Revista digital
Matemática, Educación e Internet
https://revistas.tec.ac.cr/index.php/matematica
Vol 23, No 2. Marzo, 2023 – Agosto, 2023
ISSN 1659 -0643

Análisis psicométricos de ítems de una prueba diagnóstico sobre estadística descriptiva utilizando el modelo de Rasch

| Psychometric Analysis of Diagnosis Test Items Using the Rasch Model |

Ma. Eugenia Canut Díaz Velarde
marucanut@gmail.com
FES Acatlán-UNAM
México, CDMX

Ingrid Torres-Ramos
ingrid-tr@ciencias.unam.mx
FES Acatlán- UNAM
México, CDMX

Recibido: 2 mayo 2022 Aceptado: 5 diciembre 2022

Resumen: Este trabajo presenta la construcción de un banco de ítems, que permita aplicar y evaluar de manera precisa y equilibrada aptitudes cognitivas (conceptos, razonamiento, identificación, análisis) específicas del dominio de la asignatura de probabilidad y estadística, de la licenciatura de ingeniería civil en una universidad pública. Se invitó a tres profesores como expertos del área de estudio para validar el contenido de los ítems propuestos. Se aplicó un cuestionario de 34 ítems a una muestra de 167 alumnos de tercer semestre en donde las respuestas se codificaron dicotómicamente. Los análisis realizados son la confiabilidad y validez del cuestionario, también se verificó la unidimensionalidad de la escala por medio del análisis exploratorio, y se evaluó la habilidad de los alumnos y la dificultad del ítem analizado por el modelo de Rasch. Dentro de los resultados obtenidos, se eliminaron los ítems que no cumplían con la discriminación y con el modelo; permaneciendo 10 ítems. Se evaluó el índice de separación de los ítems es 4.35 y la fiabilidad de los ítems 0.95; se considera que estos valores son adecuados. Por otro lado, los índices de separación de las personas 1.06 y el índice de fiabilidad de las personas 0.53, no son muy adecuados. Con los parámetros obtenidos se destaca que se deben mejorar las estrategias de enseñanza.

Palabras Clave: Modelo de Rasch, Teoría Clásica del Test (TCT), Teoría de la Respuesta al ítem (TRI), Winstep.

Abstract:

The aim of this study is to do a bank of elements, which allows to apply and evaluate in a precise and balanced way cognitive aptitudes (concepts, reasoning, identification, analysis) specific to the domain of the subject of probability and statistics, of the civil engineering degree in a public university. Three professors were invited as experts in the study area to validate the content of the proposed articles. A 34-item questionnaire was administered to a total number of 167 students in their 2nd year of Civil Engineering where the responses were coded dichotomously. The analyzes carried out are the reliability and validity of the questionnaire, the unidimensionality of the scale was also verified through exploratory analysis, and the ability of the students and the difficulty of the item analyzed by the Rasch model were evaluated. Within the results obtained, the elements that did not comply with the discrimination and with the model were eliminated; remaining 10 items. The separation index for the

items was 4.35 while their reliability was 0.95. Therefore, these values are considered to be adequate. On the other hand, the separation indices of 1.06 people and the reliability index of 0.53 people are not very adequate. With the parameters obtained, it is highlighted that teaching strategies should be improved. Therefore, these values were considered appropriate. Additionally, the separation index for people was 1.06 while the reliability index was 0.53. These values are not really appropriate. The parameters that were obtained underline the importance of improving learning strategies.

Keywords: Rasch Model, Classical Test Theory (CTT), Item Response Theory (IRT), Winsteps.

1. Introducción

En una universidad pública de México se imparte la carrera de ingeniería civil y en su programa de estudios incluye la asignatura de probabilidad y estadística, ubicada en el tercer semestre, en ella se observó que los alumnos presentan una baja acreditación de la materia, por lo que se preguntó si los exámenes tienen una medición correcta del objetivo general. En la universidad pública por lo general se realizan exámenes parciales que son realizados por cada profesor que imparte la materia, sin que se tenga un consenso de los temas revisados lo que ocasiona que el estudiante le dé importancia a los temas de acuerdo con el punto de vista del profesor, y no respecto a los temas propuestos de la asignatura, por lo que no se generaliza un dominio real de los conceptos. En el ámbito universitario el examen es el instrumento más utilizado para medir conocimientos curriculares, lo que conlleva a que el profesor de la materia es quién establece los criterios y procedimientos de la calificación. De modo que la asignación de puntaciones está sujeta a fuentes de variabilidad que no siempre son atribuibles al nivel de competencia de los alumnos, [27].

Esto motivo el interés por desarrollar herramientas de medición calibradas que aporten objetividad a la evaluación a fin de proporcionar una estimación adecuada del nivel de dominio con el fin de ubicar a cada uno de los estudiantes en un nivel de acuerdo con la calificación obtenida. Dependiendo del nivel que obtengan se pueden hacer recomendaciones de talleres que cubran las necesidades de la asignatura.

Los resultados de este tipo de clasificación son fuente de información importante para la carrera, dado que los profesores pueden afinar el contenido de la materia en el aula y sirve de orientación a los estudiantes, con prioridad y énfasis en las debilidades detectadas.

La evaluación de los conocimientos es un elemento constitutivo del proceso de enseñanza-aprendizaje, genera evidencia de las habilidades y conocimientos de la asignatura de acuerdo con los objetivos establecidos en un programa, de tal forma que permite tomar decisiones en el contexto educativo al contar con información confiable. De acuerdo con Scriven [29], la evaluación es el acto o proceso cognitivo por el cual se establece una afirmación respecto de la calidad, valor o importancia de cierta entidad. Sin embargo, la medición es asignar números y evaluar es hacer un juicio integral de las cualidades de un objeto de interés. Es así, que tanto la evaluación como la medición son constructos complementarios, debido a que el resultado obtenido de la medición es un insumo de la evaluación y un juicio evaluativo que permite dar sentido y significado al dato de una medición. Se tienen dos tipos de evaluaciones educativas, identificadas: la primera es llamada a gran escala la cual es diseñada y administrada por el sistema educativo cuyo objetivo es la información válida y confiable para la política educativa de un distrito, región o país; la segunda es la evaluación del aula que es suministrada por el profesor que construye los reactivos, sitúa el ejercicio en el momento que es necesario, asigna los puntos o la calificación, decide como informar y usar la información.

Para la construcción del ítem se requiere tener evidencias de la validez de contenido que se fundamentan y respaldan las interpretaciones que se tienen de las puntuaciones del test. El propósito de esta sección es el análisis de la variable y las dimensiones que lo componen para ser medidos empíricamente. Por ello que se hace necesario realizar un juicio de expertos que genere una tabla de especi-

ficaciones que ayuda a delimitar y definir el dominio de conocimiento del instrumento de evaluación. La tabla de especificaciones logra la planificación sistemática, otorga orden y da orientación en la selección de los contenidos que constituyan una muestra representativa de los aprendizajes relevantes. Para definir el contenido de la variable y sus dimensiones se analizó el programa de la asignatura, en el que se hizo una revisión de los contenidos principales. Los expertos evaluaron la representatividad de la información para cada dominio.

