UNIVERSIDAD DEL VALLE DE GUATEMALA

Facultad de Educación

Análisis psicométrico de la prueba de diagnóstico de matemática de la Universidad del Valle de Guatemala

Trabajo de graduación en modalidad de modelo de trabajo profesional

Ester Noemy Rodas Ochoa

para optar al grado académico de

Maestría en Medición Evaluación e Investigación Educativa

Guatemala

2019

Análisis psicométrico de la prueba de diagnóstico de matemática de la Universidad del Valle de Guatemala

UNIVERSIDAD DEL VALLE DE GUATEMALA

Facultad de Educación

Análisis psicométrico de la prueba de diagnóstico de matemática de la Universidad del Valle de Guatemala

Trabajo de graduación en modalidad de modelo de trabajo profesional

Ester Noemy Rodas Ochoa

para optar al grado académico de

Maestría en Medición Evaluación e Investigación Educativa

Guatemala
2019

Vo. Bo.

(f)_____

MA. Amalia Ruballos

Tribunal Examinador:

(f)_____

MA. Amalia Ruballos

(f)_____

MA. Justo Magzul

(f)_____

MA. Herbert Smith

Fecha de aprobación: Guatemala, 29 de noviembre del 2019

Contenido

ÍNDICE DE TABLAS

ÍNDICE DE DIAGRAMAS

ÍNDICE DE FÓRMULAS

ÍNDICE DE ILUSTRACIONES

ÍNDICE DE GRÁFICAS

ÍNDICE DE CUADROS

RESUMEN

En este estudio se realizó el análisis psicométrico de la prueba de diagnóstico de matemática que utilizó el Departamento de Matemática de la Universidad del Valle campus central, en los años 2016, 2017 y 2019. La prueba fue aplicada a los estudiantes de primer ingreso inscritos en el de curso Pensamiento Cuantitativo del primer ciclo en las Facultades de Ingeniería, Ciencias Sociales y Ciencias y Humanidades. El análisis psicométrico se realizó utilizando los modelos de Teoría Clásica y Teoría de Respuesta al Ítem. Además, se desarrolló un juicio de expertos utilizando la metodología de Bookmark que sugiere los puntos de corte que delimitan los niveles de desempeño y los descriptores de estos, que permiten ubicar el porcentaje de alumnos que pertenecen a cada nivel. Se encontró de manera general que la prueba tiene un Alpha de Cronbach de 0.70, hay ítems entre todas las categorías de dificultad, menos en la "muy fáciles". Por medio del modelo de Rasch se establece que $Theta = 0.15$ es el punto de corte; los valores mayores a este demuestran que los estudiantes tienen la habilidad de contestar más del 50% de los ítems de manera correcta. Mientras que los estudiantes valores menores a este demuestran la habilidad de contestar hasta 50% de los ítems de manera correcta.

I. INTRODUCCIÓN

El contexto actual, documentado en la Dirección General de Evaluación e Investigación Educativa -Digeduca- demuestra en los resultados del año 2018 que en el área de matemática únicamente el 11.44% de los estudiantes alcanza los niveles satisfactorio y excelente en la prueba estandarizada aplicada a nivel nacional. Estos resultados documentan la necesidad que presentan las instituciones de educación superior como Universidad del Valle de Guatemala -UVG- de diagnosticar el dominio de conocimientos y habilidades en el área de matemática. Para establecer las mayores dificultades manifestadas por los alumnos y crear planes estratégicos que permitan a los estudiantes mejorar su desempeño y desarrollar el aprendizaje en el área.

Una prueba de diagnóstico es un instrumento que debe cumplir con indicadores psicométricos que garanticen su calidad, para generar medidas lo más precisas posibles que permitan conocer el nivel de desempeño que poseen los alumnos al iniciar la universidad. Por esta razón, este estudio busca determinar en qué medida los resultados de la prueba de diagnóstico del Departamento de Matemática de UVG campus central, cumple con los indicadores psicométricos que demuestran la calidad de esta.

Se trabajó un estudio con enfoque cuantitativo, de tipo no experimental con diseño longitudinal de tendencia y alcance descriptivo. La muestra utilizada es de tipo no probabilístico que consideró a 1,172 estudiantes de primer ingreso que llevaron el curso de Pensamiento Cuantitativo en el primer semestre en los años 2016,2017 y 2019. Las limitaciones documentan que no se trabajó con los datos del año 2018 ya que por razones administrativas no se pudo aplicar la prueba al inicio del ciclo y otras relacionadas al juicio de expertos.

Los aportes indican que el análisis de la confiabilidad según Teoría Clásica de los Test-TCT- y según Teoría de Respuesta al Ítem -TRI- la prueba es clasificada con una confiabilidad "aceptable". El análisis del índice de dificultad y del parámetro de dificultad indica que hay únicamente 2 ítems muy difíciles, ninguno muy fácil y la mayoría distribuida de manera equitativa entre los niveles "fácil", "moderado" y "difícil". El análisis de discriminación indica que el 50% de los ítems discrimina de manera óptima. Se documentó en el análisis de distractores los ítems que tienen por lo menos un distractor poco atractivo y los que todos los distractores son más selectos que la respuesta correcta.

Según el análisis de TRI se dedujo que el ítem No. 9 se demuestra ajuste de WMS (Infit) ni de UMS (Oufit). El punto de corte está ubicado en $Theta = 0.15$ y que el 70% de los resultados de los estudiantes demuestran una habilidad para contestar hasta el 50% de los ítems de manera correcta. Los resultados también apuntan a clasificar a los estudiantes únicamente en dos estratos. Y de manera general se recomienda agregar ítems en los niveles "muy fáciles", "moderados" y "muy difíciles" para medir la habilidad de los estudiantes con habilidades extremas y diferenciar de manera más precisa los que demuestran habilidades cerca del punto de corte.

Este estudio, además de proporcionar de manera específica el comportamiento de los ítems y una interpretación de los resultados obtenidos por los estudiantes, da un indicio a la transición de realizar los análisis psicométricos con pruebas del Departamento de Matemática de UVG ya no únicamente con la TCT que es lo usual, sino con TRI, que proporciona aportes significativos en los análisis al relacionar sobre la misma escala a los ítems y las puntuaciones de las personas; a su vez permite trabajar con pruebas con menor cantidad de ítems, pruebas con ítems de otro tipo y no únicamente opción múltiple. Lo cual es desventaja de la TCT.

II. OBJETIVOS

A. Objetivo general:

Determinar en qué medida los resultados de la prueba de diagnóstico del Departamento de Matemática de la Universidad del Valle de Guatemala (UVG) Campus Central, cumplen con los indicadores psicométricos que demuestran la calidad de la misma.

B. Objetivos específicos:

1. Realizar un análisis psicométrico de la prueba de diagnóstico de matemática del Departamento de Matemática de la UVG, con los resultados de los años 2016, 2017 y 2019.

2. Diseñar criterios y descriptores de desempeño de la prueba de diagnóstico que utiliza el Departamento de Matemática de UVG por medio de un juicio de expertos con la metodología Bookmark.

3. Categorizar a los resultados de los estudiantes de los años 2016,2017 y 2019 según los niveles de desempeño establecidos por el juicio de expertos con la metodología Bookmark.

4. Proporcionar hallazgos y recomendaciones fundamentados con los análisis realizados al Departamento de Matemática respecto a la prueba de diagnóstico.

III. JUSTIFICACIÓN

De acuerdo con Martínez Arias, Hernández Lloreda, & Hernández Lloreda (2006) al momento de desarrollarse un proceso de medición, se comenten diversos errores. Por consiguiente, la confiabilidad busca identificar hasta qué punto las medidas proporcionadas por una prueba, reflejan con precisión las medidas verdaderas de los resultados. De tal manera que una prueba establece cierta consistencia en la manera en la que emite las puntuaciones, ya sea entre las puntuaciones de los mismos sujetos obtenidas en diferentes momentos o con ítems equivalentes Argibay (2016).

Para Martínez Arias, Hernández Lloreda, & Hernández Lloreda (2006) la validez indica si el nivel de uso que se pretende hacer con las puntuaciones de las pruebas está justificado, con el fin de probar la fuerza y credibilidad de los datos usando diversas fuentes de evidencia. Para Argibay (2016) la validez determina si el instrumento está midiendo realmente lo que dice medir. A diferencia de esto Aiken (2003) plantea que esta definición tradicional implica que una prueba solo tiene un tipo de validez. Sin embargo, la validez depende del propósito para el cual la prueba fue diseñada. El proceso para la validación de una prueba implica una acumulación de evidencia sólida y científica que explique el modo en el que se interpretaran los resultados, ya que no se validan los instrumentos, sino el uso que se hace de ellos Eyzaguirre (2003).

Estos conceptos y definiciones fundamentan inicialmente la importancia de que en todas las pruebas que sean creados, utilizadas o aplicadas es indispensable establecer la confiabilidad y validez. De tal manera que en los resultados se puedan proporcionar puntuaciones precisas y lo más acercadas a la realidad. Aplicar una prueba sin estudios de las propiedades psicométricas es cuestionable, y principalmente si el test tiene altas consecuencias. No es recomendable medir y, por consiguiente, tomar decisiones según la información que desconoce la precisión en sus resultados obtenidos.

Digeduca expone la situación actual de los resultados de Lectura y Matemática de los estudiantes que egresan del sistema educativo nacional en el año 2018. Dichas evaluaciones aplicadas a 158,161 estudiantes graduandos de los 4,083 establecimientos educativos de todo el país. Estos resultados se resumen en cuatro niveles de desempeño: Insatisfactorio, Debe Mejorar, Satisfactorio y Excelente establecidos para describir e interpretar el logro representado. Estos resultados demuestran en resumen que únicamente el 11.44% de los estudiantes evaluados consiguen posicionarse entre las categorías de "satisfactorio" y "excelente" en el área de matemática Digeduca (2018).

A pesar de que estos resultados son bajos, cabe mencionar que es el porcentaje más alto alcanzado en estos últimos 8 años. Siendo el intervalo de resultados de un 7.30% a un 11.44% actual la visión del alcance de los estudiantes graduandos en el área de matemática Digeduca (2018). A continuación, se presenta una tabla que describe el porcentaje de alumnos en cada nivel de desempeño y los descriptores del mismo según los resultados de las evaluaciones estandarizadas del año 2018:

Tabla 1: Resultados del área de matemática prueba Digeduca 2018

Nivel de desempeño	**Porcentaje de estudiantes**	**Descriptor del desempeño**
Excelente	7.38%	*"Los estudiantes, hacen lo del nivel satisfactorio y también resuelven operaciones combinadas, calculan áreas y perímetros de figuras combinadas, determinan probabilidades, traducen el enunciado verbal de un problema a lenguaje algebraico y pueden resolver problemas con información implícita".*
Satisfactorio	4.06%	*"Los estudiantes, además de hacer lo de los niveles inferiores, también realizan conversiones de medidas de tiempo, capacidad, longitud y peso; utilizan los números reales para resolver problemas y simplifican expresiones numéricas y algebraicas".*

Nivel de desempeño	**Porcentaje de estudiantes**	**Descriptor del desempeño**
Debe Mejorar	32.62%	*"Los estudiantes pueden ejecutar operaciones matemáticas considerando la jerarquía de las mismas, deducir secuencias numéricas, calcular perímetros y áreas, así como identificar expresiones algebraicas a partir de un enunciado".*
Insatisfactorio	55.93%	*"Los estudiantes tienen una debilidad significativa en la comprensión y aplicación de conceptos matemáticos en aritmética, geometría, álgebra y estadística. Poseen un vocabulario matemático limitado".*

Fuente: Elaboración propia según Digeduca (2018) e informes de resultados por institución.

Los resultados apuntan a que solo 1 de cada 10 de los alumnos que egresan de diversificado demuestran dominio en los conocimientos y habilidades que mide la prueba respecto al área de matemática. Siendo estos resultados un parámetro alarmante es evidente mencionar que si los estudiantes continúan sus estudios universitarios es muy probable que demuestren deficiencias en esta área. Siendo según la orientación especifica la base para muchos cursos a desarrollarse. En donde es indispensable un dominio óptimo de las competencias matemáticas previas a la etapa universitaria.

En relación con el contexto educativo actual y para contar con un panorama más específico de la población estudiantil con la que se va a trabajar, algunas instituciones educativas a nivel superior aplican pruebas de diagnóstico que los alumnos cuando estos están en su primer semestre. No está de más resaltar que dichas pruebas deben contener las propiedades psicométricas que permitan garantizar una evaluación que proporciona puntuaciones verdaderas, para identificar específicamente los contenidos y habilidades en los que se presentan mayores deficiencias. Y de esta manera crear estrategias y líneas de trabajo para que los alumnos puedan desarrollar las habilidades con dificultad y alcanzar un dominio aceptable de esta área a nivel superior.

