbal-Dionaldo, 2017; Gajjar et al., 2014) se encontró que solo el 50% de los ítems tenía un funcionamiento adecuado en cuanto a los criterios métricos de calidad propuestos. Por tal razón, a partir de este análisis, solo 19 ítems fueron considerados para los análisis posteriores.

El segundo análisis desde la TCT fue el cálculo de los parámetros psicométricos. De forma similar a otros estudios sobre test de admisión universitarios (Backhoff Escudero et al., 2000; Martínez et al., 2011) se encontró que los ítems se distribuyen entre difíciles y medianamente difíciles, lo cual es deseable para el propósito del test, donde se busca identificar si los estudiantes poseen la aptitud evaluada. Sin embargo, se encontró que una considerable proporción de ítems tenía una discriminación pobre, lo cual representa un problema crítico ya que el test debe ser completamente capaz de identificar los diferentes niveles de razonamiento mecánico para la toma de decisiones sobre el ingreso de los

aspirantes a la UPNFM. Por tal razón, fueron excluidos de los análisis posteriores 7 ítems de los 19 que se estaban considerando.

Aunado a esto, los índices de fiabilidad del Test de Razonamiento Mecánico fueron menores que 0.70 el cual podría considerarse suficiente si se tratase de instrumento de investigación o diagnóstico, pero en una evaluación de alto impacto en cuanto a la toma de decisiones importantes como el ingreso a la universidad, estos valores deberían ser mayores que 0.85 (Morales, 2012). No obstante, cabe recalcar que el Test de Razonamiento Mecánico constituye un subtest dentro del Examen de Admisión de la UPNFM y que el mismo se encuentra en etapa de pilotaje, lo que significa que estos valores pueden incrementar al incluir más reactivos en el banco de ítems. En esta línea, resultados similares se encuentran en un análisis psicométrico pionero en subtests de exámenes de admisión (Ponce y Ecurra, 2014). Por último, la baja fiabilidad se ve reflejada en el aumento en el error típico de la medida, por lo que los ítems de la escala deben tratarse con cautela, ya que no se podría garantizar la estabilidad de las puntuaciones en función de las diferentes aplicaciones del test.

6.2 Análisis desde la TRI

En cuanto a la estimación de parámetros desde la TRI, se aplicó el modelo logístico de dos parámetros. Para esta decisión se tomó en consideración la dificultad cumplir el requerimiento de homogeneidad en los parámetros de discriminación para aplicar el modelo de Rasch y la preferencia hacia la parsimonia evitó la elección de un modelo de tres parámetros (Martínez Arias et al., 2006). En este respecto, Lord (1980) aseveró que la elección del modelo se justifica siempre que los resultados empíricos lo avalen.

Otro factor a tomar en consideración es el tamaño de la muestra, si bien la TRI requiere tamaños de muestra muy amplias (aproximadamente a partir de 500 participantes) para obtener estimaciones precisas (Muñiz, 2018), se han realizado estudios que han evaluado el efecto del tamaño de la muestra y cantidad de ítems que componen el test en la estimación de parámetros utilizando diferentes modelos de la TRI, en los cuales se ha sugerido que la asociación entre la cantidad de ítems y la cantidad de la muestra son elementos que inciden en la robustez de las estimaciones. En este sentido, se considera

aceptable la utilización de una muestra de 250 en un test compuesto por 30 ítems (Sahin y Anil, 2017).

Ahora bien, al estimar los parámetros desde la TRI, se aplicó el modelo logístico de dos parámetros, de los 12 ítems examinados, solo 10 se ajustaron al modelo. El análisis factorial exploratorio permitió confirmar el supuesto de unidimensionalidad, 10 ítems presentaron un comportamiento unidimensional. El AFE mostró que hay un factor general que explica más del 20% de la varianza de todos los ítems, mientras que el segundo factor logró explicar un porcentaje considerablemente inferior. Posteriormente, se verificó el supuesto de independencia local. Como afirmaron Lord y Novick (1968), la independencia local se da por satisfecho al comprobar la unidimensionalidad del test.

Tras la comprobación de supuestos, se encontró que la mayoría de los parámetros de localización estimados se concentraron en niveles medios-bajos del rasgo y no distribuidos en todos los niveles de la aptitud, particularmente la ausencia de ítems que discriminen en los niveles más elevados de la aptitud. Es decir que, estudiantes con un nivel bajo de razonamiento mecánico tendrían un 50% de probabilidad de responder correctamente. Con relación a los parámetros de discriminación se consideran aceptables. Sin embargo, la precisión es mayor para discriminar estudiantes con niveles entre -2 y 2 del rasgo aproximadamente. Esto revela la debilidad de las propiedades psicométricas del Test Razonamiento Mecánico para discriminar en niveles altos del rasgo, considerando el propósito del test de seleccionar a los estudiantes que posean mayores niveles del rasgo. Por lo que es de vital importancia incorporar nuevos reactivos a la escala, ya que esto mejoraría significativamente los parámetros.

En general, de los 40 ítems comprendidos en la prueba, solamente 10 se ajustaron al modelo logístico de dos parámetros. Los resultados revelan las falencias del Test de Razonamiento Mecánico, además, el comportamiento de los ítems permite obtener información sobre la complejidad del constructo medido y sirve como referencia para la construcción de nuevos ítems a partir de los conformados por la escala. Por lo anterior, se concluye que solo 10 ítems cumplen con los criterios métricos de calidad, los cuales brindan información en los niveles medios - bajos de la aptitud.

6.3 Limitaciones

Este estudio permitió obtener evidencias de validez del Test de Razonamiento Mecánico de la UPNFM. Sin embargo, se identificaron algunas limitaciones.

En primer lugar, se sabe que las estimaciones de los parámetros de los ítems desde la TRI, mejoran su robustez en muestras más numerosas (aproximadamente 500 sujetos) en comparación a la cantidad de la muestra utilizada en el presente estudio (López, 1995; Muñiz, 2018). Es importante mencionar que para la realización de este estudio no se tuvo acceso al contenido de los ítems, tampoco se tuvo control sobre la muestra en relación con la recolección de datos demográficos, los cuales resultan de vital importancia para estudios adicionales como funcionamiento diferencial de los ítems.