La construcción de los ítems se realizó mediante juicio de expertos con el apoyo de la tabla de especificaciones basada en el programa de asignatura sobre el tema de estadística descriptiva utilizando los dominios: clasificación de variables, niveles de medición, distribución de frecuencias, tipos de gráficas, medidas de tendencia central y medidas de variabilidad.

De acuerdo a lo expuesto anteriormente, el objetivo de esta investigación es contar con un banco de ítems, que apoya la construcción de test basado en la necesidad de la evaluación de manera precisa y equilibrada en las aptitudes cognitivas (conceptos, razonamiento, identificación y análisis). Se utilizó el modelo de Rasch con el fin de analizar las propiedades psicométricas de los reactivos y detectar la capacidad de los ítems en medir diferentes niveles de habilidad, que proporciona un análisis detallado de los patrones de las respuestas individuales de los estudiantes que refleja los procesos de razonamiento.

El modelo de Rasch, ha sido utilizado para evaluar pruebas estandarizadas a nivel internacional como es el caso la prueba de PISA por su nombre en inglés Programm for International Student Assessment, desarrollado por la Organización para la Cooperación y el Desarrollo Económico (OCDE). A continuación se darán algunas aplicaciones del modelo de Rasch en distintas áreas de investigación: en [28] realizan una investigación para el área de educación en la construcción y validación de una prueba de comprensión de lectura; en [14] se lleva a cabo la calibración de una prueba de química, en el que se logró un buen ajuste del modelo con 10 reactivos de un total de 12 para una muestra de 219 alumnos, mostrando que la prueba tiene diferentes índices de dificultad; en [8] se estudia la adaptación de un grupo de reactivos y la conformación de un banco de ítems (BI) que permita evaluar de manera precisa y objetiva algunas aptitudes cognitivas específicas (razonamiento verbal, numérico y espacial) y generar un indicador general de inteligencia; en [9] realizaron una investigación acerca de la evaluación de las propiedades de medición del inventario de discapacidad auditiva para ancianos (HHIE), utilizando el análisis de Rasch para 25 ítems acompañado con una escala de respuesta de 3 punto, aplicado a 380 adultos con pérdida auditiva. Dentro de sus resultados encontraron una alta confiabilidad de separación de personas, sin embargo, la escala mostró un mal ajuste al modelo de Rasch y no fue unidimensional. En general, se pueden encontrar diversas aplicaciones del modelo de Rasch en la literatura científica.

2. Validez y confiabilidad

Los ítems son la base con lo que se construye el test, cada uno de ellos ha de coadyuvar a que el test constituya un buen instrumento de medida, es decir, que sea confiable y válido. Por ello los análisis de los ítems que son propuestos están en conexión con estas dos propiedades psicométricas básicas que son fundamentales de una buena medición, [1, 15, 23].

El concepto tradicional de validez referencia a una tautología, al indicar *un instrumento es válido si mide lo que con él se pretende medir*. A partir del final del siglo XX, Messick [17, 18] define la validez como el grado de propiedad de las inferencias e interpretaciones derivadas de los puntajes de los test, incluyendo las consecuencias sociales que se derivan de la aplicación del instrumento. También, considera que es un concepto unitario que recolecta diferentes tipos de evidencias de contenido, predictivas y constructo que se usan de acuerdo con los propósitos y usos de los instrumentos, todas esas evidencias contribuyen a la validez de constructo.

Nunnally y Bernstein en [23] señalan que un instrumento es confiable si al ser aplicado en las mismas condiciones a los mismos sujetos se producen los mismos resultados con respecto a precisión, consistencia y estabilidad en repeticiones. Uno de los indicadores que con mayor frecuencia se utilizan en psicométrica es el índice de discriminación, que es medido por la correlación ítem-total y el Alfa de Cronbach en la Teoría Clásica del Test (TCT), la cantidad de error de medición, y el tamaño de la función de información en la Teoría de la Respuesta Ítem (TRI) y el modelo de Rasch, [15, 21, 25].

2.1. Análisis de ítems

Existen diferentes tipos de métodos y modelos que se utilizan para analizar los ítems que conforman al test, tales como:

- Análisis de factores exploratorio y confirmatorio
- Teoría Clásica de los Tests (TCT)
- Teoría de Respuesta a los Ítems (TRI)
- Modelo de Rasch
- Teoría G (Generalizabilidad)

En el enfoque clásico uno de los aspectos esenciales es estudiar los parámetros de un test, se inicia por analizar los ítems como elementos que deben estar en conformidad para conseguir el resultado final buscado. Al constructor del test le corresponde escoger los mejores ítems de entre un conjunto de ellos mucho mayor del que sería necesario para que el test final resulte adecuado. Es decir, el análisis de los ítems permite al psicómetra decidir cuales ítems son pertinentes y cuáles no, en función de la finalidad y objetivo de medida del test total.

El análisis de los ítems depende del modelo teórico a partir del cual se hubiese construido el test, [3]. Bajo la perspectiva de la TCT se analizan las características (dificultad, discriminación, análisis de los distractores, fiabilidad, validez y dimensionalidad) más relevantes de los ítems que afectan las propiedades del test, [21].

3. Modelos de la teoría de respuesta a los ítems

La TCT y TRI tienen como objetivo estimar el error que cometemos al medir ciertas variables de naturaleza psicológica. Ello lo consiguen mediante la formulación de un modelo matemático que, como todo modelo se asienta en una serie de hipótesis. De acuerdo con Muñiz [20] la TRI surge como un nuevo enfoque en la teoría de las pruebas que permite superar algunas de las limitaciones de la TCT.

El modelo de TCT y los de TRI son funciones matemáticas que pretenden explicar y predecir, las respuestas de las personas a un test. La TCT explica la puntuación observada en un test (en el sistema real) como la suma de la puntuación verdadera más el error de medida ($X = V + \epsilon$). El objetivo principal de la TRI, como en la TCT, seguirá siendo estimar el verdadero nivel de habilidad del examinado.

Existen dos diferencias esenciales entre ambos modelos: 1) la unidad de análisis en el modelo clásico es el test (X es la puntuación observada en el test), en tanto que la unidad de análisis en la TRI es el ítem; 2) la TRI incorpora términos al modelo que describen las características de los ítems, es decir, las respuestas de los examinados a los ítems van a estar explicadas no solo por su nivel de habilidad,

sino por las características psicométricas de los ítems. Estas dos diferencias proporcionarán ventajas a la TRI sobre la TCT en el análisis de los ítems y en la construcción de test.