La UVG es una institución educativa privada de nivel superior. Uno de los valores fomentados es el desarrollo del pensamiento crítico Universidad del Valle de Guatemala (2017), que permite poner a disposición los conocimientos y habilidades para plantear soluciones. En los compromisos se menciona la veracidad de la información con documentos y registro exactos que respaldan los procesos. Por lo tanto, para el Departamento de Matemática del Campus Central, el análisis pertinente acerca de los resultados de la prueba de diagnóstico que realizan los alumnos de primer año contribuye a garantizar veracidad en las mediciones realizadas. Y promueve a la creación de estrategias docentes que buscan mejorar el desempeño y aprendizaje de los alumnos de primer ingreso que presentan dificultades.

Considerando que el proceso de admisión de UVG para la Facultades Ingeniería, Ciencias Sociales, Ciencias y Humanidades, Desing Innovation & Arts, Global Management and Business Intelligence y Colegio Universitario implica únicamente que los alumnos realicen la Prueba de Aptitudes Académicas -PAA-. Si esta prueba no es aprobada se puede optar por tomarla 4 meses después. Respecto a los criterios para ser admitido en UVG existe un comité especializado en admisiones evalúa cada caso y combina los criterios de puntuación del examen y el rendimiento académico a nivel medio. Universidad del Valle de Guatemala (2019)

En relación con la PAA:

> «*La PAA es una prueba que evalúa las habilidades y los conocimientos necesarios para hacer trabajo académico de nivel universitario. Desde sus inicios, este instrumento se ha desarrollado para predecir, junto con otros criterios, el éxito en el primer año de estudios superiores*».
>
> College Board (2019*).*

Organizada en tres componentes: Lectura y Redacción, Matemática e inglés. El componente de Matemática se busca medir de manera general razonamiento matemático y aprovechamiento en las áreas de aritmética, álgebra, geometría y análisis de datos y probabilidad. Sin embargo, a pesar de utilizar esta prueba como filtro en la admisión de alumnos, muchos de los alumnos que si son admitidos demuestran dificultades en el primer curso del área de matemática que imparte UVG.

A razón de esto, el Departamento de Matemática aplica una prueba de diagnóstico a los alumnos de primer ingreso con el fin de visualizar la tendencia de los datos y las áreas con mayor dificultad. Se recopilan los resultados para establecer una puntuación individual respecto al número de preguntas contestadas de manera correcta en relación con el total de las preguntas del test. Sin embargo, no se ha documentado de manera específica los indicadores psicométricos de esta prueba de diagnóstico y tampoco se han establecido los criterios para interpretar estos resultados, más que una puntuación individual. Dada esta razón, se presenta la necesidad de analizar los resultados de la prueba de diagnóstico de matemática del Departamento de Matemática de UVG Campus Central en los años 2016, 2017 y 2019; para establecer los indicadores psicométricos que determinan la calidad de esta y aplicar un mecanismo que permita realizar la interpretación pertinente de los resultados.

De igual manera y de acuerdo con los hallazgos y conclusiones de Ramírez & Barquero (2011) pretende que la información de la prueba sea utilizada proponer recomendaciones a favor de los alumnos como programas y planes de apoyo en el área psicoeducativa, cursos de nivelación, métodos de estudios, tutorías y otros.

IV. MARCO TEÓRICO

El marco teórico tiene como fin fundamentar o sustentar la investigación por medio del análisis, descripción y comparación de las teorías desarrolladas por expertos del tema. Hernández Sampieri (2010) nombra a este proceso «El desarrollo de la perspectiva teórica». el cual se define como:

> «*El desarrollo de la perspectiva teórica es un proceso y un producto. Un proceso de inmersión en el conocimiento existente y disponible que puede estar vinculado con nuestro planteamiento del problema, y un producto (marco teórico) que a su vez es parte de un producto mayor: el reporte de investigación*».
>
> *Yedigis y Weinbach (2005) en* Hernández Sampieri (2010)

A. Medición

La Real Academia Española (2019) define la palabra medición como la "*acción y efecto de medir".* Según Aliaga Tovar (2011):

> «*En la psicología, la educación y las ciencias sociales se trata de medir aspectos que no son físicos ni directamente observables*».

La medición según Nunnally (1987):

> «*Consiste en reglas para la asignación de números a objetos en tal forma que representen cantidades de atributos*».

Es relevante identificar que medición se visualiza como un proceso, estructurado y organizado que busca asignar una ponderación al nivel de dominio de un atributo evaluado.

La psicometría se define como:

> «*Una disciplina de la psicología cuya finalidad intrínseca es la de aportar soluciones al problema de la medida en cualquier proceso de investigación psicológica*».
>
> Aliaga Tovar (2011)

Existe toda una teoría enfocada a la medición, que organiza conceptualmente la manera en la que se realizan las inferencias a partir de las puntuaciones obtenidas de los test. Fundamentada en modelos estadísticos y teorías de la probabilidad que garantizan un análisis de calidad. De manera histórica las teorías de la medición marcan una trayectoria que ha ido profundizando los análisis y mejorando las interpretaciones, para promover tomas de decisiones fundamentadas y argumentos certeros.

La creación, uso o adaptación de test resulta un proceso sumamente complejo, ya que existen muchas variables y factores que se asocian al momento de realizar una medición, y que pueden intervenir en los resultados de la misma. La deducción incorrecta o uso inadecuado de los resultados de un test puede perjudicar en gran medida a los sujetos involucrados, principalmente por la toma de decisiones que se realizan a partir de estos. Debido al impacto que puede generar y las consecuencias que esto puede proceder, Martínez Arias, Hernández Lloreda, & Hernández Lloreda (2006) describen los problemas comunes de la medición:

- Un constructo es el "atributo o rasgo latente" que trata de medir la prueba, comúnmente en educación y psicología los constructos son rasgos no observables, como el razonamiento abstracto, inteligencia, etc. Los constructos tienen diversas maneras de ser medidos, por lo tanto, no hay una única manera de realizar la medición, y esto genera un abanico de métodos, procedimientos y conclusiones.
- Una de las principales problemáticas es determinar el número de elementos y los diversos niveles que debe contener un test para medir un dominio.
- Es común que las medidas obtenidas de un proceso de medición contengan errores. Estos errores pueden ser aleatorios que es lo normal o ya sistemáticos que conllevan a buscar que elemento está incidiendo las mediciones y controlar en medida de lo posible el error.
- No existen escalas de medidas ya establecidas que permitan comparar la manera en la que se mide un dominio, lo que conlleva a considerar la "indeterminación de las medidas" ya que no existe un referente general de comparación.
- Los constructos no son elementos aislados, por lo tanto, es indispensable establecer

las relaciones con otros constructos y con conductas que permitan ser observables.

B. Test

La palabra test se traduce del inglés como el termino: prueba, según la Real Academia Española (2019) test se define como:

> «*Prueba destinada a evaluar conocimientos o aptitudes, en la cual hay que elegir la respuesta correcta entre varias opciones previamente fijadas*».

Concepto también asociado a términos como examen y evaluación. Según SEPT (1999) en Martínez Arias, Hernández Lloreda, & Hernández Lloreda (2006) define que:

> «*Un test es un instrumento evaluativo o procedimiento en el que se obtiene una muestra de la conducta de los examinados en un dominio especificado y posteriormente es evaluada y puntuada usando un procedimiento estandarizado*».

De manera más específica Aliaga Tovar (2011) propone que:

> «*El test psicométrico es un procedimiento estandarizado compuesto por ítems seleccionados y organizados, concebidos para provocar en el individuo ciertas reacciones registrables; reacciones de toda naturaleza en cuanto a su complejidad, duración, forma, expresión y significado*»
>
> Rey (1973) en Aliaga Tovar (2011).

1. Clasificación de los test

Dependiendo los fines para los cuales han sido diseñado, los momentos de aplicación o los sujetos involucrados los test tienen diversas maneras de clasificarse. Según Martínez Arias, Hernández Lloreda, & Hernández Lloreda (2006) se pueden clasificar en función a:

a. Las consecuencias para los sujetos:

Sirven para realizar distinciones entre sujetos. Son muy utilizados para procesos de selección para algún proceso. O para recolección de datos para investigación.

b. Planteamiento del problema y tipo de respuesta:

Estos test se utilizan cuando existe una determinada manera de plantear el problema y presentarlo, así como las diferentes maneras de presentar las respuestas: elección múltiple, respuesta construida, verdadero o falso.

c. Área de comportamiento acotada:

Esto depende lo que el test busque evaluar comúnmente se clasifican como:

1) Test cognitivos:

Estos están enfocados a evaluar aspectos como aptitudes, inteligencia, rendimiento académico.

2) Test no cognitivos:

Evalúan aspectos como personalidad, motivación, actitudes.

d. Modalidad de aplicación:

Esto hace en referencia a si el test se resuelve de manera individual o grupal, si es a lápiz y papel, oral, virtual, etc.

e. Demandas temporales:

Clasifica los test en un continuo que van de rapidez a potencia.

f. Grado de aculturación o demandas específicas:

Desarrollado específicamente para un grupo o cultura requerido, para visualizar la manera en la que se resuelve el test y lograr identificar elementos investigados.

g. Modelo estadístico:

Desde la construcción se establece un modelo estadístico que permite analizar e interpretar las puntuaciones para hacer inferencias de los resultados obtenidos. La clasificación tradicional de los modelos se enfoca en tres:

- Teoría clásica de los test
- Teoría de la generalizabilidad
- Teoría de respuesta al ítem.

h. Tipo de interpretación de las puntuaciones:

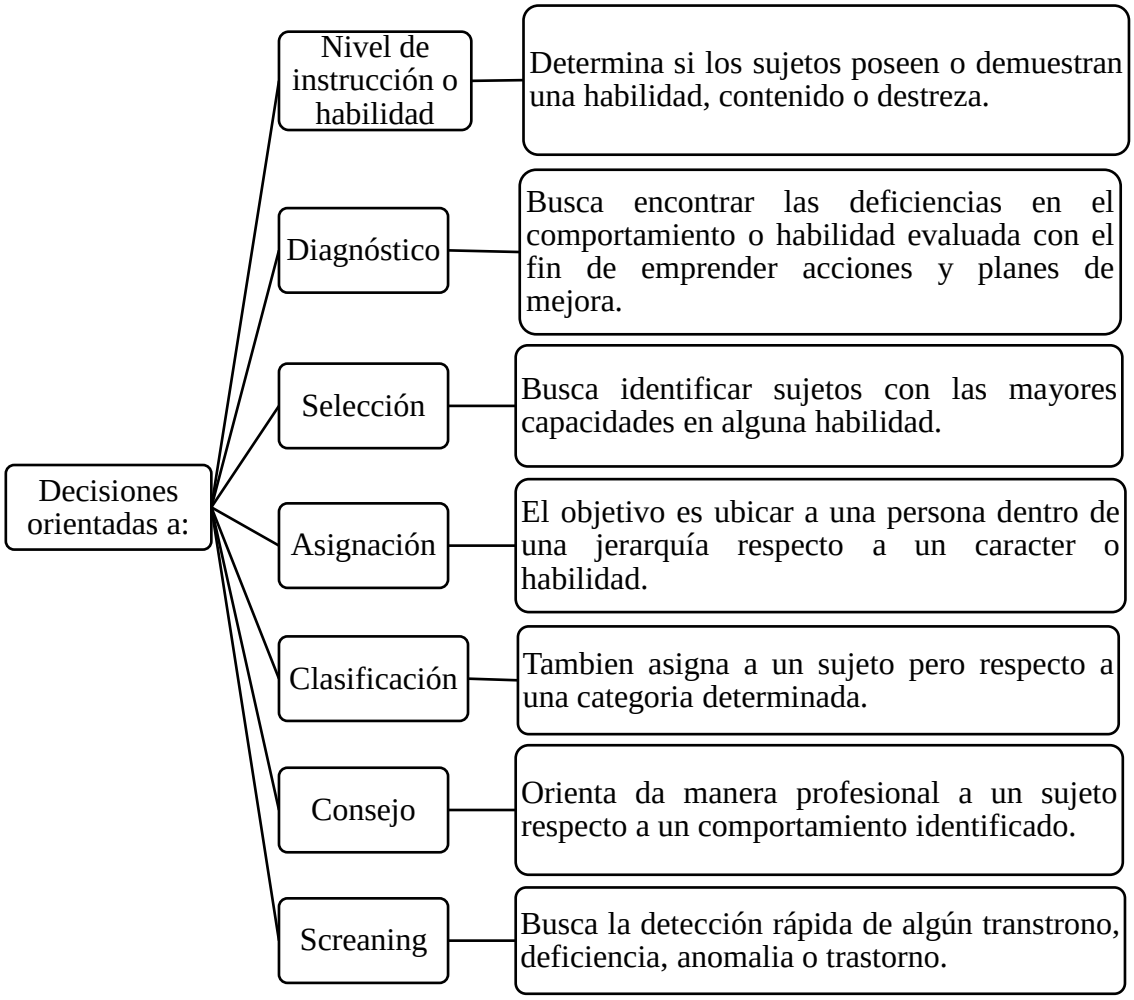
Se clasifican comúnmente en dos grupos:

- Los test que se centran en el dominio o grado de ejecución de algún criterio en específico son denominados test referidos a criterio.
- Los test que se basan en normas y sitúan a los sujetos en función a un estadístico calculado.