Por otro lado, la eliminación de aquellos ítems que muestran índices pobres y mal ajuste al modelo es completamente necesario y permite que los análisis sean más efectivos (D'Sa y Visbal-Dionaldo, 2017). Sin embargo, esta situación afecta la validez de contenido, ya que se corre el riesgo de la infrarrepresentación del constructo de razonamiento mecánico y consecuentemente puede afectar de forma significativa los resultados de la prueba (APA, AERA, NCME, 2018). La calidad de las medidas solo podrá ser justificada si los ítems representan cabalmente el constructo de medición (Prieto y Delgado, 2010). Por lo tanto, es de vital importancia nutrir el banco de ítems con reactivos homogéneos conceptual y representativamente esenciales, tomando como referencia los ya existentes.

Por último, en la actualidad se encontró poco acceso a estudios sobre propiedades psicométricas o de validación sobre exámenes de admisión relacionados específicamente al razonamiento mecánico.

6.4 Recomendaciones para estudios futuros

A pesar de que la Teoría de Respuesta al Ítem ha sido muy bien establecida en otros países, en Latinoamérica sigue predominando la aplicación de la TCT, por lo que constituye una necesidad crítica el desarrollar más estudios psicométricos basados en la TRI. Esto permitirá abrir nuevas oportunidades para el crecimiento de la psicometría en Honduras y proveer procedimientos específicos para la validación de instrumentos utilizados para la admisión a instituciones de educación superior.

En la actualidad, existe una variedad de métodos para la obtención de evidencias de validación, lo que convierte a los estudios de validación en un proceso amplio e interminable (Elousa, 2003). Por lo que se sugiere continuar con el proceso de validación, particularmente obtener evidencias de validez predictiva o relación con otras variables, como desempeño académico en las carreras donde el razonamiento mecánico es crucial. Esto permitirá obtener proyecciones confiables sobre el perfil de ingreso con respecto a la eficiencia terminal.

Capítulo 7.

Conclusiones

Esta tesis tuvo como objetivo realizar un estudio de validación del Test de Razonamiento Mecánico, diseñado por la Coordinación de Examen de Admisión de la Universidad Pedagógica Nacional. Para lo cual se realizaron una serie de análisis psicométricos, tanto en el marco de la Teoría Clásica de los Tests como en el de la Teoría de Respuesta al Ítem. Los resultados obtenidos en el presente trabajo servirán para orientar las acciones por parte de la Coordinación de Examen de Admisión para avanzar en el proceso de construcción del test y la alimentación del banco de ítems. Asimismo, este trabajo puede servir de referencia para efectuar otros estudios similares relacionados a validación de tests.

El primer objetivo específico fue analizar el funcionamiento de las opciones de respuesta de los ítems del Test de Razonamiento Mecánico. Este permitió determinar que el 50% de los ítems cumplió con los criterios de calidad en cuanto a la proporción de selección de la clave y distractores según niveles altos y bajos de la aptitud, asimismo con índices aceptables de discriminación D . Este primer análisis es clave para identificar aquellos ítems que deben ser modificados o eliminados de análisis posteriores al no ser funcionales.

El segundo objetivo fue determinar los parámetros psicométricos de los ítems del Test de Razonamiento Mecánico. De acuerdo con los análisis realizados en el marco de la Teoría Clásica de los Test, el Test de Razonamiento Mecánico presentó parámetros aceptables de dificultad y discriminación solamente para algunos ítems. Los análisis realizados desde la Teoría de Respuesta al Ítem permitieron identificar 10 ítems que se ajustan al modelo logístico de dos parámetros, permitiendo la discriminación de niveles medios-bajos de la aptitud.

El tercer objetivo fue analizar la dimensionalidad del Test de Razonamiento Mecánico, para lo cual se realizó un análisis factorial exploratorio. Los resultados mostraron la presencia de un único factor tomando en cuenta 10 ítems, por lo que el supuesto de unidimensionalidad del Test fue satisfecho para la aplicación del modelo logístico de dos parámetros.

El cuarto objetivo fue analizar la independencia local de los ítems del Test de Razonamiento Mecánico, el cual fue obtenido mediante los estadísticos estandarizados de dependencia local. Este permitió confirmar que los ítems no poseían dependencia local, por lo que se pudo cumplir con el supuesto de independencia local de los ítems.

El último objetivo fue determinar la fiabilidad de las puntuaciones del Test de Razonamiento Mecánico. Los índices de fiabilidad estimados desde la TCT resultaron bajos tomando en cuenta que es un test de alto impacto. Lo cual revela una fuerte debilidad en la consistencia interna de la prueba. Además, los resultados obtenidos por medio de la TRI demostraron que el test proporciona mediciones más precisas en los niveles medios de la aptitud.

Aunado a lo anterior, una proporción significativa (más del 60%) de ítems presentaron parámetros tales que se recomienda eliminarlos, lo cual supone una vulneración a la validez de contenido, en cuanto a la representación del constructo. Para obtener una escala con resultados más robustos será necesario incorporar de nuevos ítems alineados a la tabla de especificaciones.

En conclusión, es tarea fundamental de las instituciones educativas de nivel superior garantizar que el examen de admisión, diseñado y aplicado para seleccionar a los aspirantes, cumpla satisfactoriamente con los criterios psicométricos de validación. Esto permitirá determinar eficazmente los perfiles estudiantiles más aptos para rendir apropiadamente en su formación académica y su posterior desempeño laboral.

Capítulo 8.