Para Muñiz, [20], los modelos de TRI asumen que los datos sobre los que se aplicarán los modelos, esto es, las respuestas a los test cumplen dos supuestos: independencia local y unidimensionalidad.

El supuesto de unidimensionalidad señala que la respuesta del examinado al ítem este determinado por una única variable, denominada generalmente como rasgo latente. Por su parte el supuesto de independencia local exige que la probabilidad de responder correctamente a un ítem es independiente de la probabilidad de responder correctamente a cualquier otro.

La TRI establece la relación que existe entre la escala de aptitud o habilidad de los sujetos evaluados, y la probabilidad de acertar correctamente un ítem. El modelo TRI se representa por medio de la función logística de la Curva Característica de los Ítems (CCI).

$$\begin{aligned} P_i(\theta_s) &= c_i + (1 - c_i)\frac{e^{Da_i(\theta_s - b_i)}}{1 + e^{Da_i(\theta_s - b_i)}} \\ &= c_i + \frac{1 - c_i}{1 + e^{-Da_i(\theta_s - b_i)}} \qquad \text{con} \quad i = 1, \ldots, n \end{aligned} \tag{1}$$

donde $P_i(\theta_s)$ es la probabilidad de que la persona s responda correctamente el ítem i, $D > 0$ es una constante, usualmente se le asigna el valor de 1.7 (para buscar semejanza con la función de distribución normal); θ_s es el valor del constructo o rasgo que se desea estimar en cada examinado, a_i es el parámetro de discriminación, b_i es el parámetro de dificultad y c_i es la probabilidad de acertar el ítem.

4. Modelo de Rasch

El análisis de Rasch es un modelo propuesto por el matemático danés del mismo nombre en el año 1960, se aplicó por primera vez para la medición de la inteligencia de los soldados daneses y ha sido utilizado de manera muy extensa en distintas áreas para medir resultados educativos [16] y otros fenómenos en el ámbito económico. Así como en áreas de inteligencia, capacidades y rasgos personales no observables directamente (lo que se considera variable latente). Este tipo de variables son medidas a través de las respuestas de los individuos ante distintas preguntas formuladas en un test (ítems). Wilson [32], señala que este modelo es un referente para la construcción de un instrumento que ayuda a medir la variable de interés.

Rasch como modelo de análisis establece la probabilidad de respuesta de una persona ante un estímulo dado, en términos de la diferencia entre la medida del rasgo de una persona y la medida del estímulo utilizado. Se trata de un modelo estocástico (no determinista) donde la medida del rasgo de una persona y la medida del estímulo aplicado, quedan ubicadas en una misma escala lineal con un origen común. La variable de interés es la diferencia de ambas medidas, esto es, que se establece la medida del rasgo de la persona de manera independiente del conjunto de personas a las que se somete. El rigor es la diferencia de medida de rasgo y medida de estímulo que es independiente del instrumento o de la población. Por último, el modelo requiere que la variable sea unidimensional, ordenada e inclusiva [31], a partir de una serie de puntuaciones obtenidas para distintos ítems por diferentes individuos.

Los ítems que componen al test son evaluados de acuerdo con las propiedades psicométricas de un instrumento en relación con las propiedades específicas de cada ítem que componen al test, [19]. Las puntuaciones que se obtienen de las pruebas vienen dadas en función de los ítems y de las personas que contestan los mismos. Es así como el modelo da información acerca de la capacidad de una persona porque se centra en la dificultad de los ítems, más que la cantidad de ítems que son contestados correctamente por cada participante. Por ello, se refiere a que la habilidad de una persona queda

establecida al interactuar con la dificultad del ítem y obtener una puntuación para cada sujeto en la medida, [12].

Los fundamentos del modelo de Rasch son:

- El atributo que se desea medir puede representarse en una única dimensión en la que se situarían conjuntamente las personas y los ítems.
- El nivel de la persona en el atributo (habilidad) y la dificultad del ítem determinan la probabilidad de que la respuesta sea correcta.

Este modelo supone que la probabilidad de acertar el ítem es cero, $c_i = 0$, el parámetro de discriminación es la constante uno para todos los ítems, $a_i = 1$ y $D = 1$. Aplicando estos supuestos a la ecuación 1 y despejando el termino $\theta_s - b_i$ se tiene

$$\theta_s - b_i = ln\left(\frac{P_i(\theta_s)}{1 - P_i(\theta_s)}\right) \qquad \text{con} \qquad i = 1, \ldots, n$$

donde $P_i(\theta_s)$ es la probabilidad de que la persona s responda correctamente el ítem i, θ_s es el nivel de habilidad (conocimiento) de la persona s, y b_i es el nivel de dificultad del ítem i. Expresado en palabras, la ecuación indica que la probabilidad de una respuesta correcta es una función de la diferencia en el atributo entre el nivel de la persona, θ_s, y el nivel de dificultad del ítem, b_i.

Así, cuando una persona responde a un ítem equivalente a su umbral de competencia, tendrá la misma probabilidad de una respuesta correcta y de una respuesta incorrecta. En este caso, la dificultad del ítem es equivalente al nivel de competencia de la persona, $\theta_s - b_i = 0$. Si la competencia del sujeto es mayor que la requerida por el ítem, $\theta_s - b_i > 0$, la probabilidad de una respuesta correcta será mayor que la de una respuesta incorrecta. Por el contrario, si la competencia del sujeto es menor que la requerida por el ítem, $\theta_s - b_i < 0$, la probabilidad de una respuesta correcta será menor que la de una respuesta incorrecta.

El Modelo de Rasch requiere que los ítems tengan un valor constante e igual en el parámetro de discriminación, $a_i = 1$, es decir que es igual para todos los ítems. En [10], se señala que en este modelo los ítems deben encontrarse en datos intervalares, para su análisis y evaluar así varias características como: el nivel de ajuste del modelo, la dificultad y el orden jerárquico de los ítems, la fiabilidad de las personas e ítem, los índices de separación y el funcionamiento diferencial del ítem (DIF, por sus siglas en inglés). En [24], se comenta que los datos empíricos deben sujetarse al modelo propuesto, para evaluar estas características. Por ello el ajuste del ítem se refiere a que tan bien un reactivo mide el constructo de interés como se menciona en [4] y se cuantifica mediante medidas de infit y outfit, lo que permite asegurar que el instrumento pueda evaluar de forma correcta el constructo que pretende. De ahí que los parámetros estadísticos permiten identificar el grado de relación que existe entre el patrón de respuestas observadas y las expectativas establecidas por el modelo. De tal forma que los índices determinan si los parámetros estimados de los ítems pueden ser considerados como un resumen del patrón de repuesta observado.