2. Finalidad de los test

La construcción de un test inicia con la determinación del objetivo y el constructo que se pretenda medir. Haciendo la respectiva selección de la población, la modalidad de aplicación y la visión de la toma de decisiones que se proporcionará a partir de los resultados del mismo. Esta toma de decisiones permite establecer diversos fines para los cuales está diseñado un test:

Diagrama 1: Finalidad de los test



Fuente: Elaboración propia según Martínez Arias, Hernández Lloreda, & Hernández Lloreda (2006)

3. Test de diagnóstico

Dando énfasis en los test de diagnóstico para dar el fundamento teórico de este estudio, es importante resaltar que este tipo de test no proporciona únicamente la medición de las habilidades o conceptos que domina el alumno. Sino que su fin es más amplio, enfocado a crear planes de acción, estrategias, líneas base, etc. Que permitan desarrollar y mejorar las

habilidades de los alumnos y hacer el aprendizaje más centrado y específico para el docente y las áreas que tenga que reforzar.

Un test de diagnóstico se realiza al iniciar un periodo académico. Considerando que hay un conjunto de habilidades, destrezas, conocimientos y experiencias de aprendizaje previas que el alumno ha desarrollado en su vida académica. Y busca determinar en qué medida el alumno puede demostrar dominio de estas habilidades o conocimientos. Así como también la identificación de las deficiencias y áreas para desarrollar.

> *«Este tipo de evaluación tiene una función diagnóstica o exploratoria y sirve justamente para evaluar las características que los estudiantes traen al proceso de enseñanza, es decir, sus conocimientos previos, los cuales se relacionan directamente con el aprendizaje Dochy y Alexander (1995), habilidades y competencias, intereses, motivaciones y disposición para el estudio de los contenidos en cuestión*».
>
> Bombelli (2011)

Con el fin utilizar los resultados y dar una interpretación pertinente, se busca tomar decisiones que permitan la mejora del proceso educativo tanto para el alumno como para el docente.

> *«El diagnóstico educativo, orienta la intervención del docente en distintos aspectos; por ejemplo, en cuanto al tiempo que dedicará a los temas; en una palabra, a la práctica docente*».
>
> Bombelli (2011)

C. El ítem

Bajo el concepto psicológico y educativo, el ítem es una unidad que forma parte de un conjunto de elementos que integran un test o cuestionario. Busca obtener respuestas de los sujetos que desea evaluar o investigar. La construcción de los mismos no puede ser de

manera intuitiva, debe ser de manera sistemática y enfocada a que los mismos cumplan con ciertas propiedades que garanticen la calidad de la medición.

> *«Es la unidad básica de observación de una prueba objetiva. Se utiliza para medir conocimientos formales, habilidades cognitivas adquiridas a través de la experiencia y aprendizajes complejos producto de las dos primeras. No requiere de juicios personales del evaluador o de interpretaciones para calificar las respuestas correctas. Posee una respuesta única previamente establecida y acordad de manera colegiada».*
>
> Sánchez Restrepo & Espinosa Rodríguez (2012)

Los objetivos de los ítems se dividen en dos grandes grupos:

> «*Uno es estudiar el nivel óptimo o máximo del sujeto en determinadas competencias o rendimientos, ya sea de tipo memorístico o de razonamiento, y otro de las actitudes, personalidad preferencias o emociones típicas de cada persona*».
>
> Crocker y Algina (1986); Cronbach (1985); Nunnally y Berstein (1995) en Muñiz, Fidalgo, García-Cueto, Martínez, & Moreno (2005)

Muñiz, Fidalgo, García-Cueto, Martínez, & Moreno (2005) propone componentes de ítems y características:

1. Componentes

Los componentes son los elementos que integran de manera básica los ítems, organizados en tres principales

a) La base o cuerpo:

Es la proposición que representa una situación hipotética o problema.

b) Opciones de respuesta:

Son las alternativas y distractores que presenta el ítem, en función de lo que se está evaluando.

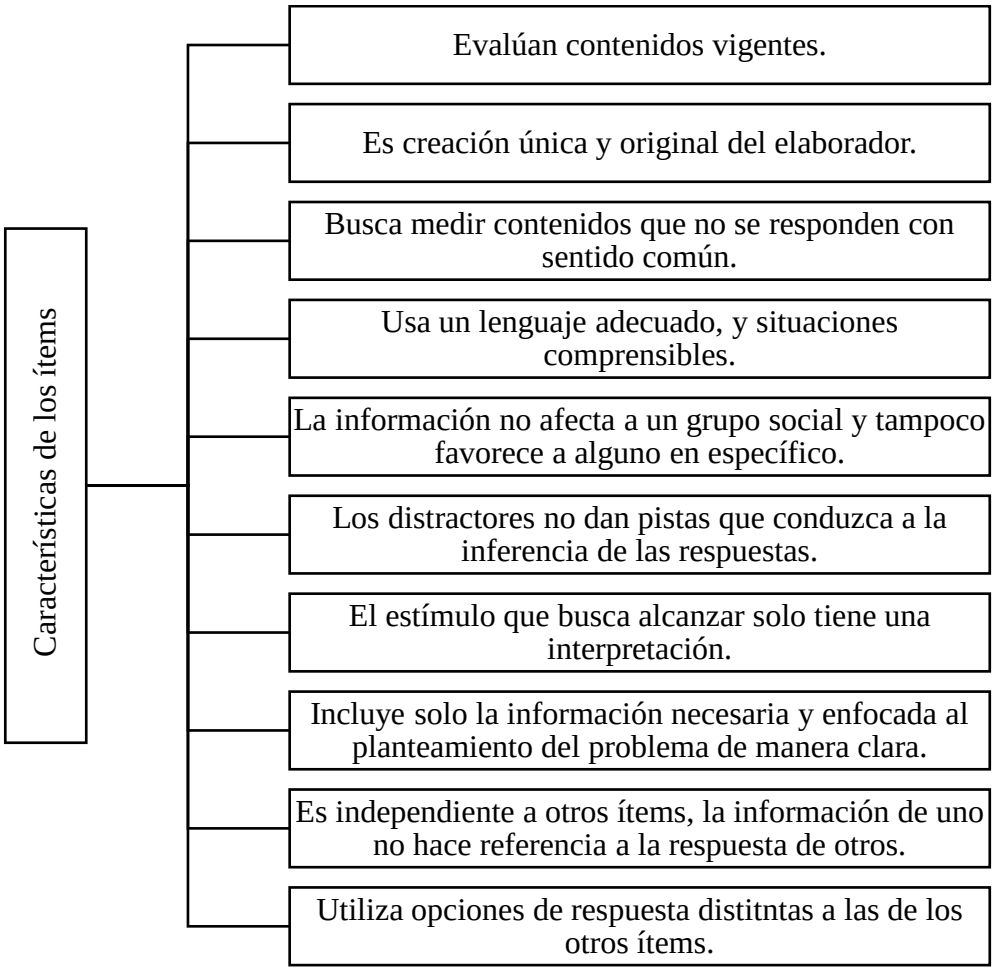
c) Las argumentaciones:

Son las explicaciones que fundamentan cada una de las opciones de respuesta propuestas.

2. Características de los ítems:

Estas describen la diversidad de los aspectos con los que deben cumplir los ítems a pesar de las diferencias de los mismos y los diferentes formatos en los que se pueden presentar:

Diagrama 2: Características de los ítems



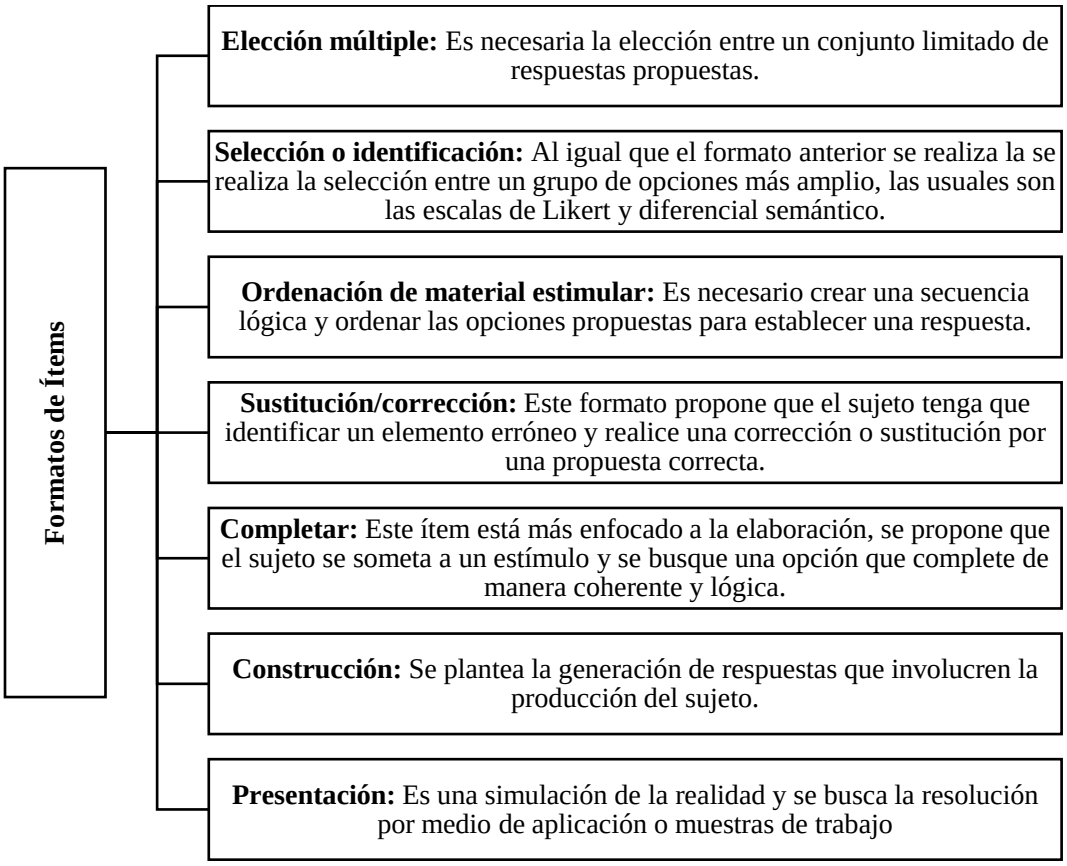
Fuente: Elaboración propia según Muñiz, Fidalgo, García-Cueto, Martínez, & Moreno (2005)

3. Tipos de ítems

Muñiz, Fidalgo, García-Cueto, Martínez, & Moreno (2005) exponen con respecto a la presentación de los ítems existen dos grandes grupos para clasificarlos. El primer grupo propone un formato que involucra la "elaboración de un producto" son los ítems en donde los sujetes deben completar, describir, exponer y expresar ideas, temas o situaciones problemáticas. El segundo grupo propone un formato en donde el sujeto hace una "*selección entre las opciones propuestas*", se puede presentar en diferentes formatos como selección de verdadero-falso, emparejamiento, elección múltiple, etc.

Bennett (1993) en Martínez Arias, Hernández Lloreda, & Hernández Lloreda (2006) hace una propuesta de los posibles formatos que pueden presentar los ítems:

Diagrama 3: Formatos de los ítems



Fuente: Elaboración propia según Bennett (1993) en Martínez Arias, Hernández Lloreda, & Hernández Lloreda (2006)

Es importante resaltar que al establecer una escala se recomienda colocar un número par de opciones para motivar una mayor discriminación. Cuando se utilicen etiquetas verbales en las escalas estas deben representar diferencias especificas entre sí.

D. Propiedades psicométricas

Martínez Arias, Hernández Lloreda, & Hernández Lloreda (2006) establece que las propiedades psicométricas son los *«principios que garantizan la calidad de las medidas»* proporcionadas por los ítems. Estas que, a su vez, también son conocidas como principios psicométricos, que marcan los criterios básicos que se deben identificar para considerar que un test cumple con las condiciones necesarias para ser utilizado.

Existen diversas listas que proponen las propiedades psicométricas, que dependiendo del autor consultado se denominan como "indispensables". Por ejemplo, Mislevy (2003) en Martínez Arias, Hernández Lloreda, & Hernández Lloreda (2006) propone que son: validez, fiabilidad, comparabilidad, equidad. Sin embargo, Muñiz, Fidalgo, García-Cueto, Martínez, & Moreno (2005) establecen que las "propiedades globales" son la fiabilidad y la validez, porque:

> «*Pues mal se pueden analizar y evaluar los ítems de un test si no se tiene claro el fin que se persigue, que no es otro, como acabamos de señalar, que obtener test fiables y validos*».

1. Fiabilidad

Fiabilidad también conocida como confiabilidad se define como:

> "*La confiabilidad (o consistencia) de un test es la precisión con que el test mide lo que mide, en una población determinada y en las condiciones normales de aplicación*".
>
> Anastasi (1982); Aiken (1995) en Aliaga Tovar (2011).

Por otra parte Martínez Arias, Hernández Lloreda, & Hernández Lloreda (2006) la definen como:

> «*Este principio tiene que ver con los errores cometidos en el proceso de medición, por lo que responde al problema de hasta qué punto las cantidades observadas reflejan con precisión la puntuación verdadera (puntuación del universo o aptitud) de la persona*».