Referencias

- AcciónEducar. (2019). *Análisis comparado de los sistemas de acceso a la educación superior*. Santiago, Chile.
- Aguilar Parra, J. M., Álvarez, J. y Lorenzo, J. J. (2016). Estudio sobre las pruebas de la oposición de acceso a la función pública docente. Variables influyentes en cada fase de la oposición. *Educación XX1*, 19(1), 357–379.
- Aguirre-Gómez, F., Dévora-Rodríguez, C. y Valenzuela-Dorado, E. (2015). Las condiciones de ingreso a la profesión docente: un factor para el logro de la calidad educativa. *Ra Ximhai*, 11(4), 405–412.
- Aiken, L. (2003). *Test Psicológicos y evaluación* (11ª ed.). Pearson Education.
- Anastasi, A. y Urbina, S. (1998). *Test Psicológicos*.
- Anderson, M., Meyer, B. y Olivier, P. (2002). *Diagrammatic Representation and Reasoning: Mechanical Reasoning about Gear-and-belt Diagrams: Do Eye-movements Predict Performance?* Springer.
- APA, AERA, NCME. (2018). *Estándares para pruebas educativas y psicológicas (Spanish Edition)*. American Educational Research Association. Edición de Kindle. Original work published 2014
- Ato, M., López-García, J. J. y Benavente, A. (2013). Un sistema de clasificación de los diseños de investigación en psicología. *Anales De Psicología*, 29(3). <https://doi.org/10.6018/analesps.29.3.178511>
- 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), 179–188.
- Auné, S. E., Abal, F. J. P. y F Attorresi, H. (2019). Teoría de la respuesta al ítem: Su utilización en América Latina. Supuestos de unidimensionalidad e independencia local. *International Journal of Psychological Research*, 12(1), 49–56. <https://doi.org/10.21500/20112084.3753>

- Backhoff Escudero, Eduardo, Larrazolo Reyna, Norma, Rosas Morales y Martín (2000). Redalyc. Nivel de dificultad y poder de discriminación del Examen de Habilidades y Conocimientos Básicos (EXHCOBA). *Revista Electrónica de Investigación Educativa*. *Revista Electrónica De Investigación Educativa*, 2(1), 11–28.
- Barber, M. y Mourshed, M. (2007). *How the world's best performing school systems come out on top*. McKinsey & Company. http://www.mckinsey.com/locations/ukirland/publications/pdf/Education_report.pdf
- Bennet, G. K. (2008). *Bennett Mechanical Comprehension Test*. Pearson Education. <https://doi.org/10.4135/9781412952675.n16>
- Brown, F. G. (1999). *Principios de la medición en Psicología y Educación*.
- Burga León, A. (2006). La unidimensionalidad de un instrumento de medición: perspectiva factorial. *Revista De Psicología De La PUCP*, 14(1).
- Caiceo Escudero, J. (2013). El Sistema de Selección Universitaria Chilena en crisis: Discusión para la elaboración de la P.A.A. *HISTEDBR on-Line*(51), 3–13.
- Caiceo Escudero, J. (2016). Los sistemas estandarizados de evaluación en Chile: participación de Mario Leyton Soto y Erika Himmel König. *Historia De La Educación*, 34(0), 357-371. <https://doi.org/10.14201/hedu201534357371>
- Cangur, S. y Ercan, I. (2015). Comparison of Model Fit Indices Used in Structural Equation Modeling Under Multivariate Normality. *Journal of Modern Applied Statistical Methods*, 14(1), Artículo jmasm.eP1810, 152–167. <https://doi.org/10.22237/jmasm/1430453580>
- CEA. (2018). *Test Vocacional de Acceso a la Formación Docente: Manual para el aspirante*. Tegucigalpa. Coordinación de Examen de Admisión. <https://web.upnfm.edu.hn/admision/phocadownload/ManualAspirante2018.pdf>
- CEA. (2019a). *Base de datos de Coordinación de Examen de Admisión*.
- CEA. (2019b). *El Proceso de Admisión en la UPNFM: Boletín Informativo*. Coordinación del Examen de Aptitud.

- Chen, W.-H. y Thissen, D. (1997). Local Dependence Indexes for Item Pairs. *Journal of Educational and Behavioral Statistics*, 22(3), 265–289.
- Cizek, G. J., Bowen, D. y Church, K. (2010). Sources of Validity Evidence for Educational and Psychological Tests: A Follow-Up Study. *Educational and Psychological Measurement*, 70(5), 732–743. <https://doi.org/10.1177/0013164410379323>
- Costa Neiva, K. M. (1996). *Manual de Pruebas de Inteligencia y Aptitudes* (Primera). Editorial Plaza y Valdés.
- Digeduca. (2015). *Análisis de ítems con jMetrik: Serie de cuadernillos Técnicos*. Guatemala.
- D'Sa, J. L. y Visbal-Dionaldo, M. L. (2017). Analysis of Multiple Choice Questions: Item Difficulty, Discrimination Index and Distractor Efficiency. *International Journal of Nursing Education*, 9(3), 109–114. <https://doi.org/10.5958/0974-9357.2017.00079.4>
- Duriez, M., López, V. y Moncada, G. (2016). *Educacion Superior en Iberoamerica: Informe 2016*.
- Egido Gálvez, I. (2010). El acceso a la profesión docente en España en perspectiva europea.: Algunas reflexiones orientadas a la mejora de la selección del profesorado. *Educación XXI*, 13(2), 47–67.
- Elacqua, G., Hincapié, D., Vegas, E., Alfonso, M., Montalva, V. y Paredes, D. (2017). *Profesión: Profesor en América Latina: ¿Por qué se perdió el prestigio docente y cómo recuperarlo?* Banco Interamericano de Desarrollo.
- Elousa, P. (2003). Sobre la validez de los tests. *Psicothema*, 15(2), 315–321.
- Ferrando y Joan, P. (1996). Evaluación de la unidimensionalidad de los ítems mediante análisis factorial. *Psicothema*, 8(2), 397–410.
- Ferrando, P. J. y Anguiano-Carrasco (2010). El análisis factorial como técnica de investigación en psicología. *Papeles Del Psicólogo*, 31(1), 18–33.
- Frias-Navarro, D. (2021). *Apuntes de consistencia interna de las puntuaciones de un instrumento de medida*. <https://doi.org/10.17605/OSF.IO/KNGTP>
- Gajjar, S., Sharma, R., Kumar, P. y Rana, M. (2014). Item and Test Analysis to Identify Quality Multiple Choice Questions (MCQs) from an Assessment of Medical Students