La jerarquización de los ítems consiste en el ordenamiento de los mismos en niveles de dificultad (del más fácil al más difícil). Este ordenamiento de ítems es un principio fundamental de la edición dado que nos permite determinar si un alumno posee mayor o menor habilidad con respecto a otro, [4]. Si el ítem no está en escala logit y ordenado de manera jerárquica, las puntuaciones obtenidas por un alumno en el test pueden ser engañosas. El contar con un orden jerárquico de los ítems nos permite identificar ítems redundantes o niveles de dificultad no cubiertos que disminuyen la precisión y la eficacia del instrumento. La precisión de la medida y de los índices de separación de personas

depende de que tan bien los ítems del instrumento permitan diferenciar los niveles de habilidad y de lo bien que el instrumento puede diferenciar a las personas en la medida.

El análisis de Rasch ofrece estadísticas de fiabilidad y separación para los ítems y las personas. En [13], Linacre señala que la fiabilidad significa el grado de reproducibilidad de las habilidades relativas o de las dificultades estimadas. El hecho que se tenga un índice alto en fiabilidad para personas nos indica que existe una alta probabilidad de que las personas identificadas por el test tengan alta habilidad y que existen otras que no las tienen. De manera semejante, alta fiabilidad en los ítems significa que los ítems establecidos como de alta dificultad tienen realmente alta dificultad.

El índice de separación indica el número de diferentes estratos de rendimiento que la prueba puede identificar [33]. El DIF puede ser conceptualizado como el hecho de que la respuesta a un ítem está sujeta a cambios en función de diferentes grupos de personas, [7]. En otras palabras, un ítem presenta DIF cuando la probabilidad de respuesta correcta no depende únicamente del nivel de la persona en el rasgo intencionadamente medido por el test, [4].

5. Metodología

Es una investigación referida a estudios exploratorios, en el que se pretende examinar el nivel de conocimientos de alumnos de la materia probabilidad y estadística de la licenciatura de ingeniería civil, en búsqueda de evidencias teóricas y empíricas de confiabilidad y validez del instrumento.

- Primera fase: construcción ítems.

La validación de ítems se realizó mediante juicio de expertos, en el que se utilizó una tabla de especificaciones, a partir del programa de la asignatura de la materia de probabilidad y estadística como marco de referencia para una evaluación, se espera que a través de ésta se pueda obtener una clara definición del constructo u objeto de medida. Esto permite responder claramente a la pregunta de qué se debe evaluar, y nos guía a la pregunta del cómo hacerlo.

Se definió el constructo u objeto de medida, es decir, puso en práctica su significado con el objetivo de realizar el proceso de evaluación. Esto significa que se puedan tomar decisiones sobre el modo en que será entendido y observado en la situación evaluativa, y delimitar la extensión que se espera abarcar con la evaluación. Para lograr esto se trabajó con un grupo de tres especialistas en la materia, quienes a partir del conocimiento de la materia y de la población describieron los contenidos y procesos representados en los ítems. El proceso sólo se realizó para la parte de estadística descriptiva.

Estas decisiones formulan el marco de especificaciones, que funcionará como puente entre el referente y el instrumento o dispositivo con el que se levantará la información acerca de los alumnos evaluados. En la tabla de especificaciones, se consideraron los dominios conceptuales en el siguiente orden: escalas de medida, distribución de frecuencias, tipos de gráficas, medidas de tendencia central y medidas de variabilidad. Los tres objetivos de aprendizajes indicados son: comprensión (capacidad de adquirir, reflexionar, identificar la información referente al tema, es decir, se busca una comprensión básica de hechos); saber o reproducción (aplicación en el que se hace uso del conocimiento, permite resolver problemas mediante la aplicación de una secuencia de acciones); saber hacer o aplicación (examina con detalle y descompone la información en partes identificando los motivos o causas).

Los 34 ítems propuestos que conforman el instrumento se organizaron considerando que el bloque de comprensión tiene una representatividad del 20 % con un total de 7 ítems, el bloque de aplicación contiene 45 % con 15 ítems y el último bloque de análisis abarca un 35 % con 12 ítems con lo que se obtuvo un total de 34 ítems propuestos.

Los indicadores utilizados son:

1. Identificar en una variable si es numérica o categórica.
2. Si la variable es numérica identificar si es discreta o continua.
3. En una variable categórica (métrica) tipos de operaciones básicas a utilizar.
4. Clasificación de una variable en escalas de medida (nominal, jerárquico, intervalos y razón).
5. Identificar tipos de gráficos para variables categóricas y numéricas.
6. Pasos para realizar una tabla de distribución de frecuencias.
7. Identificar las medidas de tendencia central en datos agrupado y no agrupados.

- Segunda fase: estructura de la prueba.

La prueba se estructuró con base al contenido especificado con reactivos de opción múltiple, es decir, se plantea una pregunta (problema) con 4 alternativas de respuestas, dónde sólo una es correcta. Se utilizaron 34 ítems con este formato de pregunta.

- Tercera fase: aplicación de la prueba.

Evaluación de conocimientos, se decidió realizar la prueba a la mitad del semestre, momento en el que los cuatro grupos tienen estudiados el tema a evaluar, considerando el programa de estudio de la materia. El cuestionario se aplicó en línea utilizando la plataforma de Moodle, en la que se abrió un espacio dedicado al proceso de evaluación con un total de 167 alumnos, cada estudiante contaba con dos horas para contestar.

- Cuarta fase: recolección y preparación de la información.

Se llevó a cabo la recolección y preparación de la información obtenida del cuestionario. Los datos recogidos se revisaron y ordenaron para ser examinados en los programas de SPSS 26 (en donde se analizó la confiabilidad y validez) y Winsteps (para estudiar el modelo de Rasch).

- Quinta fase: análisis de información.

Se realizó el análisis de la información y se dio una interpretación.

6. Resultados

El índice de dificultad de un ítem es un indicador de la dificultad de este. La clasificación de la dificultad de los ítems se presenta en la Tabla 1.

El coeficiente de fiabilidad es un indicador global de la precisión con el que, el test está midiendo una determinada variable. Evalúa en que grado los ítems de un test convergen, es decir que están interrelacionados, se refiere a la consistencia interna, se mide a través del Coeficiente alfa de Cronbach se obtuvo al utilizar el SPSS 26, a los 34 ítems propuestos, Tabla 2.

La clasificación del índice de discriminación del ítem, de acuerdo al TCT, se obtuvo cuales ítem distinguen entre los estudiantes que conocen sobre el tema y el que no, Tabla 3.

Tabla 1: Índice de dificultad del ítem. Elaboración propia basada en el análisis computacional.

Ítem	*Rango*	*Categoría*
P1; P6; P19; P27; P30; P31; P32	ID < 0.25	Muy difícil
P4; P10; P13; P15; P16; P20; P21; P23; P26; P28; P29; P33	$0.25 <$ ID < 45	Difícil
P3; P8; P24; P25; P26	$0.45 <$ ID < 0.55	Normal
P2; P7; P9; P11; P14; P18; P22	$0.55 <$ ID < 0.75	Fácil
P5; P12; P17; P34	ID > 0.75	Muy Fácil

Tabla 2: Estadísticas de fiabilidad. Elaboración propia basada en el análisis computacional.