Respecto a los errores cometidos en el proceso de medición, la falta de confiabilidad de un test se relaciona con el error.

> «*Se considera que el error es cualquier efecto irrelevante para los fines o resultados de la medición que influye sobre la falta de confiabilidad de tal medición*».
>
> Aliaga Tovar (2011)

El error puede ser de dos tipos:

a) Error sistemático:

Hace referencia a un error constante, que proporciona mediciones mayores o menores a lo que realmente deben ser.

b) Error causal:

Este se produce cuando algunas mediciones son mayores o menores a las que realmente deben ser. Este error puede verse afectado por factores que inciden en el sujeto que realiza el test como fatiga, tensión emocional, condiciones externas, situaciones o condiciones personales.

Muñiz, Fidalgo, García-Cueto, Martínez, & Moreno (2005) exponen diversas maneras de analizar y calcular la confiabilidad, esto depende del modelo psicométrico se esté utilizando como fundamento. La Teoría clásica analiza la confiabilidad por medio de un "coeficiente de confiabilidad", el cual proporciona una estimación global de la confiabilidad del test. Mientras que la Teoría de respuesta al ítem deja a un lado la estimación global y utiliza una "función de información para cada uno de los ítems". "*Esta función indica la precisión con la que el ítem está midiendo a cada nivel de la variable medida*". Muñiz, Fidalgo, García-Cueto, Martínez, & Moreno (2005)

a. Coeficiente de confiabilidad

El coeficiente de confiabilidad es una medida psicométrica que proporciona un parámetro global que describa la precisión con la cual el test mide lo que debe medir. Comúnmente el complemento de esta medida a uno es lo que se considera como error.

> *«Es un coeficiente de correlación entre dos grupos de puntajes e indica el grado en que los individuos mantienen sus posiciones dentro de un grupo. Abarca valores desde 0 a 1. Cuanto más se acerque el coeficiente a 1, más confiable será la prueba. El coeficiente de confiabilidad señala la cuantía en que las medidas del test están libres de errores casuales o no sistemáticos».*
>
> Aliaga Tovar (2011)

Según Muñiz, Fidalgo, García-Cueto, Martínez, & Moreno (2005) el coeficiente de confiabilidad se puede ver afectado por tres factores elementales:

- La longitud del test: Se comprende como el número de ítems que contiene el test. La teoría establece que, a un mayor número de ítems, mayor es la confiabilidad del
- test. Este crecimiento no se realiza de manera lineal, sino siguiendo la relación establecida por la fórmula de Spearman-Brown.
- La variabilidad de la muestra: Al aumentar la variabilidad de la muestra la confiablidad también tiende a aumentar.
- El nivel del sujeto en la variable medida: Hace referencia a los diversos grupos de puntuaciones en los que puede ubicarse un sujeto.

b. Métodos para calcular el coeficiente de confiabilidad:

Muñiz, Fidalgo, García-Cueto, Martínez, & Moreno (2005) establece que hay tres métodos para hacer la estimación empírica del coeficiente de confiablidad.

- Test paralelos: Se aplican dos formas equivalentes del test, o test paralelos (que miden los mismos constructos) y se correlacionan los resultados entre sí, con el coeficiente de correlación de Pearson.
- Test-retest: El mismo test se aplica en dos oportunidades espaciadas por un periodo de tiempo previamente establecido. Se correlacionan los resultados por medio del Coeficiente de Pearson.
- Consistencia interna: Es uno de los procedimientos más utilizados, ya que implica una única aplicación del test. Se puede hacer con diversos procedimientos como correlación entre dos mitades de los puntajes del test o con el cálculo de Alfa de Cronbach. El coeficiente indica el grado en qué los ítems de un test convergen o están intercorrelacionados, cada ítem con el total de test.

2. Validez

El concepto de validez es amplio y complejo; ya que según el autor puede hacerse énfasis en diversos métodos, cuestiones sociales y sustantivas del constructo evaluado. Para Martínez Arias, Hernández Lloreda, & Hernández Lloreda (2006) en Muñiz, Fidalgo, García-Cueto, Martínez, & Moreno (2005) la validez:

> «*Es el más importante de los principios y nos habla del grado en que el uso que pretendemos hacer de las puntuaciones de los test está justificado. Supone examinar la red de creencias y teorías sobre las que se asientan los datos y probar su fuerza y credibilidad por medio de diversas fuentes de evidencia*».

De manera breve «*Se dice que un test es válido si mide realmente aquello que pretende medir*» Elosúa (2003); Muñiz (2003); Navas (2001) en Muñiz, Fidalgo, García-Cueto, Martínez, & Moreno (2005) .

Por otro parte, Popham (2000) en Martínez Arias, Hernández Lloreda, & Hernández Lloreda (2006) propone valiosas conclusiones acerca de lo que no es validez:

- No se valida el test, sino el uso de las puntuaciones e interpretaciones de estas.
- La validez no es un valor o punto de corte. Por lo que no es pertinente mencionar resultados "validos" o "inválidos" ya que la validez hace referencia a un nivel o grado.
- La validez es una propiedad especifica al uso o interpretación de un test, por lo que no es posible generalizar.
- Para realizar un proceso de validación no se considera solo un tipo de evidencia, sino la integración de varias que permitan argumentar y explicar las puntuaciones.

También es relevante mencionar la validez no se comprende como un índice o coeficiente como lo hace la confiablidad. Por lo tanto, el concepto actual de validez se enfoca más en:

> «*La adecuación, significado y utilidad de las inferencias específicas hechas con las puntuaciones de los tests. La validación de un test es el proceso de acumular evidencia para apoyar tales inferencias. Una variedad de evidencias puede obtenerse de las puntuaciones producidas por un test dado, y hay muchas formas de acumular evidencia para apoyar una inferencia específica*».
>
> Aliaga Tovar (2011)

a. Tipos de evidencia

Bajo el concepto de que la validez se relaciona con la interpretación e inferencias hechas a partir de las puntuaciones de un test. De tal manera que esta se centra en el grado en que la acumulación de evidencias apoya a esas inferencias. Y considerando que la acumulación de evidencias según Aliaga Tovar (2011) hace referencia a *"evidencia teórica, estadística, empírica y conceptual*". Existe una clasificación respecto al tipo de evidencias, que no desunifican el concepto integral de validez, sino que lo aborda desde otros elementos.

La Asociación de Psicología Americana (APA) propone que los tipos de validez según la evidencia que proporcionan pueden ser:

1) Validez de contenido:

Este tipo de validez busca identificar si los ítems utilizados son suficientes y significativos para evaluar un dominio de interés. Este proceso de validación se puede realizar mediante procedimientos de consulta de juicio de expertos y calcular el porcentaje de ítems que representan el dominio.

2) Validez predictiva:

> *«Se refiere a la comprobación de que el test predice un criterio externo. Se operativiza mediante el cálculo de una correlación entre el test y el criterio, que recibe el nombre de coeficiente de validez*».
>
> Muñiz, Fidalgo, García-Cueto, Martínez, & Moreno (2005).

La validez predictiva está relacionada con las evidencias referidas a criterio, lo diferencia con la validez concurrente es la temporalidad, que existen entre tomar un test y luego de un periodo tomar la medida de un criterio. Con las puntuaciones del test y las medidas del criterio se realizan correlaciones para identificar si las puntuaciones del test predicen en algún grado las medidas del criterio.

3) Validez concurrente:

Este tipo de validez también hace referencia a las evidencias de criterio al igual que la predictiva. Sin embargo, Aliaga Tovar (2011) describe que este tipo de validez difiere en que las medidas de comparación son tomadas en momentos contemporáneos y el uso se enfoca en encontrar sustitos del test si fuera conveniente. Los principales retos para las evidencias que se relacionan con un criterio es encontrar criterios confiables para poder realizar las comparaciones.

4) Validez de constructo:

Para definir la validez de constructo es importante comprender que el término constructo es un concepto hipotético, que sirve para explicar elementos de la conducta humana que son abstractos pero observables como: la creatividad, la inteligencia, la personalidad, etc. La validez de constructo es la obtención de evidencias que apoya la identificación de estas conductas observadas. Según Muñiz, Fidalgo, García-Cueto, Martínez, & Moreno (2005) algunos procedimientos para hacer una validez de constructo es el análisis factorial que representa la validez factorial, y la matriz multirasgo-multimétodo que representa la validez convergente-discriminante.

Martínez Arias, Hernández Lloreda, & Hernández Lloreda (2006) definen la validez convergente y la validez discriminante:

- Evidencia de validez convergente: Un test que proporciona buenas medidas del constructo, mostrará correlaciones altas con resultados que midan el mismo constructo.
- Evidencia de validez discriminante: Un test que proporciona buenas medidas del constructo, mostrará correlaciones bajas cuando se relaciona con otros test que miden otros constructos que no son el evaluado.

E. Validación y juicio de expertos

Uno de los métodos usuales cuando se realizan procesos de validación es el juicio de expertos. Según Escobar-Pérez & Cuervo-Martínez (2008) el juicio de expertos se define como la exposición de una "opinión informada" de un grupo de personas que demuestran dominio y trayectoria en un tema específico, que tiene como fin dar información, evidencias, juicios, valoraciones y toma de decisiones de determinado tema. Bolado, Ibáñez, & Lantarón (1998) relacionan el juicio de expertos con el término "incertidumbre de conocimiento" que surge de la necesidad de establecer información para completar sistemas en donde no hay muchas evidencias de parámetros, hipótesis, procesos, etc.

En el ámbito de medición se utilizan los juicios de expertos para validar por ejemplo el contenido y temas de una prueba. En donde se valora el porcentaje de ítems que evalúan cierto tema y los niveles de los mismos. De igual manera los juicios de expertos se pueden utilizar para establecer estándares y puntos de corte de desempeño de una prueba. Haciendo énfasis en el fundamento teórico de este estudio se profundiza en el proceso para establecer estándares de desempeño.

1. Estándares de desempeño una prueba

Cuando se habla de estándares de desempeño se relaciona el concepto "evaluación de calidad", que hace referencia a establecer criterios que permitan determinar de manera objetiva la adquisición de ciertas habilidades y destrezas. Apunta a garantizar que los procesos de aprendizaje están alineados a los mismos objetivos y que evidencian elementos fundamentales. Por medio de un juicio de experto se pueden establecer los estándares que describen el desempeño en una prueba. En relación con los estándares la Secretaría de Evaluación Educativa, Ministerio de Argentina (2016) del Ministerio de Educación en Argentina describe que hay cuatro etapas usuales para establecer estándares de una prueba, organizados de la siguiente manera:

- Primera etapa: Se debe seleccionar el número de niveles de desempeño que va a contener la prueba, se recomienda que la prueba contenga tres o como máximo cuatro niveles de desempeño, considerando que siempre se pueda diferenciar entre cada nivel.

- Segunda etapa: Se hace la elección de los nombres y etiquetas de cada nivel de desempeño. Algunas recomendaciones que propone la Secretaría de Evaluación Educativa, Ministerio de Argentina (2016) son: "Básico, competente, avanzado, elemental, Satisfactorio, Sobresaliente, por debajo de lo suficiente, alcanza objetivo satisfactorio".

- Tercera etapa: Se realiza la redacción de descriptores que implica definir los principios y habilidades generales que se alcanzan en cada nivel, sin hacer referencia a los contenidos específicos de la prueba.

- Cuarta etapa: La última etapa implica establecer los puntos de corte para cada nivel, de tal manera que se pueda clasificar a los estudiantes. Para realizar este procedimiento existen diversos métodos ya definidos. Estos procedimientos son realizados por expertos que buscar hacer una revisión del material e indicadores. Según la Secretaría de Evaluación Educativa, Ministerio de Argentina (2016) los métodos más conocidos son dos en específico: Método de Angoff (1971) y Método de Bookmark (Lewis, Mitzel y Green, 1996).

F. Método Bookmark

En la Secretaría de Evaluación Educativa, Ministerio de Argentina (2016) se describe que, en 1996 en Estados Unidos, Lewis propone una serie de procedimientos para describir e interpretar los resultados obtenidos por los alumnos en una prueba y ubicarlos acorde a una categoría que describa de mejor manera el desempeño alcanzado. Este procedimiento se realiza por medio de un juicio de expertos en donde se establecen una clasificación de los niveles desempeños representados por los estudiantes de una manera cualitativa. Luego de esto se le entrega a cada experto el test ordenado según dificultad, es decir de fácil a difícil. Con el fin de ubicar en los ítems "separadores" que establezcan la diferencia entre los niveles de desempeño propuestos.

Comúnmente se realizan por medio de tres rondas, en cada ronda los expertos hacen sus propuestas para los separadores, luego se ingresan los datos y se hace un consenso y la respectiva representación gráfica. Se repite el procedimiento similar en la segunda ronda y para la tercera se espera que la propuesta de los separadores entre los expertos demuestre más congruencia.

Es válido mencionar que en la aplicación de esta metodología existan algunos expertos que no estén de acuerdo con la propuesta del orden de los ítems según su dificultad, y esto se puede adjudicar a: la aplicación de diferentes currículos educativos si corresponde diferentes distritos o localizaciones, dificultad en los expertos para interpretar el índice

dificultad de los ítems y por último ignorar características del ítem por ejemplo si es de opción múltiple la dificultad también está representada por los distractores propuestos que hacen que un ítem que se visualiza aparentemente fácil resulte difícil para los estudiantes.