- of Ahmedabad, Gujarat. *Indian Journal of Community Medicine : Official Publication of Indian Association of Preventive & Social Medicine*, 39(1), 17–20. <https://doi.org/10.4103/0970-0218.126347>
- Geiser, S. (2016). Medición y evaluación para los procesos de admisión de la educación superior: Hallazgos desde California. *Pensamiento Educativo: Revista De Investigación Educativa Latinoamericana*, 53(1), 1–18. <https://doi.org/10.7764/PEL.53.1.2016.7>
- Gierl, M. J., Bulut, O., Guo, Q. y Zhang, X. (2017). Developing, Analyzing, and Using Distractors for Multiple-Choice Tests in Education: A Comprehensive Review. *Review of Educational Research*, 87(6), 1082–1116. <https://doi.org/10.3102/0034654317726529>
- González, G., Castro Solano, A. y González, F. (2008). Perfiles Aptitudes, Estilos de Pensamiento y Rendimiento Académico. *Anuario De Investigaciones*, 8, 33–41. <http://www.redalyc.org/articulo.oa?id=369139944035>
- Güler, N., Kaya Uyanık, G. y Taşdelen Teker, G. (2014). Comparison of classical test theory and item response theory in terms of item parameters. *International Association of Social Science Research*, 2(1), 1–6.
- Hegarty, M. (2004). Mechanical reasoning by mental simulation. *Trends in Cognitive Sciences*, 8(6), 280–285. <https://doi.org/10.1016/j.tics.2004.04.001>
- Hernández Sampieri, R., Fernández Collado, C. y Baptista Lucio, P. (2010). *Metodología de la investigación* (6a ed.). McGraw-Hill.
- Hidalgo-Montesinos, M. D. 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.
- Hu, L. y Bentler, P. M. (1998). Fit indices in covariance structure modeling: Sensitivity to underparameterized model misspecification. *Psychological Methods*, 3(4), 424–453. <https://doi.org/10.1037/1082-989X.3.4.424>

- Hurtado, L. L. (2018). Relación entre los índices de dificultad y discriminación. *Revista Digital De Investigación En Docencia Universitaria*, 12(1), 273–300. <https://doi.org/10.19083/ridu.12.614>
- Injoque-Ricle, I., Formoso, J., Calero, A. y Caruso, G. (2019). Razonamiento mecánico, memoria de trabajo y velocidad de procesamiento. *Liberabit: Revista Peruana De Psicología*, 25(1), 71–84. <https://doi.org/10.24265/liberabit.2019.v25n1.06>
- Juliá, C. y Antolí, J. Ó. (2018). Enhancing spatial ability and mechanical reasoning through a STEM course. *International Journal of Technology and Design Education*, 28(4), 957–983. <https://doi.org/10.1007/s10798-017-9428-x>
- Katz, S. y Frish, Y. (2016). A new admission system model for teacher colleges. *Quality Assurance in Education*, 24(4), 463–489. <https://doi.org/10.1108/QAE-01-2014-0003>
- Lázaro Tortosa, F. R. (2015). *Relación de Aptitudes Musicales, Intelectuales y Rasgos de Personalidad e Identificación del Talento Musical en escolares de diez a doce años*.
- Ledesma, R., Ferrando, P. y Tosi, J. (2019). Uso del Análisis Factorial Exploratorio en RIDEP. Recomendaciones para Autores y Revisores. *Revista Iberoamericana De Diagnóstico Y Evaluación - E Avaliação Psicológica*, 52(3). <https://doi.org/10.21865/RIDEP52.3.13>
- Liu, A. S. y Schunn, C. D. (2017). Applying math onto mechanisms: Mechanistic knowledge is associated with the use of formal mathematical strategies. *Cognitive Research: Principles and Implications*, 2(1), 6. <https://doi.org/10.1186/s41235-016-0044-1>
- Lloret-Segura, S., Ferreres-Traver, A., Hernández-Baeza, A. y Tomás-Marco, I. (2014). El análisis factorial exploratorio de los ítems: una guía práctica, revisada y actualizada. *Anales De Psicología*, 30(3). <https://doi.org/10.6018/analesps.30.3.199361>
- López, J. A. (1995). Estimación de parámetros en la TRI: Una evaluación de BILOG en muestras pequeñas. *Psicothema*, 7(1), 173–185.
- Lord, F. M. (1980). *Applications of Item Response Theory to Practical Testing Problems*. Lawrence Erlbaum.

- Lord, F. M. y Novick, M. R. (1968). *Statistical theories of mental tests scores*. Addison-Wesley.
- Mainieri, A. (2008). Reseña Histórica de la Prueba de Aptitud Académica (PAA-U.C.R.). *Revista Electrónica De Historia*(numero especial).
- Martínez, E. A. C., Parra, L. A. G. y Moreno, R. R. (2011). *Análisis de las propiedades psicométricas de un examen de admisión para aspirantes a ingeniería*. XI Congreso Nacional de Investigación Educativa.
- Martínez Arias, M. R., Hernández Lloreda, M. J. y Hernández Lloreda, M. V. (2006). *Psicometría* (Edición electrónica). Alianza Editorial.
- Matas, A. (2010). Introducción al análisis de la Teoría de Respuesta al Ítem.
- Mavrou, I. (2015). Análisis factorial exploratorio Cuestiones conceptuales y metodológicas. *Revista Nebrija De Lingüística*, 19.
- McKenna, A. F. y Agogino, A. M. (2004). Supporting Mechanical Reasoning with a Representationally-Rich Learning Environment. *Journal of Engineering Education*, 93(2), 97–104. <https://doi.org/10.1002/j.2168-9830.2004.tb00794.x>
- McKinsey. (2007). *Cómo hicieron los sistemas educativos con mejor desempeño del mundo para alcanzar su objetivo*. McKinsey & Company.
- Mendiola, M., García Minjares, M., Martínez González, A. y Buzo Casanova, E. (2020). El Examen de Ingreso a la Universidad Nacional Autónoma de México: Evidencias de Validez de una Prueba de Alto Impacto y Gran Escala. *Revista Iberoamericana De Evaluación Educativa*, 13(2), 107–128. <https://doi.org/10.15366/riee2020.13.2.006>
- MINEDUC. (2019). *¿Qué sabemos sobre admisión a la educación superior? Una revisión para implementación del nuevo Sistema de Acceso en Chile*. Santiago, Chile.
- Montero Rojas, E. (2013). Referentes conceptuales y metodológicos sobre la noción moderna de validez de instrumentos de medición: implicaciones para el caso de personas con necesidades educativas especiales. *Actualidades En Psicología*, 27(114), 113–128.
- Morales, P. (2007). *La fiabilidad de los tests y escalas*.
- Morales, P. (2012). Análisis de ítems en las pruebas objetivas.