Alfa de Cronbach	*Número de elementos*
.667	34

Tabla 3: Índice de discriminación del ítem. Elaboración propia basada en el análisis computacional.

Ítems	*Valores*	*Interpretación*
P1; P9; P12; P19; P20; P29; P33; P34; P31	Menores que 0.10	Ítem carece de utilidad para discriminar
P7; P15; P17; P18; P22; P23	$0.10 - 0.19$	Ítem límite, se debe mejorar
P2; P4; P5; P10; P13; P14; P16; P27; P28; P30; P32	$0.20 - 0.29$	Ítem discrimina poco
P3; P6; P11; P21; P24; P25; P26	$0.30 - 0.39$	Ítem discrimina bien
P8	Mayor o igual que 0.40	Ítem discrimina muy bien

Tabla 4: Estadísticas de fiabilidad. Elaboración propia basada en el análisis computacional.

Alfa de Cronbach	*Número de elementos*
.740	22

En algunos de los ítems, la correlación con el total del test (índice de discriminación) era baja, por lo que se decidió eliminarlos, obteniendo así alfa de Cronbach de .740 con 22 ítems, Tabla 4.

Se continuó con el análisis factorial exploratorio, con el propósito de tener evidencias de la validez asociada a la estructura factorial y analizar en que grado razonable, se cumple el supuesto de unidimensionalidad.

Se utilizo el programa SPSS 26, para realizar un Análisis factorial utilizando la extracción de componentes principales (ACP), y con rotación de varimax. El índice de medida de adecuación muestral Kaiser-Meyer-Olkin (KMO) obtenido fue de .683 con un valor de significancia de .00, por que se procedió a realizar el análisis factorial.

En la Tabla 5, se presenta la varianza total explicada. Se observa que se tiene aproximadamente un

22.658 % de la varianza total es explicada por el primer componente. El segundo componente tiene sólo un 10.970 %, el tercero 10.261 %.

Tabla 5: Varianza total explicada. Elaboración propia basada en el análisis computacional.

	Valores propios iniciales			*Sumas de cargas al cuadrado de la extracción*		
Componente	*Total*	*Varianza (%)*	*Acumulado (%)*	*Total*	*Varianza (%)*	*Acumulado (%)*
1	2.492	22.658	22.658	2.492	22.658	22.658
2	1.207	10.970	33.628			
3	1.129	10.261	43.889			
4	1.055	9.594	53.483			
5	.936	8.511	61.994			
6	.893	8.114	70.108			
7	.841	7.649	77.756			
8	.760	6.905	84.661			
9	.667	6.062	90.723			
10	.562	5.112	95.835			
11	.458	4.165	100			

En la Figura 1, se presenta el gráfico de sedimentación del instrumento. Se puede observar que con este criterio el número de factores se representan por el punto en el que se presenta un cambio importante en la trayectoria de la caída de la pendiente. Cattell, citado en [6], expone que se consideren todos aquellos factores situados antes de este punto. En el gráfico se observa la existencia de un solo componente.

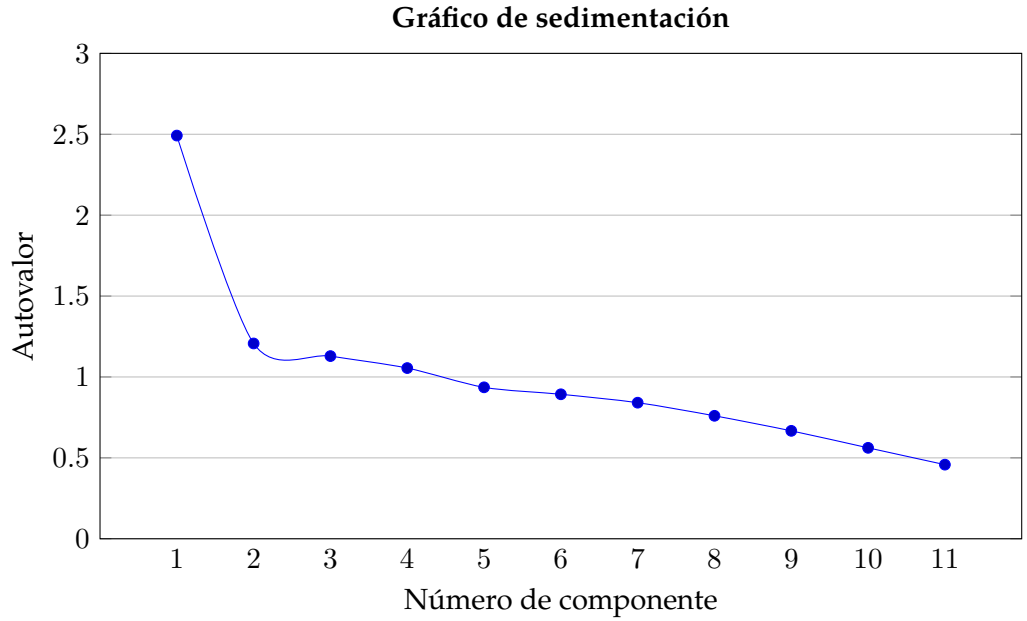


Figura 1: Gráfico de sedimentación. Elaboración propia basada en el análisis computacional.

7. Aplicación del modelo de Rasch

El estudio se realizó a partir de las respuestas obtenidas de 34 ítems aplicados a 167 estudiantes. Se anularon los ítems de las personas que no cumplían con las expectativas del modelo, obtenido un

total de 10 ítem y 168 individuos. Realizando nuevamente el análisis computacional, Winsteps, a la base obtenida. En la Tabla 6, se presentan las estadísticas de confiabilidad tanto para personas como para los ítems obtenidos aplicando el modelo de Rasch, de acuerdo con el modelo, la medida de confiabilidad de los examinados indica que tan consistentes son los resultados, es decir, si al mismo grupo de examinados se les aplicara otro conjunto de ítems del mismo universo, al que pertenece el conjunto que se analiza, se obtendrían los mismos resultados. Para el examen de estadística 2020, la confiabilidad de las personas fue 0.53 que es un valor que se considera bajo para la prueba de diagnóstico.

En cuanto a la confiabilidad de los ítems, indica que tan consistentes son las estimaciones del parámetro de dificultad en el mismo conjunto de ítems. Para este caso, el valor de la confiabilidad de los ítems es de .95, lo cual indica que las estimaciones de Rasch son muy consistentes.

Tabla 6: Estimación de Rasch. Elaboración propia basada en el análisis computacional.

Número de		*Índice de confiabilidad de*		*Índice de separación de*	
casos	*ítems*	*personas*	*ítems*	*personas*	*ítems*
167	10	0.53	0.95	1.06	4.35

En la Figura 2, se observa el análisis de ítems obtenido en el modelo de Rasch, se tiene como medida el índice Infit MNSQ que se calcula con las medias cuadráticas sin estandarización, Wright y Linacre en [34] proponen como valores aceptables de Infit MNSQ los valores ubicados entre 0.8 y 1.2.