Por último, para terminar el proceso se establecen los denominados "puntos de corte", estos no son iguales que los separadores. Una de las estrategias establece que los puntos de corte se establecen haciendo uso del Modelo de la Teoría de Respuesta al ítem que permite calcular "la habilidad" para acertar en cada separador con una probabilidad del 0.67 que utiliza esta metodología como valor elemental.

> «*Los puntajes de corte son obtenidos mediante la sumatoria y promedio de los valores de habilidad theta (θ), correspondientes a la ubicación de los marcadores para cada nivel de desempeño*».
>
> Secretaría de Evaluación Educativa, Ministerio de Argentina (2016)

Aunque existen otros procedimientos para establecer los puntos de corte como tomar literalmente los separadores propuestos por los jueces. Sin embargo, Jornet Meliá y Backhoff (2006) en la Secretaría de Evaluación Educativa, Ministerio de Argentina (2016) proponen establecer "la mediana del nivel de habilidad" que corresponde a los separadores propuestos por los expertos. Para presentar los resultados de esta metodología se ubica el porcentaje de alumnos que corresponden a cada nivel.

G. Análisis de ítems

El elemento principal para el análisis de los parámetros psicométricos de un test se fundamenta en el análisis de los ítems que componen el mismo.

> *«El análisis de los ítems se define como el estudio de aquellos de sus parámetros cuyas características estén relacionadas con las propiedades y la finalidad última del test. Mejorar los ítems, o elegir los más adecuados, solo tiene como finalidad mejorar las propiedades psicométricas del test».*
>
> Muñiz, Fidalgo, García-Cueto, Martínez, & Moreno (2005)

Este análisis permite seleccionar cuáles de los ítems no aportan de manera estadística, mantener o mejorar la calidad del test.

"El análisis de los ítems depende del modelo teórico a partir del cual se hubiese construido el test." Bechger, Maris, Verstralen y Béguin, (2003); Ellis y Mead (2003); Sinar y Zickar (2002) en Muñiz, Fidalgo, García-Cueto, Martínez, & Moreno (2005).

Los modelos de medida relacionan las respuestas con el constructo que los fundamentó, los usualmente utilizados según Martínez Arias, Hernández Lloreda, & Hernández Lloreda (2006) son:

- Modelo de Teoría Clásica: Se enfoca en la puntuación total de la prueba, que se determina mediante la proporción de respuestas contestadas de manera correcta entre la cantidad total de ítems de la prueba. Tiene indicadores globales que determinan la calidad de las mediciones realizadas.
- Modelos de Teoría de Teoría de Respuesta al Ítem: Se enfoca en el análisis de cada ítem y los patrones de respuesta presentados en este. Utiliza funciones que adicionan parámetros que determinan escalas relativas que describen el comportamiento de la habilidad para responder correctamente.

H. Modelo de Teoría Clásica

Este modelo fue formulado en casi la mayoría de sus fundamentos por Spearman. La teoría estadística que lo fundamentó fue la correlacional Pearsoniana. Luego Lord y Novich propusieron la nueva formulación. El modelo propone que al momento de realizar una medición existe una puntuación observada que es el resultado de una puntuación verdadera combinada con el error. La formulación de esta teoría según se puede establecer por medio del siguiente modelo:

Fórmula 1: Modelo de Teoría Clásica

$$X_{i=}V_i + E_i$$

> *«El modelo expresa, simplemente, que la puntuación observada surge de una puntuación verdadera, V_i que es la cantidad que el sujeto posee del atributo más un error de medida, E_i. Como puede observarse, la relación entre X y E es aditiva, dando lugar a un modelo lineal*».
>
> Martínez Arias, Hernández Lloreda, & Hernández Lloreda (2006)

El modelo de Teoría Clásica propone que para garantizar la calidad de una prueba existen diversos indicadores, los más comunes según Muñiz, Fidalgo, García-Cueto, Martínez, & Moreno (2005) son:

1. Índice de dificultad:

El índice de dificultad es un valor relativo que está determinado por las personas todas las personas que intentan responder un ítem, los que lo aciertan y los que lo fallan. Se considera un ítem fácil cuando la mayoría de las personas intentan y aciertan al responderlo. Si el ítem es de opción múltiple la dificultad también se ve influenciada por el número de alternativas que propone el ítem, a esto se le conoce como efecto del azar.

Fórmula 2: Fórmula para calcular el Índice de dificultad

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

$A = Número\ de\ personas\ que\ aciertan.$
$N = Número\ de\ personas\ que\ intentaron\ responder\ el\ ítem.$

Fórmula 3: Fórmula para calcular el índice de Dificultad Corrección (efectos del azar)

$$ID = p - \frac{q}{k-1}$$

$p = La\ proporción\ de\ aciertos$
$q = La\ proporción\ de\ fallos$
$k = Número\ de\ alternativas\ que\ tenga\ el\ ítem.$

Para la interpretación del coeficiente de confiabilidad se utiliza como referencia a Vallejo (2008)

Tabla 2: Tabla de interpretación del índice de dificultad

Intervalo	**Interpretación**
0.00-0.15	Se considera muy difícil
0.15-0.40	Se considera difícil
0.40-0.60	Se considera moderado
0.60-0.85	Se considera fácil
0.85-1.00	Se considera muy fácil

Fuente: Vallejo (2008)

2. Índice de discriminación:

Discriminar implica que cada ítem contribuya a diferenciar entre las personas que tienen puntuaciones más altas y las que han obtenido puntuaciones más bajas. Un ítem que presenta una buena discriminación comúnmente es acertado por las personas que tienen puntuaciones más altas y fallado por las personas con puntuaciones bajas. Lo cual implica una correlación positiva entre las puntuaciones obtenidas en el ítem y la puntuación obtenida en la prueba. Kelley (1939) en Muñiz, Fidalgo, García-Cueto, Martínez, & Moreno (2005) propone dividir a los grupos en dos extremos, el 27% con los resultados superiores. Y el 27% de los resultados inferiores.

Fórmula 4: Fórmula para calcular el Índice de Discriminación

$$D = P_+ - P_-$$

P_+ = Proporción de acertantes al ítem del grupo superior.
P_- = Proporción de acertantes al ítem del grupo inferior.

Se propone una tabla para la interpretación del índice de discriminación Ebel (1965) en Muñiz, Fidalgo, García-Cueto, Martínez, & Moreno (2005):

Tabla 3: Interpretación del índice de discriminación

Valores	**Interpretación**
Igual o mayor que 0.40	El ítem discrimina muy bien
Entre 0.30 y 0.39	El ítem discrimina bien
Entre 0.20 y 0.29	El ítem discrimina poco
Entre 0.10 y 0.19	Ítem límite. Se debe mejorar
Menor de 0.10	El ítem carece de utilidad para discriminar

Fuente: Ebel (1965) en Muñiz, Fidalgo, García-Cueto, Martínez, & Moreno (2005)

3. Análisis de distractores:

Muñiz, Fidalgo, García-Cueto, Martínez, & Moreno (2005) describe que en una prueba de selección múltiple todas las alternativas propuestas como respuestas incorrectas representan un distractor. Estos distractores tienden a resultar igual de atractivos para todas las personas que contestan el ítem. Para verificar que todas las opciones resulten atractivas se puede verificar mediante una prueba de chi-cuadrado. La cual se verifica comúnmente con un 95% de confianza.

Fórmula 5: Fórmula para la prueba de independencia χ^2

$$\chi^2 = \sum_{i=1}^{K} \frac{(FT - FO)^2}{FT}$$

$FT = Las\ frecuencias\ teóricas.$
$FO = Las\ frecuencias\ observadas.$

Teóricamente se establece que de todos los distractores tienen que ser electos por lo menos por el 5% de las personas que contestaron el ítem. Además, se establece de todas las opciones que presenta el ítem el mayor porcentaje electo tiene que coincidir con la respuesta correcta.

4. Coeficiente de confiabilidad Alpha de Cronbach:

Existen diversos métodos ya descritos para medir el coeficiente de confiabilidad sin embargo al contar con la aplicación del test en un único momento, uno de los métodos usuales es el cálculo del coeficiente que determine la consistencia interna: coeficiente Alpha de Cronbach. (Martínez Arias, Hernández Lloreda, & Hernández Lloreda, 2006) exponen que el coeficiente Alpha de Cronbach es el equivalente a calcular la media de los coeficientes basados en pruebas de mitades y utilizar la fórmula de Spearman-Brown para estimar la fiabilidad si las partes se consideran como pruebas paralelas.

Fórmula 6: Fórmula para calcular el coeficiente Alpha de Cronbach:

$$\alpha = \left(\frac{k}{k-1}\right)\left(1 - \frac{\sum_{j=1}^{n} \sigma^2 j}{\sigma^2 x}\right)$$

$k = \ número\ de\ ítems$

$\sum_{j=1}^{n} \sigma^2 j = \ suma\ de\ varianza\ de\ los\ ítems$

$\sigma^2 x = \ varianza\ de\ los\ totales$

Para la interpretación del coeficiente Alpha de Cronbach se utiliza la propuesta de Cortina (1993) en Tarqui (2017)

Tabla 4: Tabla para interpretación del coeficiente Alpha de Cronbach

Alpha de Cronbach	**Nivel de confiabilidad**
α = 0.9	Excelente
0.8 = α < 0.9	Bueno
0.7 = α < 0.8	Aceptable
0.6 = α < 0.7	Cuestionable
0.5 = α < 0.6	Poble
α < 0.5	Inaceptable

Fuente: Cortina (1993) en Tarqui (2017)

I. Modelo de Teoría de Respuesta al Ítem

La Teoría de Respuesta al ítem (TRI) al igual que la Teoría Clásica de los Test (TCT) tiene como fin estimar el nivel de habilidad que posee el evaluado, el denominado rasgo latente. Tal y como su nombre lo indica en TRI la unidad de análisis son los ítems, mientras que en TCT son las puntuaciones obtenidas en el test. TRI permite visualizar en el ámbito educativo una prueba desde una perspectiva diferente.

> «*La utilidad de esta teoría en el campo educativo radica en determinar si un estudiante consigue responder correctamente a cada una de las preguntas (ítems) y no al puntaje bruto obtenido en la prueba (test)*».
>
> Carvajal Álzate, Méndez Sánchez, & Torres Angulo (2016)

1. Diferencias entre TRI y TCT:

Embrestson y Reise (2000) en Martínez Arias, Hernández Lloreda, & Hernández Lloreda (2006) propone las diferencias elementales entre las reglas de TRI respecto TCT.

- TRI no generaliza el error ya que difiere entre puntuaciones o patrones de respuesta.
- Los test cortos también son fiables.
- No es necesario que los test sean paralelos para poder compararse.
- Los parámetros están dados por los ítems y no por las muestras.
- Para hacer la interpretación de los resultados se compara la distancia entre ítems.
- Para establecer las propiedades de las escalas de intervalo se aplica el modelo de Rasch.
- Se pueden utilizar formatos diferentes para ítems.
- Puntuaciones de cambio son fácilmente establecidas.
- El análisis factorial de ítems lleva al análisis factorial de la información.
- Las propiedades psicométricas se enfatizan en los ítems.

De igual manera Muñiz, Fidalgo, García-Cueto, Martínez, & Moreno (2005) propone dos diferencias más entre TRI y TCT.

- TRI incorpora parámetros al modelo para describir de manera más precisa las características de los ítems.
- Los supuestos sobre los que descansan los supuestos en TRI son diferentes a los de TCT.

2. Supuestos

En TRI se asume que los datos de las respuestas propuestas en los ítems deben cumplir con cierto supuestos. Antes de la aplicación de TRI se establece necesario la verificación de los mismos. Según Muñiz, Fidalgo, García-Cueto, Martínez, & Moreno (2005) dos:

- Supuesto de unidimensionalidad: Este supuesto establece que cada ítem debe medir únicamente el rasgo latente establecido, por lo que cada respuesta se ve influenciada en representar una única dimensión evaluada, esto es posible de verificar por medio de un análisis factorial.
- Supuesto de independencia local: Se considera que responder cada ítem es un evento de probabilidad independiente, por lo que la probabilidad de responder correctamente un ítem es independiente a la probabilidad de responder cualquier ítem.

3. Modelos:

Tal y como se ha descrito anteriormente el modelo TRI se enfoca en el análisis de los ítems, una de las principales herramientas para su análisis es la de nominada Curva Característica del ítem (CCI).