- Muñiz, J. (2010). Las teorías de los Tests: Teoría Clásica y Teoría de respuesta a los Ítems. *Papeles Del Psicólogo*, 31(1), 57–66.
- Muñiz, J. (2018). *Introducción a la Psicometría: Teoría Clásica y TRI*. Pirámide.
- Muñiz, J. y Fonseca-Pedrero, E. (2019). Ten steps for test development. *Psicothema*, 31(1), 7–16. <https://doi.org/10.7334/psicothema2018.291>
- Navas Ara, M. J. (2001). *Métodos, diseños y técnicas de investigación psicológica*. UNED.
- Ortiz Romero, G. M., Díaz Rojas, P. A., Llanos Domínguez, O. R., Pérez Pérez, S. M. y González Sapsin, K. (2015). Dificultad y discriminación de los ítems del examen de Metodología de la Investigación y Estadística. *EduMeCentro*, 7(2), 19–35.
- Pérez y Díaz, M. E. (2006). La Prueba de Aptitud Académica: Una visión de la subprueba de Comprensión de Lectura. *Investigación Y Postgrado*, 21(2), 143–176.
- Pérez, Ortiz, L. y Parra, P. (2011). Prueba de Selección Universitaria, rendimiento en enseñanza media y variables cognitivo-actitudinales de alumnos de Medicina. *Rev Educ Cienc Salud*, 3(2), 120–127.
- Pérez, E. R. y Medrano, L. (2010). Análisis Factorial Exploratorio: Bases Conceptuales y Metodológicas. *Revista Argentina De Ciencias Del Comportamiento*, 2(1), 58–66.
- Ponce, C. y Ecurra, M. (2014). Estudio psicométrico sobre el examen de admisión 2008-I a la Universidad Nacional Mayor De San Marcos. *Revista De Investigación En Psicología*, 11(1), 137. <https://doi.org/10.15381/rinvp.v11i1.3886>
- Prieto, G. y Delgado, A. R. (2010). Fiabilidad y Validez. *Papeles Del Psicólogo*, 31(1), 67–74. <https://www.redalyc.org/articulo.oa?id=778/77812441007>
- Puryear, J. (2015). Preparando a docentes de alta calidad en América Latina: Nota de Política PREAL.
- Remmers, H. H. y Smith, J. M. (1936). Reliability and practice effect in the O'Connor Wiggly Block Test: *Journal of Applied Psychology*, 20(5), 591–598. doi:10.1037/h0054862. *Journal of Applied Psychology*, 20(5), 591–598. <https://doi.org/10.1037/h0054862>

- Rios, J. y Wells, C. (2014). Validity evidence based on internal structure. *Psicothema*, 26(1), 108–116. <https://doi.org/10.7334/psicothema2013.260>
- Rojas Torres, L. (2014). Evidencias de validez de la Prueba de Aptitud Académica de la Universidad de Costa Rica basadas en su estructura interna. *Actualidades En Psicología*, 28(116), 15–26.
- Rosales López, P. P. (2012). *Uso del test de aptitudes mecanicas de Mac Quarrie en la selección de personal para mejorar la productividad.*
- Sahin, A. y Anil, D. (2017). The Effects of Test Length and Sample Size on Item Parameters in Item Response Theory. *Educational Sciences: Theory & Practice*, 17(1), 321–335. <https://doi.org/10.12738/estp.2017.1.0270>
- Toland, M. D. (2014). Practical Guide to Conducting an Item Response Theory Analysis. *The Journal of Early Adolescence*, 34(1), 120–151. <https://doi.org/10.1177/0272431613511332>
- UNAH. (2018). *Manual de Procedimientos Académicos de la UNAH: Sección de procedimientos académicos de la UNAH.*
- UPNFM. (2005). *Reglamento del Regimen Académico para programas de pregrado.* Universidad Pedagógica Nacional Francisco Morazán. <https://web.upnfm.edu.hn/test/phocadownload/Registro/REGLAMENTOS/REGLAMENTO%20DE%20REGIMEN%20ACADEMICO%20PARA%20PROGRAMAS%20DE%20PREGRADO.pdf>
- UPNFM. (2019). *Memoria Estadística UPNFM.*
- Velásquez Ospina, F. A. (2008). ¿En qué consiste la quinta edición del DAT? Una aproximación actualizada a la medición de actitudes. *Avances En Medición*(6), 175–182.
- Ventura-León, J. L. y Caycho-Rodríguez, T. (2017). El coeficiente Omega: un método alternativo para la estimación de la confiabilidad. *Revista Latinoamericana De Ciencias Sociales, Niñez Y Juventud*, 15(1), 625–627. <https://doi.org/10.11600/1692715x.12117101613>

- Wesley, F. (1937). *Determining the reliability and validity of the Clinton mechanical ability tests.*
- Yen, W. M. (1984). Effects of Local Item Dependence on the Fit and Equating Performance of the Three-Parameter Logistic Model. *Applied Psychological Measurement*, 8(2), 125–145. <https://doi.org/10.1177/014662168400800201>

Declaración de ética de la investigación

Esta investigación ha sido realizada tomando en consideración el código de ética de investigación. No se reportan conflictos de interés.

DEGT-UNAH

Capítulo 9.