En la columna Infit ZSTD se aprecian los valores estandarizados, estadígrafo de media cuadrática de los residuales representados en logaritmo natural. Los valores que toma se ubican entre el rango -2 a +2, los valores están en el intervalo de lógitos aceptable para determinar ajuste razonable tanto en examinados como en personas, se puede observar que los datos están dentro del intervalo.

ENTRY NUMBER	TOTAL SCORE	TOTAL COUNT	MEASURE	MODEL S.E.	INFIT MNSQ	INFIT ZSTD	OUTFIT MNSQ	OUTFIT ZSTD	PTMEASUR-AL CORR.	PTMEASUR-AL EXP.	EXACT MATCH OBS%	EXACT MATCH EXP%	ITEM
1	114	167	-1.43	.19	.99	-.07	1.21	1.39	A .40	.42	78.4	73.5	P2
10	73	167	-.13	.18	1.08	1.15	1.13	1.25	B .40	.46	68.5	69.6	P28
9	33	167	1.34	.22	1.05	.43	.89	-.41	C .43	.44	80.2	82.7	P27
6	62	167	.22	.18	1.04	.58	.98	-.12	D .44	.46	67.9	72.1	P13
4	64	167	.15	.18	.99	-.07	1.02	.22	E .46	.46	71.6	71.6	P10
5	108	167	-1.22	.18	1.00	.01	.95	-.36	e .44	.43	72.2	71.8	P11
3	36	167	1.20	.21	.98	-.15	.90	-.43	d .46	.44	82.1	81.2	P6
2	81	167	-.38	.18	.97	-.39	.94	-.63	c .48	.46	71.0	68.0	P3
8	74	167	-.16	.18	.96	-.55	.96	-.41	b .48	.46	72.8	69.4	P26
7	56	167	.42	.19	.92	-.91	.90	-.76	a .51	.46	76.5	73.6	P21
MEAN	70.1	167.0	.00	.19	1.00	.0	.99	.0			74.1	73.3	
P.SD	25.2	.0	.85	.01	.04	.6	.10	.7			4.7	4.6	

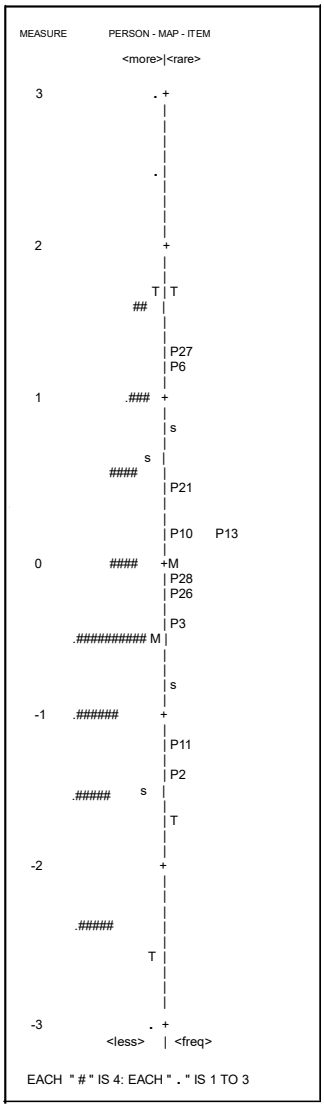
Figura 2: Estadística del ítem. Elaboración propia resultante del análisis computacional.

El estadígrafo outfit, es una mediada cuadrática no ponderada sensible a los comportamientos extremos no esperados en los patrones de respuesta. Se mide en los mismos términos del infit, como se observa outfit MNSQ, el item P2 se encuentra en el límite del valor 1.2 pero outfit ZSTD, se encuentra entre el intervalo de -2 a +2. Measure, es la calibración de la dificultad del reactivo expresada en lógitos. El ítem P27 resultó ser el más difícil, con una dificultad de 1.34, en escala logit, seguido de P6 con dificultad de 1.20 y el ítem más fácil es P2 con valor de -1.43 en la escala logit.

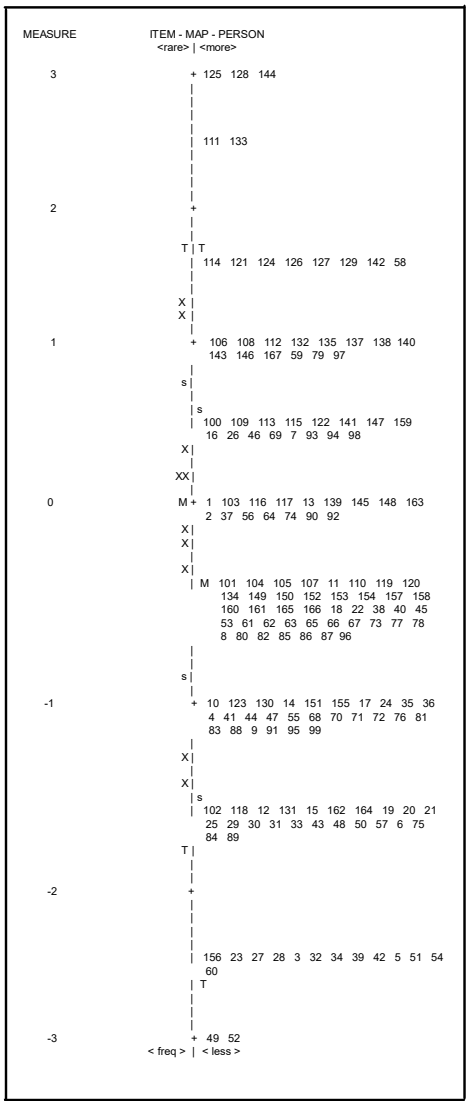
Estos mismos datos se pueden observar en la Figura 3a. Este mapa se compone de dos grandes histogramas (graficados de forma vertical). El lado izquierdo presenta el histograma de la distribución de habilidades de las personas. El lado derecho presenta el histograma de la distribución de dificultades de los ítems. Los dos histogramas comparten el eje de valores (línea quebrada vertical), genéricamente los resultados se presentan en lógitos, [5]. Distribución de personas o el mapa de personas (lado

izquierdo) presenta las puntuaciones ordenadas de forma jerárquica. En la parte superior se ubican las personas de mayor habilidad y en la parte inferior se ubican las personas de menor habilidad. Distribución de ítems. El mapa de ítems (lado derecho) presenta las dificultades de forma jerárquica. En este mapa se debe tener una consideración adicional, los ítems P27 y P6 son los más difíciles y los ítems P2 y P11 son los más fáciles, en el promedio (M) tenemos los ítems P28, P26, P3. Los ítems que están a una desviación estándar por arriba del promedio es P21, P10 junto P13. Los ítems que se ubican uno al lado del otro, como es el caso de los ítems P10 y P13 en la Figura 1, son ítems que tienen similar grado de dificultad. Estos ítems proporcionan poca información adicional en relación al desempeño de las personas. El mapa de ítems permite interpretaciones basadas en criterio de dificultad.