> "*La CCI es una función matemática que relaciona la probabilidad de responder correctamente al ítem con el nivel de habilidad que tiene en la variable medida por el ítem quien responde a él*".
>
> Muñiz, Fidalgo, García-Cueto, Martínez, & Moreno (2005)

Este modelo está determinado por una función logística:

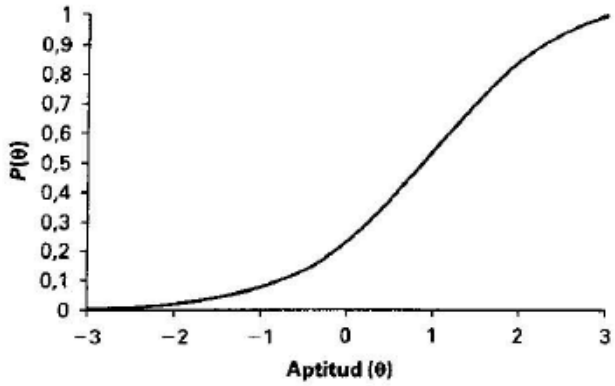
Fórmula 7: Modelo de Teoría de Respuesta al Ítem

$$y = \frac{e^x}{1 + e^x}$$

$e = \ es\ una\ constante, la\ base\ de\ los\ logaritmos\ neperianos\ 2.718$

$x \ = \ es\ cualquier\ valor\ o\ función$

Ilustración 1: Curva característica del ítem



Fuente: Martínez Arias, Hernández Lloreda, & Hernández Lloreda (2006)

La teoría de los ítems implica establecer que la probabilidad de responder correctamente un ítem está determinada por el nivel de habilidad que posee la persona para determinado rasgo latente, por consiguiente, una persona con mayor nivel de habilidad tiene una mayor posibilidad de responder correctamente el ítem. Sin embargo, es importante mencionar que los ítems suelen diferir entre diversos parámetros como lo son: dificultad, discriminación y la probabilidad de contestar al azar. Por lo que la CCI es diferente para cada ítem y los modelos de TRI al adicionar cada parámetro permiten hacer una descripción más precisa d de la medida del dominio rasgo latente.

4. Modelo de Rasch

Según Muñiz, Fidalgo, García-Cueto, Martínez, & Moreno (2005) el Modelo de Rasch es el modelo más simple de TRI. Este es un modelo logístico que incorpora el parámetro de la dificultad para la medida del rasgo latente. Este modelo considera que no hay aciertos al azar y que los ítems poseen la misma discriminación.

Fórmula 8: Función que describe el Modelo de Rasch

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

$P_i(\theta) = \ Probabilidad\ de\ acertar\ al\ ítem\ i\ para\ un\ valor\ \theta$
$b_i = \ Índice\ de\ dificultad\ de\ ítem$
$e = \ Base\ de\ los\ logaritmos\ neperianos$
$D \ = \ Constante,\ cuando\ toma\ el\ valor\ 1.7,$
$la\ función\ logística\ se\ aproxima\ a\ la\ normal\ acumulada.$

Para comprender el comportamiento y análisis de los modelos de TRI es importante definir que representa los parámetros utilizados.

a. El parámetro θ:

La representación de la letra griega θ Theta hace referencia en TRI al constructo o rasgo latente que mide el test. Muñiz, Fidalgo, García-Cueto, Martínez, & Moreno (2005) establece que *"Fundamentalmente solo podamos determinar las posiciones relativas de los individuos en el rasgo latente, sus distancias y en ningún caso comprobar directamente los valores en dicho rasgo."* Esto se debe a que las propiedades métricas de θ hacen referencia a una escala de intervalo. Por lo que el "0" no es absoluto y no representa la ausencia del constructo. En este caso específico un $\theta = 0$ representa a una persona con habilidad promedio. Y por lo tanto los valores negativos de θ indican que la habilidad de la persona está por debajo del promedio, mientras que θ positivo indica que la habilidad de la persona está por arriba del promedio.

b. Parámetro b:

Según Muñiz, Fidalgo, García-Cueto, Martínez, & Moreno (2005) este parámetro hace referencia a la dificultad del ítem en TRI. De igual manera medido en la misma escala que θ . Este parámetro indica que cuanto más difícil sea un ítem mayor será el nivel de habilidad necesario para tener una probabilidad de acertarlo correctamente.

V. ANTECEDENTES

Los antecedentes proporcionan la revisión bibliográfica que describe el estado del arte de estudios similares a esta investigación en los últimos 10 años. Con el fin de visualizar de manera general los alcances, las limitaciones, metodologías, procedimientos y los hallazgos de los mismos. Y de esta manera fundamentar el modelo a utilizarse en esta investigación y tomar como referencias las experiencias que sean adaptables para este contexto.

Bombelli (2011) establece que un diagnóstico educativo permite orientar los procesos de intervención del docente. Aspectos esenciales como el tiempo que se dedicará a cada tema y el cambio a las prácticas docentes. En su estudio realizado para conocer: La Importancia de la Evaluación Diagnóstica en Asignaturas a Nivel Superior con conocimiento Preuniversitario. Se concluye que es alto el porcentaje de los alumnos que obtienen bajos resultados en estas pruebas y que llevaron estas asignaturas en nivel medio. Lo más relevante es que se reconoce que el aprendizaje es complejo y multicausal y que el bajo rendimiento de los alumnos en estas pruebas se asocia a la calidad de enseñanza que se imparte en el nivel medio.

Según Zamora Araya (2013) la prueba de diagnóstico de matemática en la Universidad Nacional Autónoma de Costa Rica surge como una necesidad de obtener información ante los bajos resultados obtenidos por los alumnos en los cursos introductorios de matemática. Su investigación propone un análisis psicométrico de la prueba utilizando el modelo de Rasch. Dicho modelo, seleccionado por trabajar de menor manera ciertas dificultades que presenta normalmente la Teoría Clásica. El análisis permitió brindar recomendaciones sobre la construcción de la prueba, estas recomendaciones por consiguiente mejoraron y aumentaron la confiabilidad de esta en comparación a la aplicación del 2010 y la del 2012, en donde el Alfa de Cronbach aumentó 0.27.

Según Jiménez Alfaro & Montero Rojas (2013) en su estudio de: La aplicación del modelo de Rasch, en el análisis psicométrico de una prueba de diagnóstico en matemática; se realizó un análisis a los casos de la prueba de diagnóstico de conocimientos y destrezas en matemática del estudiante para ingresar a la Universidad de Costa Rica en el 2018, utilizando el modelo de Rasch para sus análisis. Se determinó que, al aplicar este modelo se obtiene una medida de confiabilidad de los evaluados y que la correspondiente de los ítems resultaron bastante consistentes y existe una medición bastante precisa en cuanto a la dificultad de los ítems. Esto es relevante ya que el objetivo de la prueba es contribuir a la solución de los problemas de bajo rendimiento académico que tienen los estudiantes al ingresar a la universidad mediante el diagnóstico de la prueba. También se encontró que el uso de este modelo permite hacer una mejor selección de expertos que conocen los constructos que analizan y clasifican los ítems, según su dificultad, contenido y procesos presentes en su solución.

Así mismo, en el estudio del análisis de una prueba de diagnóstico en matemática del Instituto Tecnológico de Costa Rica, aplicada a estudiantes de nuevo ingreso, Ramírez & Barquero (2011) cuentan con un análisis de probabilidad condicional relacionado con el rendimiento de cursos introductorios de matemática, con relación a variables como sexo, tipo de financiamiento del colegio donde egresa, el lugar de procedencia, etc. Además, de un análisis psicométrico que expone la validez de tipo predictiva con relación al rendimiento académico de los alumnos. Para estos cursos en donde se demuestra una problemática de aprobación. Entre los hallazgos relevantes se concluye que los bajos resultados obtenidos en esta prueba, se deben al reflejo del “pobre bagaje matemático” que adquirieron los estudiantes, antes de ingresar al ITCR.

De acuerdo con Cerdas Núñez & Montero Rojas (2017) el modelo de Rasch permite la construcción de pruebas más adecuadas y eficientes en la elaboración de pruebas de admisión. En su estudio: *Uso del Modelo de Rasch Para la Construcción de Tablas de Especificaciones: Propuesta metodológica aplicada a una prueba de selección universitaria*; se determinó que el uso de un modelo psicométrico permite evidenciar el proceso de validación de los ítems y como estos pronosticarán el desempeño de un

estudiante de nuevo ingreso a la Universidad de Costa Rica. Dentro de los aportes que se lograron se establece: Una primera aproximación de tabla de especificaciones para el Programa Permanente Prueba de Aptitud Académica, se definieron contenidos y procesos representativos del conjunto de ítems, a partir de esta validación se propuso que estos pueden servir como una herramienta para la creación y juzgamiento análisis de nuevos ítems.

En el estudio: *Calibración de Instrumentos de Evaluación, Clasificación en Matemáticas en la Universidad Jorge Tadeo Lozano* de Bogoya, Barragán, Contento, & Ocaña (2014) se expone acerca la prueba de conocimientos básicos que aplican a los estudiantes de nuevo ingreso. En 2011 se empleó TRI, con un parámetro (modelo de Rasch) para el análisis de datos. Dentro de las conclusiones se encontró que el modelo de Rasch con el parámetro, el de la dificultad, fue una herramienta valiosa para procesar los datos del examen porque logró un análisis conjunto de los parámetros de los ítems y del instrumento. Se logró mejor diagnóstico individual para los evaluados sobre sus conocimientos básicos de matemáticas y pudo optimizar los procesos para promover a los estudiantes para que pudiera continuar con su plan de estudios que tenían previsto.

En primera instancia los estudios mencionados anteriormente, fundamentan la importancia de la aplicación de una prueba de diagnóstico para el área de matemática a nivel universitario. Siendo esta prueba una herramienta para hacer recolección valiosa de información en referencia a dominio básico de la materia. Además, con los estudios de Zamora Araya (2013), Jiménez Alfaro & Montero Rojas (2013), Cerdas Núñez & Montero Rojas (2017), Bogoya, Barragán, Contento, & Ocaña, (2014) como fundamento se puede establecer que el modelo de Rasch es funcional para este tipo de análisis y que puede proporcionar elementos para mejorar la prueba y de esta forma aumentar la confiabilidad en futuras aplicaciones.

Por lo tanto, el énfasis por analizar una prueba de diagnóstico de matemática se deriva de la idea de que la prueba sea lo suficiente válida y confiable; con la capacidad de identificar las habilidades que presentan mayor dificultad los alumnos y buscar estrategias para mejorar los aprendizajes. No con el fin de enfocarse en incrementar los índices de aprobación, sino en la búsqueda de las estrategias pertinentes y funcionales a los docentes que permitan a los alumnos la mejora de su aprendizaje partiendo de su situación actual. Además de justificar la importancia de la aplicación de una prueba de diagnóstico de matemática a los estudiantes de primer ingreso en la universidad, verificar que la prueba cumpla con propiedades e indicadores psicométricos que garanticen calidad de la misma y la importancia del análisis de los resultados utilizando el modelo de Rasch, este estudio también se fundamenta en aplicar la metodología Bookmark para establecer los niveles de desempeños en los cuales se ubican los estudiantes.

La Secretearía de Evaluación Educativa, Ministerio de Argentina (2016) desarrolló una serie de talleres denominados: Aprender 2016 Bookmark establecimiento de puntos de corte. Con el reto de trabajar con 201 docentes y establecer los separadores de diversas pruebas. Se tomó referencia la clasificación de niveles según el Operativo Nacional de Evaluación: Por debajo del nivel básico, básico, satisfactorio, avanzado. La diversidad de elección de separadores fue llegando a un punto en común en medida que incrementaban las rondas. Entre las conclusiones se menciona que establecer los estándares apunta a generar evaluaciones confiables y validas, que permiten interpretar y visualizar los resultados, además de promover el diseño e implementación de cambios en cuanto a la planificación e incluso políticas educativas que beneficien la mejora de la calidad educativa.

De igual manera el método de Bookmark ha sido aplicado en Guatemala, la Dirección General de Evaluación e Investigación Educativa, Ministerio de Educación (2010) utilizó este procedimiento para establecer los “separadores o Bookmark” que clasifican los niveles desempeño en las pruebas para las carreras de Secretaríado y Perito Contador. Con el fin de interpretar de una manera pertinente los resultados obtenidos por los estudiantes y clasificar a la población con relación al nivel detectado. Los niveles de desempeño ya

establecidos son: Insatisfactorio, Debe mejorar, Satisfactorio, Excelente. Además de explicar de manera detallada la aplicación de la metodología, segmentar la prueba con separadores y visualizar el porcentaje de alumnos que pertenecen en cada nivel se encontró que, luego de la clasificación el mayor porcentaje de alumnos se concentraba en el nivel Insatisfactorio en cada ronda. Como propuesta final se determinó que para la prueba que consta de 45 ítems se concluyó el primer punto de corte está ubicado en el ítem 9 (debe mejorar), el segundo punto de corte se ubica en el ítem 21 (satisfactorio), el último punto de corte se ubicó en el ítem 31 (excelente). En esta última y consensuada ronda permitió distribuir de una manera bastante equitativa a los alumnos en cada nivel.

VI. METODOLOGÍA

La metodología de investigación permite describir el procedimiento utilizado para el desarrollo de un estudio. Responde a las preguntas: ¿cómo se realizó?, ¿de qué manera?,¿quiénes son los involucrados? y las variables a estudiar. Con la posibilidad de replicar estos procedimientos para futuras investigaciones. Para esta investigación la metodología pretende, organizar una serie de pasos que permitan realizar el análisis de confiablidad y validez de una prueba de diagnóstico de matemática.