Apéndice

Apéndice 1. Funcionamiento de las opciones de respuesta de los ítems

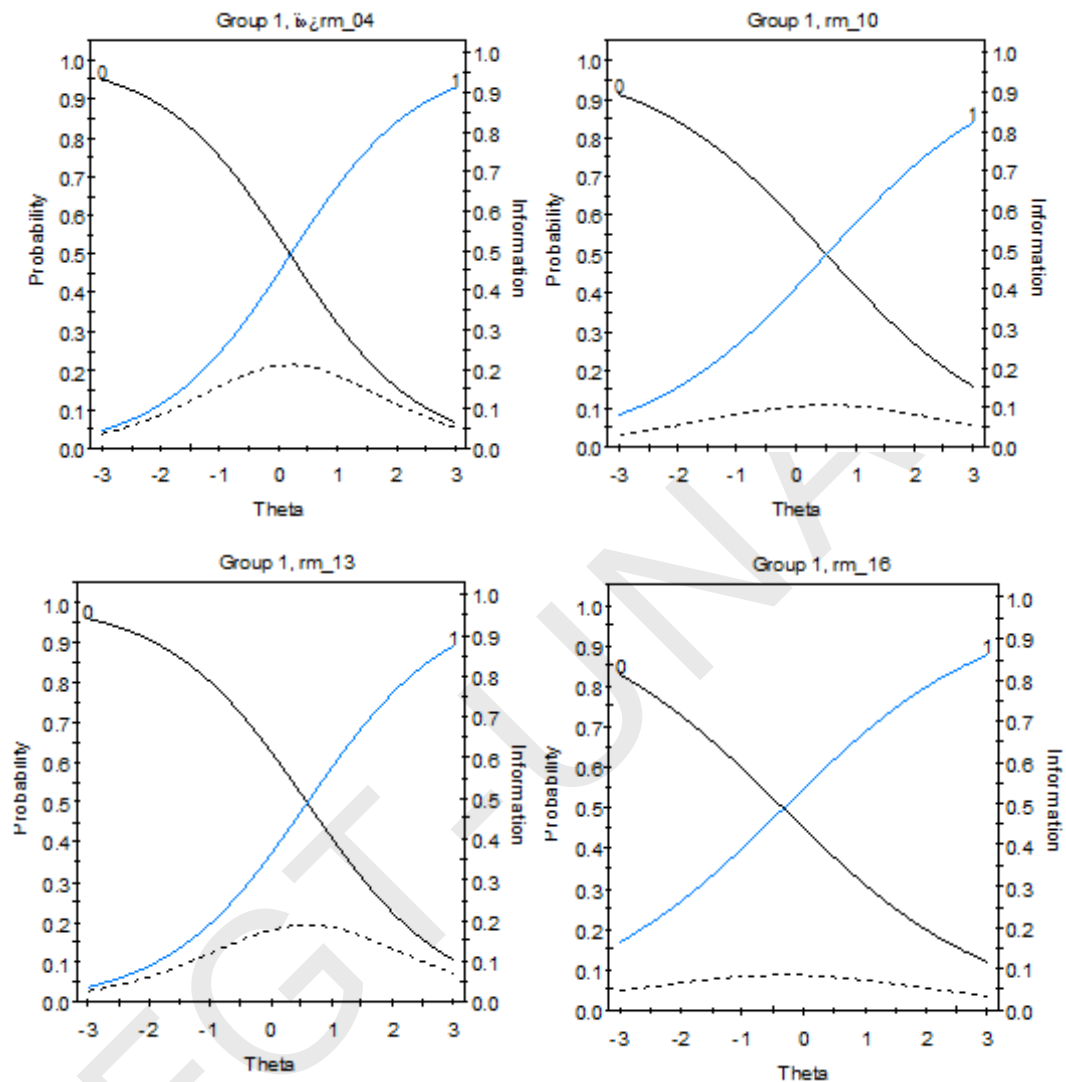
Ítem	Opciones	Proporción de respuestas seleccionadas	Índice de discriminación	Baja aptitud	Alta aptitud
Ítem 1	1	0.15	-0.10	0.18	0.07
	2	0.27	-0.24	0.41	0.17
	3***	0.56	0.41	0.35	0.76
Ítem 2	1	0.01	-0.03	0.03	0.00
	2	0.01	0.00	0.01	0.01
	3***	0.97	0.04	0.95	0.99
Ítem 3	1***	0.26	0.21	0.19	0.40
	2	0.55	-0.07	0.56	0.49
	3	0.18	-0.13	0.24	0.11
Ítem 4	1***	0.46	0.44	0.23	0.67
	2	0.40	-0.01	0.26	0.26
	3	0.29	-0.42	0.49	0.07
Ítem 5	1	0.17	-0.20	0.24	0.04
	2	0.11	-0.21	0.21	0.00
	3***	0.71	0.42	0.54	0.96
Ítem 6	1	0.03	-0.05	0.07	0.01
	2***	0.90	0.17	0.80	0.97
	3	0.06	-0.10	0.11	0.01
Ítem 7	1	0.21	-0.19	0.32	0.13
	2***	0.52	0.34	0.31	0.64
	3	0.27	-0.15	0.37	0.23
Ítem 8	1***	0.42	0.21	0.32	0.53
	2	0.34	-0.17	0.40	0.23
	3	0.22	-0.04	0.26	0.23
Ítem 9	1***	0.58	0.27	0.52	0.79
	2	0.33	-0.15	0.35	0.20
	3	0.07	-0.06	0.08	0.01
Ítem 10	1***	0.43	0.36	0.24	0.60
	2	0.49	-0.22	0.60	0.39
	3	0.07	-0.12	0.13	0.01
Ítem 11	1***	0.47	0.19	0.38	0.57
	2	0.39	-0.05	0.42	0.37
	3	0.13	-0.13	0.19	0.06
Ítem 12	1	0.44	0.04	0.43	0.47
	2***	0.23	0.19	0.15	0.34