En la Figura 3a, se representa el mapa de las personas y su ubicación de acuerdo a la escala de logitos en la que se puede ver las personas con nivel de habilidad más alto, ubicadas en + 3 logitos y son los que tienen la probabilidad de contestar correctamente todos los ítems, que corresponde a las personas 125, 128 y 144, es decir que existe que hay 1.79 % de las personas que poseen una habilidad superior a 1.34 lógitos del ítem que resultó ser el más difícil, y los que se encontraron en el nivel de habilidad (θ) más bajo, que tienen una alta probabilidad de fallar todos los ítems son las personas 49 y 52 ubicados en - 3 lógitos. Las personas que se ubicaron en una escala promedio (M) con $\theta = 0$ son 9.58 % del total de alumnos.



(**a**) Distribución de personas-ítems



(**b**) Distribución de ítems-personas

Figura 3: Mapa de la distribución de la persona y el ítems. Elaboración propia resultante del análisis computacional.

8. Conclusiones y recomendaciones

Se evaluó las propiedades psicométricas de un instrumento de medida para el área de estadística de la materia Probabilidad y Estadística de la carrera de Ingeniería Civil, en el que participaron los profesores que imparten la materia, con la idea de lograr un instrumento que sea justo, en el grado de dificultad respecto al tipo de conocimiento que se requiere; es decir que se mida la habilidad del alumno ubicado en una escala que permita tener una comparación entre la habilidad del alumno y la dificultad de los ítem utilizados a la hora de emitir una respuesta de los mismos. De manera que se puedan tomar decisiones adecuadas para el mejoramiento del aprendizaje. Utilizando la metodología utilizada fue Teoría Clásica del Test (TCT) y el modelo de Rasch (TRI).

Utilizando la TCT, se permitió reconocer que entre los ítems propuestos hay, un 44 % no discriminan entre los alumnos que saben de los que no saben. Los ítems que discriminan poco son un 32 %, con solo 24 % que discriminan adecuadamente.

La confiabilidad del cuestionario obtenida con el alfa de Cronbach, con los 34 ítems es de .667. Teniendo 34 % de los ítems entre muy fáciles y fáciles y hay 56 % de ítems se encuentran entre muy difíciles y difíciles y un 10 % que resultaron tener una dificultad media. Al eliminar ítems que no discriminan, se mantuvieron 22 ítems y la confiabilidad del cuestionario con el alfa de Cronbach, aumento a .774.

Como parte del análisis de las propiedades psicométricas del cuestionario, se utilizó el modelo de Rasch, que permite asegurar que los parámetros de las personas y de los ítems se expresen en las mismas unidades (medición conjunta), ajustar los datos al modelo demostrando qué personas son independientes de los ítems administrados (objetividad específica), y que la escala presenta propiedades de intervalo (propiedades de medida) como es el tipo logit, [30]. Para utilizar el modelo se realizó el análisis factorial para probar la unidimensionalidad y con ella se obtuvo 22.658 % de la varianza total es explicada por el primer componente que contenía 10 ítems. También se produce evidencia de que los ítems miden el constructo interés y que los alumnos poseen las habilidades que mide el instrumento.

De los 22 ítems preseleccionados, se desecharon 12 que no cumplían con las especificaciones del modelo y se aplicó el análisis Rasch a los 10 ítems restantes. Se obtuvieron niveles de ajuste adecuado para los 10 ítems, tanto para el índice Infit MNSQ como para el outfit MNSQ.

PTMEA se refiere a la correlación punto – media, mide el grado de asociación entre el puntaje particular observado para el reactivo (o examinado), es un indicativo que el reactivo trabaja en conjunto con la escala a la que pertenece. Los valores encontrados están entre 0.40 a 0.51.

En la Figura 3a, aparece una representación gráfica del escalamiento conjunto. Puede observarse una representación de los objetos (personas e Ítems) en un rango de valores entre -3 y 3 logit. se puede observar que los ítems P11 y P2 se consideran muy fáciles, en la media se encuentran los ítems P28, P26, P310 y que los ítems que se encuentran a una desviación estándar son P21, P10 y P13. Los ítems que resultaron ser difíciles P27 y P6.

En el mapa Figura 3b de las distribuciones conjuntas de los individuos y los ítems, se puede observar que el promedio del nivel de habilidad de los examinados está por debajo de la dificultad promedio de los ítems, la mayor parte de la población se ubicó por debajo de la dificultad promedio de los ítems, esto indica que la prueba resultó levemente difícil para los examinados.

El uso del modelo de Rasch, permite establecer el control de calidad del instrumento. Se obtuvieron el índice de separación de los ítems de 4.35 y la fiabilidad de los ítems .95, adecuado, lo que nos indica que la muestra utilizada es suficientemente grande como para confirmar la jerarquía de dificultad del ítem (validez de constructo) del instrumento, [13]. Por otro lado, el índice de separación de las personas (1.06) indica la aptitud del instrumento para discriminar a las personas en la variable medida y el índices de fiabilidad de las personas de 0.53 son considerados bajos, por lo que es necesario

aumentar el número de preguntas para cubrir otros niveles de habilidad, ya que este conjunto de ítems no es suficiente para distinguir entre sujetos de alto y de bajo rendimiento. El modelo confirma que los alumnos tienen una habilidad baja y que se debe reforzar el conocimiento mejorando estrategias de aprendizaje.

8.1. Discutir sobre la tabla de especificaciones y la unidimensionalidad

En [2, 11, 35] se comenta que existen diversas líneas metodológicas para evaluar la unidimensionalidad. Sin embargo, el uso de la herramienta del análisis factorial puede ser evaluada, debido a que es utilizada para estudiar la dimensionalidad de un conjunto de ítems, [22]. Como existen diversos criterios sobre la unidimensionalidad de la medición de un instrumento. El criterio que se toman en cuenta es la varianza explicada por el primer factor extraído. Así un conjunto de ítems será unidimensional si el primer factor explica por lo menos el 20 % de la varianza, [26]. Se inició con la verificación de la unidimensionalidad de los datos, la cual dice que un único constructo es suficiente para explicar los resultados de examinados y las relaciones entre ítems. Mediante el programa SPSS 22.0, se efectuó un análisis factorial exploratorio bajo el método de factorización componentes principales y se obtuvo que la varianza total explicada presentaba un gran factor que explicaba casi el 22.658 % de los datos de la matriz de correlación observada, de igual forma, en el gráfico de sedimentación se apreció la existencia de un factor predominante de acuerdo con el criterio de Reckase.