A. Formulación del problema

Análisis psicométrico de la prueba de diagnóstico de matemática de la Universidad del Valle de Guatemala.

B. Preguntas de investigación

1. Pregunta central:

¿En qué medida los resultados de la prueba de diagnóstico del Departamento de Matemática de la UVG Campus Central, cumplen con los indicadores psicométricos que demuestran la calidad de la misma?

2. Preguntas secundarias:

- ¿Cuáles son los indicadores psicométricos de la prueba de diagnóstico del Departamento de Matemática de la UVG con los resultados años 2016,2017 y 2019?
- ¿Cuáles son los niveles y los descriptores que determinan el desempeño de los estudiantes en la prueba de diagnóstico de matemática que aplica el Departamento de Matemática de UVG?
- ¿Qué porcentaje de estudiantes se ubican en cada nivel de desempeño de la prueba de diagnóstico de matemática que aplica el Departamento de Matemática de UVG?

C. Enfoque de investigación

Para este estudio el enfoque seleccionado es el cuantitativo ya que de acuerdo con Hernández Sampieri (2010) este enfoque utiliza diversos análisis numéricos de variables seleccionadas que permiten comprobar hipótesis y teorías. Se asocia con la búsqueda de la verdad objetiva. Aravena, Kimelman, Micheli, Torrealba, & Zúñiga (2006)

Las características fundamentales de este enfoque son:

- Este tipo de investigación es lo más objetiva posible.
- Estos estudios tienen un patrón predecible y estructurado.
- Se pretenden generalizar los resultados de las muestras a poblaciones.
- La meta principal es la construcción y demostración de teorías que "explican y predicen".
- El proceso es riguroso con ciertas reglas lógicas, lo que permite que los datos generados tengan validez y confiabilidad.

Específicamente este estudio se fundamenta en el enfoque cuantitativo. Dado que, se pretenden determinar mediante un análisis de ítems que se fundamenta en la Teoría Clásica y Teoría de Respuesta al Ítem, en qué medida los resultados de la prueba de diagnóstico de matemática, que utiliza el Departamento de Matemática de la UVG cumplen con indicadores y parámetros psicométricos, que garantizan la calidad de esta. La explicación de los resultados de este análisis está acompañada por un juicio de expertos, que por medio de la metodología de Bookmark establece los puntos de corte que determinan los niveles y descriptores de desempeño. Con el fin de clasificar a los alumnos en el nivel que describe su desempeño alcanzado en la prueba e interpretar de mejor manera las puntuaciones obtenidas.

Este enfoque es el más conveniente para esta investigación pues busca visualizar de manera precisa si la prueba en realidad está funcionando, si discrimina entre los alumnos de diferente nivel de dominio, si principalmente proporciona puntuaciones lo más apegadas a la realidad. Por otra parte, el juicio de expertos desarrollado con la metodología Bookmark se trabaja de una manera muy sistemática y busca llegar a un consenso de la ubicación de puntos de corte que segmentan los niveles de desempeño. Y de esta manera se clasifique el porcentaje de estudiantes que alcanzan determinado nivel para diseñar estrategias que busquen mejorar el aprendizaje. Dado que, el objetivo es determinar las medidas en las que se está cumpliendo la calidad e interpretar las mismas y la única manera de visualizar esto es bajo el análisis de datos trabajados con el enfoque cualitativo.

D. Tipo de investigación

El tipo de investigación seleccionado para este estudio es no experimental, este por concepto permite observar situaciones ya existentes, realizadas en su entorno natural y sin ninguna manipulación de variables; con el fin de analizarlas después que ya sucedieron. Es el análisis de hechos que ya sucedieron con sus respectivos efectos, este tipo de investigación se hace pertinente en estudio de tipo retrospectivos y prospectivos. El estudio no experimental abordado específicamente con un diseño longitudinal de tendencia que permite analizar cambios a través del tiempo con un principal en énfasis en las poblaciones, en donde los participantes pueden no ser los mismos a lo largo de los años, pero sí pertenecientes a la misma población Hernández Sampieri (2010).

La selección específica de que el estudio sea de tipo no experimental surge de contar con los resultados ya existentes de las pruebas de diagnóstico que utiliza el Departamento de Matemática de la Universidad del Valle de Guatemala Campus Central. Esta prueba explora el dominio de los conocimientos que los alumnos de primer ingreso demuestran en el área de matemática principalmente en aritmética y algebra. Al contar con esa información no se realiza una manipulación de los datos, porque estos ya fueron recopilados en años anteriores.

Se seleccionó el diseño longitudinal de tendencia ya que se cuenta con los resultados de los alumnos del 2016,2017 y 2019. Con dichos resultados se puede identificar la tendencia de los resultados a lo largo de los años y además es posible realizar diversos análisis que permiten determinar la medida en que la prueba de diagnóstico demuestra calidad con una muestra representativa de diversos grupos. Además, permite ubicar el porcentaje de alumnos que corresponden a cada nivel de desempeño ya sea de manera general o por año.

E. Población, muestra y unidad de análisis o sujetos de investigación.

En este estudio la población está comprendida por los 1741 estudiantes de primer ingreso de la Universidad del Valle de Guatemala Campus Central que según los datos del Registro Académico fueron inscritos en el primer ciclo de los años 2016, 2017 y 2019 pertenecientes a la Facultades Ingeniería, Ciencias Sociales, Ciencias y Humanidades.

La muestra representa un subgrupo de la población. Para este estudio la elección de la muestra es de tipo no probabilístico. Ya que de acuerdo con Hernández Sampieri (2010) este tipo de muestra no depende de la probabilidad sino de las características de la población. Por consiguiente, para este estudio la muestra está comprendida por los 1172 estudiantes que tienen la característica de estar asignados al curso de Pensamiento Cuantitativo del primer ciclo en los años 2016, 2017 y 2019 y que se contaba con los resultados documentados en las bases de datos en esos años. El curso de Pensamiento Cuantitativo es el curso inicial en el área de matemática que reciben los alumnos de primer ingreso, impartido en las Facultades de Ingeniería y Facultad de Ciencias y Humanidades. La distribución de los estudiantes se organiza de la siguiente manera:

Tabla 5: Distribución de estudiantes de la muestra según año

Año	Frecuencia	Porcentaje
2016	303	25.9
2017	406	34.6
2019	463	39.5
Total	1172	100.0

Fuente: Elaboración propia según los resultados de la aplicación de la prueba de diagnóstico de los años 2016, 2017 y 2019

La unidad de análisis son las respuestas proporcionadas por los estudiantes de primer ingreso en la prueba de diagnóstico de matemática en los años 2016, 2017 y 2019. Estas se obtienen al aplicar la prueba de diagnóstico al inicio del primer semestre en el mes de enero. Recopiladas por los catedráticos del Departamento de Matemática de la Universidad del Valle de Guatemala que imparten el curso de Pensamiento Cuantitativo. Otra unidad de análisis son los ítems que integran la prueba, estos como elementos que se someten a análisis con relación a las respuestas documentadas por los estudiantes; que permiten determinar los indicadores psicométricos de la prueba.

F. Instrumentos o técnicas que utilizará la recolección de datos.

El principal instrumento que permite recolectar la información para esta investigación es la prueba de diagnóstico de matemática utilizada en los años 2016, 2017 y 2019 por el Departamento de Matemática de la UVG campus central. La prueba tiene como objetivo identificar conocimientos y habilidades básicas en el área, que los alumnos de primer ingreso poseen al momento de iniciar su carrera.

Los antecedentes de la prueba utilizada indican que no es una prueba estandarizada, sino que es la integración de dos pruebas diseñadas, piloteadas, validadas y aplicadas por la directora del Departamento de Matemática y una de las catedráticas que imparten cursos en el mismo; elaboradas de manera individual como trabajo de graduación para optar al grado de Maestría en Medición Evaluación en Investigación Educativa en el año 2009 en UVG. Se realizaron seis formas de la prueba, para generar dos propuestas finales. Con el fin de proponer una visión que diagnóstica los alcances de los alumnos que ingresan a la universidad en área de matemática. Una de las pruebas fue trabajada bajo la Taxonomía de Marzano y otra con la Taxonomía de Bloom, ambas validadas con un juicio de expertos en relación al contenido y en uno de los casos se utilizó específicamente el método de Angoff para esta validación. Para ambas pruebas los estudiantes hicieron uso de un tiempo entre los entre veinte minutos hasta una hora, para la resolución de la misma. Las propuestas finales luego del análisis respectivo cuentan con 30 y 22 ítems. Con una confiablidad alcanzada entre (0.72-0.79) según el coeficiente del Alpha de Cronbach, valores aceptables

para no ser una prueba estandarizada. Aplicadas aproximadamente en uno de los casos a una muestra de 349 sujetos en 4 instituciones. Y en el otro de los casos a 513 estudiantes de dos instituciones. En el año 2014 se decide crear una versión integrada de prueba de diagnóstico de matemática para los para los estudiantes de primer ingreso de UVG específicamente. De las pruebas originales se seleccionaron 26 ítems que pretenden medir las habilidades de los alumnos de primer ingreso respecto a las áreas de algebra y aritmética. Cabe mencionar que, si bien para este estudio no se realizó un pilotaje con la prueba de los 26 ítems, es porque se trabajó el análisis de los resultados ya existentes. Se buscó realizar un análisis para ver la calidad de la versión que había sido utilizada en la realidad, con el fin explorar y describir resultados que permitan de tomar decisiones y modificaciones respecto a esta prueba utilizada. Se puede mencionar que, aunque la prueba como tal no fue piloteada los ítems de la misma sí.

Otro instrumento se denomina “cuadernillo de Bookmark” que es la misma prueba de diagnóstico, pero ordenada según el índice de dificultad. Este cuadernillo es utilizado en la aplicación de la metodología de Bookmark que establece que por medio de un juicio de expertos una revisión de los ítems del test y la selección de los “separadores” que permiten establecer los puntos de corte, entre los niveles de desempeño que proporcionan una interpretación de los resultados obtenidos en la prueba de diagnóstico. Esta metodología se realiza comúnmente en tres rondas en donde los expertos identifican los ítems que consideran como “separadores”. luego se realiza un consenso de los resultados y se repite la actividad.

Los métodos estadísticos utilizados están regidos por la Teoría Clásica que proporciona indicadores psicométricos que permiten establecer la calidad de la prueba. Y también la Teoría de Respuesta al Ítem, específicamente con el modelo de Rasch. Para Zamora Araya (2013) el modelo Rasch propuesto por Georg Rasch propone solventar muchas deficiencias que presenta la Teoría clásica. Relaciona matemáticamente la habilidad de la persona y la dificultad del ítem. Que permite realizar comparaciones sobre la misma escala. Los análisis fueron realizados en el programa especializados en análisis estadístico y psicométrico: Jmetrik.

G. Alcances y limitaciones del modelo de trabajo.

El alcance de este estudio es de tipo descriptivo, de acuerdo con Hernández Sampieri (2010), los estudios que tienen alcance descriptivo permiten identificar propiedades, características, perfiles y rasgos importantes de un fenómeno que se analiza, además profundizan con precisión en el análisis variables o aspectos específicos. La característica valiosa de estos estudios es que buscan medir información de manera independiente entre las variables que se analizan, sin necesidad de buscar relaciones entre las mismas. Según Hernández Sampieri (2010) los estudios con este tipo de alcance no suelen proponer hipótesis, debido que es complicado proponer o predecir un valor que se manifiesta en una variable. Para esta investigación lo que se pretende en el análisis es identificar el nivel o medida en que la prueba de diagnóstico proporciona resultados confiables y válidos. Y a su vez, establecer niveles de desempeño que permitan realizar interpretaciones de las puntuaciones obtenidas por los alumnos.

Entre las limitaciones para este estudio se puede mencionar que por razones administrativas en el año 2018 no se realizó la aplicación de la prueba de diagnóstico en enero y por consiguiente no se consideró este año para dicho análisis. Tampoco se cuenta con otro "criterio confiable" ya sea nota final del curso u otra prueba, que permita verificar la validez de tipo predictiva que pueda tener la prueba de diagnóstico con relación a la predicción del rendimiento o éxito en el curso de Pensamiento Cuantitativo.

Con respecto al desarrollo del juicio de expertos utilizando la metodología de Bookmark cabe mencionar que únicamente se trabajó con cinco catedráticos del Departamento de Matemática ya que por razones de diversidad de horario otros catedráticos que también tienen amplia experiencia en el curso de Pensamiento Cuantitativo no pudieron participar en la actividad a pesar de ser invitados. Por otra parte, al momento de hacerle la propuesta a los catedráticos del cuadernillo de Bookmark estos manifestaron desacuerdo en el orden según dificultad y esto genero una discusión de lo incensario que era separar los niveles. Sin embargo, luego de varias discusiones lograron

coincidir en ciertos aspectos y establecer los niveles, pero solo se desarrolló en una sola ronda ya que la coincidencia fue múltiple en todos los casos de las separaciones.

H. Pasos o fases de investigación.

Para el desarrollo de esta investigación se utilizarán las siguientes fases:

1. Fase de selección del tema

Con el fin de contribuir a resolver una necesidad e interrogante real. Se establece el tema de investigación mediante discusiones con catedráticos y directora del Departamento de matemática, la necesidad de analizar los resultados de la prueba de diagnóstico en un rango determinado de años.