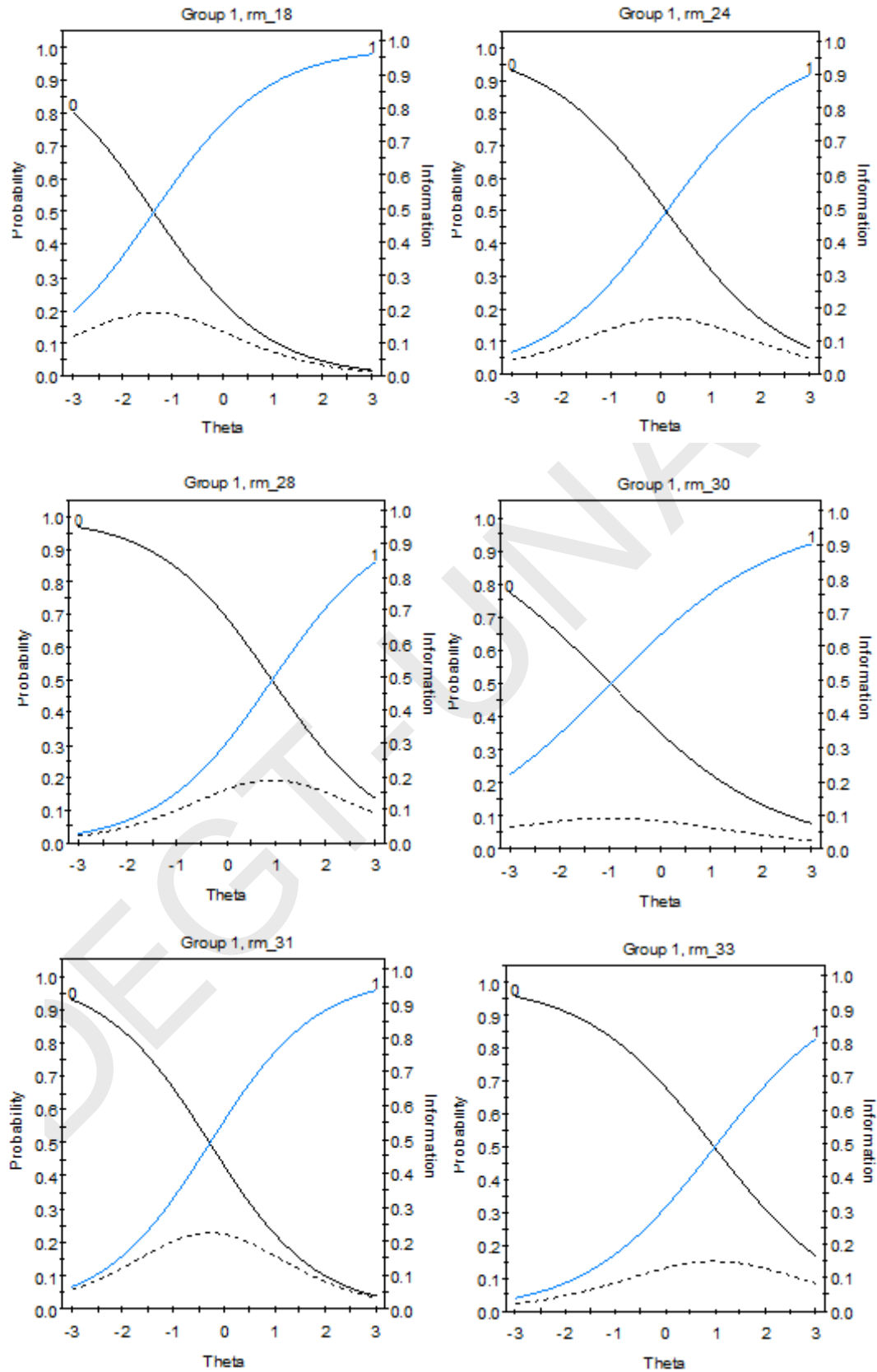
Ítem	Opciones	Proporción de respuestas seleccionadas	Índice de discriminación	Baja aptitud	Alta aptitud
Ítem 13	3	0.31	-0.20	0.38	0.19
	1***	0.39	0.51	0.21	0.71
	2	0.13	-0.15	0.21	0.06
	3	0.47	-0.33	0.56	0.23
Ítem 14	1***	0.13	0.15	0.09	0.24
	2	0.74	-0.09	0.75	0.66
	3	0.13	-0.06	0.16	0.10
Ítem 15	1	0.47	0.01	0.51	0.51
	2***	0.33	0.18	0.21	0.39
	3	0.17	-0.15	0.23	0.09
Ítem 16	1***	0.55	0.40	0.37	0.77
	2	0.14	-0.05	0.15	0.10
	3	0.30	-0.32	0.45	0.13
Ítem 17	1	0.27	0.04	0.26	0.30
	2***	0.62	0.13	0.53	0.66
	3	0.09	-0.11	0.15	0.04
Ítem 18	1***	0.74	0.36	0.54	0.90
	2	0.02	-0.05	0.05	0.00
	3	0.24	-0.31	0.41	0.10
Ítem 19	1***	0.16	0.13	0.11	0.24
	2	0.66	0.08	0.62	0.70
	3	0.14	-0.12	0.18	0.06
Ítem 20	1	0.23	-0.07	0.25	0.19
	2***	0.48	0.05	0.45	0.50
	3	0.29	0.04	0.27	0.31
Ítem 21	1	0.06	-0.13	0.13	0.00
	2	0.12	-0.04	0.11	0.07
	3***	0.81	0.19	0.74	0.93
Ítem 22	1	0.05	0.00	0.04	0.04
	2	0.05	-0.04	0.05	0.01
	3***	0.90	0.06	0.88	0.94
Ítem 23	1***	0.46	0.17	0.34	0.51
	2	0.17	-0.18	0.26	0.09
	3	0.36	0.02	0.38	0.40
Ítem 24	1	0.22	-0.06	0.23	0.17
	2***	0.48	0.43	0.27	0.70
	3	0.30	-0.34	0.47	0.13
Ítem 25	1	0.21	-0.05	0.25	0.20
	2***	0.33	0.35	0.15	0.50
	3	0.45	-0.26	0.56	0.30
Ítem 26	1***	0.59	0.17	0.52	0.69
	2	0.06	0.03	0.04	0.07