9. Bibliografía

[1] AERA (American Educational Research Association), American Psychological Association and National Council for Measurement in Education [AERA, APA and NCME] (2014). The Standards for Educational and Psychological Testing. Washington, D.C.: AERA.

[2] Anderson, J., Gerbin, D. and Hunter, J. (1987). On the assessment of unidimensional measurement: Internal and external consistency, and overall consistency criteria. *Journal of Marketing Research*, 24(4), 432-437.

[3] Bechger, T. M., Maris, G. Verstralen, H. H. F. M., and Béguin, A. A. (2003). Using Classical Test Theory in Combination With Item Response Theory. *Applied Psychological Measurement*, 27(5), 319–334. https://doi.org/10.1177/0146621603257518

[4] Bond, T. G. and Fox, C. M. (2003). Applying the Rasch model: Fundamental measurement in the human sciences. *Journal of Educational Measurement*, 40(2), 185-187. https://doi.org/10.1111/j.1745-3984.2003.tb01103.x

[5] Bond, T. and Fox, C. M. (2015). *Applying the Rasch Model: fundamental measurement in the human sciences* (*Third*). Routledge: New York

[6] Cea D'Ancona, M. (2002). *Análisis multivariable*. España: Editorial Síntesis, S.A.

[7] De Ayala, R. J. (2009) *The theory and practice of item response theory*. New York, New York: The Guilford Press. http://goo.gl/VLZzWJ

[8] Ghio, F.B., Morán, V.E., Garrido, S.J., Azpilicueta, A.E., Córtez, F. and Cupani, M. (2020) Calibración de un banco de ítems mediante el modelo de Rasch para medir razonamiento numérico, verbal y espacial. *Avances en Psicología Latinoamericana*, 38(1), 157-171. Doi: http://dx.doi.org/10.12804/revistas.urosario.edu.co/apl/a.7760

[9] Heffernan, E., Weinstein, B.E., and Ferguson, M.A.(2020). Application of Rasch Analysis to the Evaluation of the Measurement Properties of the Hearing Handicap Inventory for the Elderly. *Ear and hearing*, 41(5), 1125–1134. Doi: https://doi.org/10.1097/AUD.0000000000000832

[10] Kleinman, M. and Teresi, J.A. (2016). Differential item functioning magnitude and impact measures from item response theory models. *Psychological Test and Assessment Modeling*, 58(1), 79–98. https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5505278/

[11] Linacre, J. (1994). DIMTEST diminuendo. *Rasch Measurement Transactions*, 8(3), 384.

[12] Linacre, J. M. (2002). Optimizing rating scale category effectiveness. *Journal of Applied Measurement*, 3(1), 85-106. http://dx.doi.org/10.1.1.424.2811

[13] Linacre, J. M. (2016). *Winsteps®* (*Version 3.92.0*) [*Computer Software*]. Beaverton, Oregon. http://www.winsteps.com/

[14] Martín, N., Díaz, C., Córdoba, G. and Picquart, M. (2011). Calibración de una prueba de química por el modelo de Rasch. *Revista electrónica de investigación educativa*, 13(2), 132-148. Recuperado http://www.scielo.org.mx/scielo.php?script=sci_arttext&pid=S1607-40412011000200009&lng=es&tlng=es

[15] Martínez, M. R., Hernández, M.J. and Hernández, M. V. (2006). *Psicometria*, Madrid: Alianza Editorial.

[16] Masters, G.N. and Keeves, J.P., (1999). *Advances in measurement in educational research and assessment*. Amsterdam; New York: Pergamon.

[17] Messick, S. (1989a). Meaning and values in test validation: The science and ethics of assessment. *Educational Researcher*, 18(2), 5-11.

[18] Messick, S. (1989b). *Validity*. In R.L. Linn (Ed.), Educational measurement (3rd ed.). New York: Macmillan.

[19] Messick, S. (1994). Foundations of validity: Meaning and consequences in psychological assessment. *European Journal of Psychological Assessment*, 10, 1-9.

[20] Muñiz, J. (1997). *Introducción a la Teoría de Respuesta a los ítems*. Madrid: Ediciones Pirámide, S.A.

[21] Muñiz, J. (2003a). *Teoría Clásica de los Tests*. Madrid: Ediciones Pirámide, S.A.

[22] Muñiz, J. (2003b). La validación de los tests. *Metodología de las Ciencias del Comportamiento*, 5, 119-139.

[23] Nunnally, J. C. and Bernstein, I. J. (1995). *Teoría psicométrica, (3ra ed)*. México, D.F.: Editorial McGrawHill Latinoamericana

[24] Prieto, G. and Delgado, A. R. (1999). *Medición cognitiva de las aptitudes*. En J. Olea, V. Ponsoda y G. Prieto (Eds.) Tests informatizados: Fundamentos y aplicaciones. (207-226) Madrid: Pirámide.

[25] Prieto, G. and Delgado, A. R. (2003). Análisis de un test mediante el modelo de Rasch. *Psicothema*, 15(1), 94-100.

[26] Reckase, M. (1979). Unifactor latent trait models applied to multifactor tests: Results and implications. *Journal of Educational Statistics*, 4(3), 207-230.

[27] Rodríguez-Ayán Mazza, M. N. (2007). *Análisis multivariado del desempeño académico de estudiantes universitarios de química* (Tesis doctoral). Universidad Autónoma de Madrid, Madrid, España. https://repositorio.uam.es/bitstream/handle/10486/1800/5491_rodriguez_ayan.pdf

[28] Salas, J.S., and Rojas, E.M. (2011). Construcción y validación de una prueba de comprensión de lectura mediante el modelo de Rasch. *Revista Electrónica Actualidades Investigativas en Educación*. 11(2), 1-27.

[29] Scriven, M. (2013). *The future of evaluation in society: A tribute to Michael Scriven*. Stewart I. Donaldson.

[30] Schulz, W. and Fraillon, J. (2011). The analysis of measurement equivalence in international studies using the Rasch model. *Educational Research and Evaluation*, 17(6), 447-464. http://dx.doi.org/10.1080/13803611.2011.630559

[31] Tristán, A. (2002). *Análisis de Rasch para todos*, Ed. Ceneval, México.

[32] Wilson, M. (2005). *Constructing measures: An item response modeling approach*, Mahwah, New Jersey: Lawrence Erlabaum Associates

[33] Wright, B. D. (1996). Comparing Rasch measurement and factor analysis. *Structural Equation Modeling: A Multidisciplinary Journal*, 3(1), 3-24. http://dx.doi.org/10.1080/10705519609540026

[34] Wright, B. D. and Linacre, J. M. (1994). Reasonable mean-square fit values. *Rasch Measurement Transactions*, 8, 370-371

[35] Wright, B. & Linacre, J. (1995). MESA Research memoramdum 44. *Archives of physical, medicine and rehabilitation*, 70 (12) 857-860.