Ítem	Opciones	Proporción de respuestas seleccionadas	Índice de discriminación	Baja aptitud	Alta aptitud
	3	0.33	-0.19	0.42	0.23
	1	0.19	-0.18	0.24	0.06
Ítem 27	2***	0.44	0.37	0.29	0.66
	3	0.37	-0.15	0.44	0.29
	1***	0.33	0.37	0.20	0.57
Ítem 28	2	0.41	-0.05	0.37	0.33
	3	0.25	-0.31	0.41	0.10
	1	0.28	-0.20	0.33	0.13
Ítem 29	2***	0.62	0.36	0.49	0.86
	3	0.09	-0.13	0.14	0.01
	1	0.28	-0.18	0.31	0.13
Ítem 30	2***	0.64	0.37	0.51	0.87
	3	0.07	-0.13	0.13	0.00
	1***	0.56	0.36	0.41	0.77
Ítem 31	2	0.32	-0.18	0.38	0.20
	3	0.10	-0.14	0.16	0.03
	1	0.31	-0.20	0.42	0.21
Ítem 32	2	0.08	-0.10	0.10	0.00
	3***	0.60	0.32	0.46	0.79
	1***	0.34	0.40	0.14	0.54
Ítem 33	2	0.12	-0.01	0.11	0.10
	3	0.53	-0.39	0.74	0.34
	1	0.42	-0.04	0.48	0.44
Ítem 34	2	0.25	-0.04	0.21	0.17
	3***	0.3	0.13	0.24	0.37
	1	0.22	-0.15	0.30	0.14
Ítem 36	2	0.19	0.01	0.18	0.19
	3***	0.54	0.20	0.43	0.63
	1	0.42	0.12	0.4	0.51
Ítem 37	2	0.29	-0.02	0.26	0.24
	3***	0.26	-0.03	0.27	0.24
	1***	0.46	0.33	0.29	0.61
Ítem 38	2	0.24	-0.13	0.33	0.20
	3	0.29	-0.17	0.35	0.19
	1***	0.16	0.10	0.10	0.20
Ítem 39	2	0.66	0.15	0.59	0.74
	3	0.16	-0.23	0.29	0.06
	1	0.25	-0.19	0.33	0.14
Ítem 40	2***	0.70	0.31	0.55	0.86
	3	0.04	-0.09	0.09	0.00

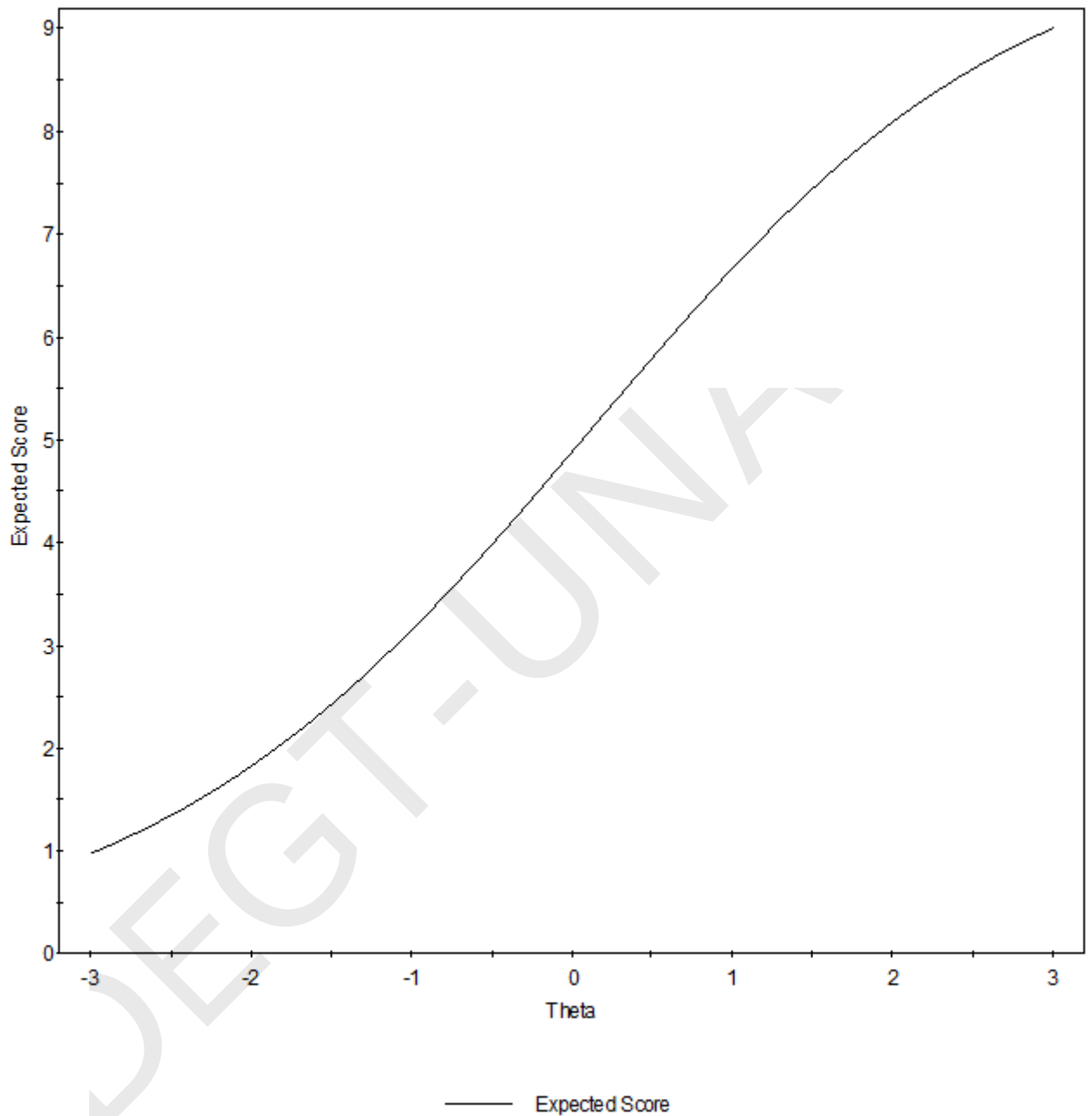
Nota. *** Respuesta correcta. Los valores de bajo desempeño representan los estudiantes ubicados en el quintil 5, los de alto desempeño representan los estudiantes ubicados en el primer quintil.

Apendice 2. Curvas Características de los ítems y su nivel de información





Apéndice 3 . Curva característica del test



Apéndice 4. Curva de información total