2. Fase de revisión bibliográfica

Para esta revisión los temas a investigar están relacionados con: confiabilidad, validez, validación de pruebas de matemática y técnicas para el análisis de ítems con modelo de Teoría Clásica y modelo de Teoría de Respuesta al ítem. Así como también la revisión del estado del arte de estudios similares.

3. Fase de planteamientos de investigación

Para esta fase se procede a plasmar los elementos que le dan estructura a la investigación como el tipo, diseño, problema, hipótesis, población, muestra, unidad de análisis, objetivos e instrumentos.

4. Fase de análisis de ítems

Organización de las bases de datos existentes, con los resultados de las pruebas de diagnóstico por año. Luego se realizará un análisis de ítems según el modelo de Teoría

Clásica y de Teoría de Respuesta al ítem. Dicho análisis permitirá conocer los indicadores psicométricos de la prueba.

5. Fase de validación por juicio de expertos

Con los resultados de los indicadores psicométricos de los resultados de la prueba de diagnóstico años 2016, 2017, 2019 se organizan los ítems de la prueba según el índice de dificultad. Se realiza un juicio de expertos aplicando la metodología de Bookmark para establecer los puntos de corte que separan los niveles de desempeño y que permiten interpretar las puntuaciones de la prueba.

6. Fase de categorización de los estudiantes

Se categoriza a los estudiantes según su desempeño alcanzado, con base a los niveles establecidos por el juicio de expertos con la metodología de Bookmark.

7. Fase de hallazgos y resultados

Se establecen las conclusiones y hallazgos encontrado del análisis de los indicadores psicométricos y el juicio de expertos. Se proponen recomendaciones al Departamento de Matemática en relación con la prueba de diagnóstico que realizan los alumnos de primer ingreso.

VII. PRESENTACIÓN Y ANÁLISIS DE RESULTADOS

A continuación, se presenta el análisis de resultados de este estudio; que como se mencionó anteriormente, se trabajó con los resultados de la prueba de diagnóstico de matemática que aplicó el Departamento de Matemática de UVG Campus Central a 1172 estudiantes de primer ingreso de los años 2016, 2017 y 2019. Los resultados se analizaron según el modelo de análisis de los ítems TCT y TRI, además de la documentación de la aplicación de la metodología Bookmark. Los análisis de ítems realizados se procesaron en el programa Jmetrik.

A. Análisis de ítems Teoría Clásica de los Test

El análisis de ítems de la TCT implica la revisión de los indicadores psicométricos y el comportamiento que tienen los resultados en cada uno de estos.

1. Índice de dificultad:

En la Tabla 1 se presenta el índice de dificultad para cada uno de los ítems acompañado de la respectiva interpretación Vallejo (2008). Con los resultados expuestos se pude identificar que 8 de los 26 ítems de la prueba de diagnóstico que utiliza el Departamento de Matemática se encuentran en la categoría moderado respecto a su índice de dificultad. Es importante resaltar que los ítems 24 y 25 son clasificados como muy difíciles. Y según el contenido que estos miden se puede mencionar que los alumnos presentan dificultad en la simplificación de expresiones logarítmicas y exponenciales.

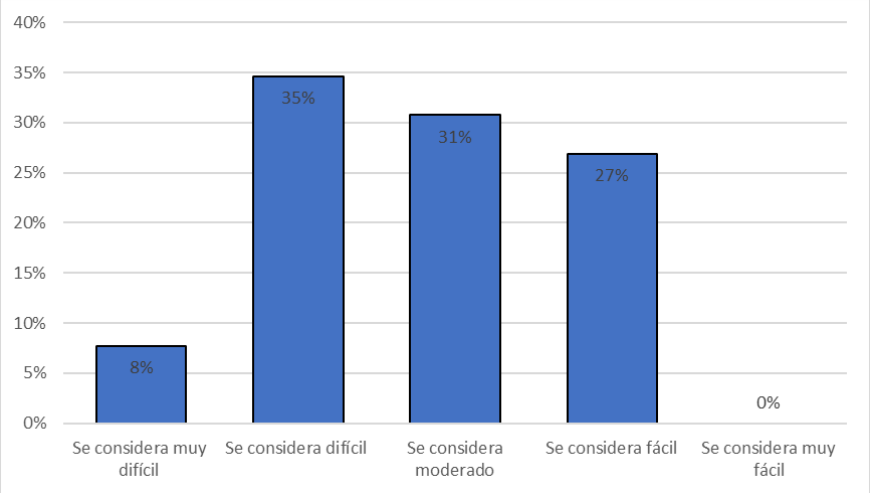
Tabla 6: Índice de dificultad de cada ítem de la prueba

No. de Ítem	Índice dificultad TCT	Interpretación
P1	0.55	Se considera moderado
P2	0.17	Se considera difícil
P3	0.19	Se considera difícil
P4	0.64	Se considera fácil
P5	0.54	Se considera moderado
P6	0.55	Se considera moderado
P7	0.24	Se considera difícil
P8	0.21	Se considera difícil
P9	0.67	Se considera fácil
P10	0.55	Se considera moderado
P11	0.72	Se considera fácil
P12	0.25	Se considera difícil
P13	0.56	Se considera moderado
P14	0.31	Se considera difícil
P15	0.43	Se considera moderado
P16	0.52	Se considera moderado
P17	0.33	Se considera difícil
P18	0.80	Se considera fácil
P19	0.63	Se considera fácil
P20	0.59	Se considera moderado
P21	0.17	Se considera difícil
P22	0.73	Se considera fácil
P23	0.10	Se considera muy difícil
P24	0.07	Se considera muy difícil
P25	0.34	Se considera difícil
P26	0.63	Se considera fácil

Fuente: Elaboración propia según los resultados obtenidos de la prueba de diagnóstico años 2016, 2017 y 2019

En la Figura 1 se muestra la distribución porcentual de los ítems según la categoría de correspondencia. Cabe mencionar que ninguno de los ítems se clasificó como “muy fácil”. En relación con los ítems que se categorizan como fáciles los contenidos que se buscan medir son: la simplificación algebraica, aplicación de reglas de los radicales y resolución de desigualdades lineales. Si se considera común que en una prueba existan ítems clasificados como difíciles, moderados y fáciles; según los resultados de la Figura 1 los ítems de esta prueba apuntan a que el 92% se encuentren entre estas categorías, distribuidos de una manera bastante equitativa. Aunque cabe mencionar que el 43% de los ítems se encuentran entre muy difícil y difícil.

Gráfica 1: Distribución porcentual de ítems según índice de dificultad



Fuente: Elaboración propia según los resultados obtenidos de la prueba de diagnóstico años 2016, 2017 y 2019

2. Índice de discriminación:

En la Tabla 2 se representa el índice de discriminación para cada uno de los ítems, con la respectiva interpretación según Ebel (1965) en Muñiz, Fidalgo, García-Cueto, Martínez, & Moreno (2005). Es relevante mencionar que 13 ítems se encuentran en las categorías de discriminar bien y discriminar muy bien. Mientras que los ítems 3,21,24 indican que discriminan poco; los ítems 10,11,20,23 se encuentran al límite del índice de discriminación y; y los ítems 9,18,19,22,25,26 carecen utilidad para discriminar.

Tabla 7: Índice de discriminación de cada ítem de la prueba

No. de ítem	**Índice de discriminación**	**Interpretación según Ebel (1965) en Muñiz, Fidalgo, García-Cueto, Martínez, & Moreno (2005)**
Ítem 1	0.37	El ítem discrimina bien
Ítem 2	0.57	El ítem discrimina muy bien
Ítem 3	0.25	El ítem discrimina poco
Ítem 4	0.36	El ítem discrimina bien
Ítem 5	0.37	El ítem discrimina bien
Ítem 6	0.31	El ítem discrimina bien
Ítem 7	0.57	El ítem discrimina muy bien
Ítem 8	0.62	El ítem discrimina muy bien
Ítem 9	-0.41	El ítem carece de utilidad para discriminar
Ítem 10	0.18	Ítem límite. Se debe mejorar
Ítem 11	0.17	Ítem límite. Se debe mejorar
Ítem 12	0.35	El ítem discrimina bien
Ítem 13	0.35	El ítem discrimina bien
Ítem 14	0.50	El ítem discrimina muy bien
Ítem 15	0.37	El ítem discrimina bien
Ítem 16	0.41	El ítem discrimina muy bien
Ítem 17	0.57	El ítem discrimina muy bien
Ítem 18	0.05	El ítem carece de utilidad para discriminar
Ítem 19	-0.19	El ítem carece de utilidad para discriminar
Ítem 20	0.12	Ítem límite. Se debe mejorar
Ítem 21	0.21	El ítem discrimina poco
Ítem 22	-0.01	El ítem carece de utilidad para discriminar
Ítem 23	0.15	Ítem límite. Se debe mejorar
Ítem 24	0.21	El ítem discrimina poco
Ítem 25	0.06	El ítem carece de utilidad para discriminar
Ítem 26	-0.19	El ítem carece de utilidad para discriminar

Fuente: Elaboración propia según los resultados obtenidos de la prueba de diagnóstico años 2016, 2017 y 2019

3. Análisis de distractores:

En la Tabla 3 se describe el análisis de distractores de cada ítem. Entre los resultados se puede identificar que los ítems 1,6,9,10,15,16 y 19 cumplen con que todos sus distractores sean atractivos y que la respuesta correcta fue electa por la mayoría de los estudiantes. Mientras que los ítems 2,7,8, 12 y 17 contienen un distractor que es poco atractivo y, también hay un distractor más electo que la respuesta correcta. Los ítems 4,18 y 22 contienen dos distractores que resultan poco atractivos para los estudiantes. Por otra parte, los ítems 21,23,24 demuestran que todos sus distractores son más atractivos que la respuesta correcta.

Tabla 8: Análisis de distractores de cada ítem

Ítem	**Porcentaje de estudiantes que seleccionaron la respuesta correcta**	**Distractor más atractivo que la respuesta**	**Porcentaje de distractor más atractivo que la respuesta**	**Distractor poco atractivo (menos del 5%)**	**Observación**
Ítem 1	55%	No	---	No	Todos los distractores son atractivos y la respuesta correcta es electa por la mayoría de los estudiantes.
Ítem 2	17%	Sí	70%	Sí (1)	Un distractor es poco atractivo y hay un distractor más electo que la respuesta correcta.
Ítem 3	19%	Sí	40%	No	Todos los distractores son atractivos y hay un distractor más electo que la respuesta correcta.
Ítem 4	64%	No	---	Sí (2)	Dos distractores son poco atractivos y la respuesta correcta es electa por la mayoría de los estudiantes.
Ítem 5	54%	No	---	Sí (1)	Un distractor es poco atractivo y la respuesta correcta es electa por la mayoría de los estudiantes.
Ítem 6	55%	No	---	No	Todos los distractores son atractivos y la respuesta correcta es electa por la mayoría de los estudiantes.

Ítem	Porcentaje de estudiantes que seleccionaron la respuesta correcta	Distractor más atractivo que la respuesta	Porcentaje de distractor más atractivo que la respuesta	Distractor poco atractivo (menos del 5%)	Observación
Ítem 7	24%	Sí	63%	Sí (1)	Un distractor es poco atractivo y hay un distractor más electo que la respuesta correcta.
Ítem 8	21%	Sí	68%	Sí (1)	Un distractor es poco atractivo y hay un distractor más electo que la respuesta correcta.
Ítem 9	67%	No	---	No	Todos los distractores son atractivos y la respuesta correcta es electa por la mayoría de los estudiantes.
Ítem 10	55%	No	---	No	Todos los distractores son atractivos y la respuesta correcta es electa por la mayoría de los estudiantes.
Ítem 11	72%	No	---	Sí (1)	Un distractor es poco atractivo y la respuesta correcta es electa por la mayoría de los estudiantes.
Ítem 12	25%	Sí	38%	Sí (1)	Un distractor es poco atractivo y hay un distractor más electo que la respuesta correcta.
Ítem 13	56%	No	---	Sí (1)	Un distractor es poco atractivo y la respuesta correcta es electa por la mayoría de los estudiantes.
Ítem 14	31%	Sí	46%	No	Todos los distractores son atractivos y hay un distractor más electo que la respuesta correcta.
Ítem 15	43%	No	---	No	Todos los distractores son atractivos y la respuesta correcta es electa por la mayoría de los estudiantes.
Ítem 16	52%	No	---	No	Todos los distractores son atractivos y la respuesta correcta es electa por la mayoría de los estudiantes.

Ítem	Porcentaje de estudiantes que seleccionaron la respuesta correcta	Distractor más atractivo que la respuesta	Porcentaje de distractor más atractivo que la respuesta	Distractor poco atractivo (menos del 5%)	Observación
Ítem 17	33%	Sí	55%	Sí (1)	Un distractor es poco atractivo y hay un distractor más electo que la respuesta correcta.
Ítem 18	80%	No	---	Sí (2)	Dos distractores son poco atractivos y la respuesta correcta es electa por la mayoría de los estudiantes.
Ítem 19	63%	No	---	No	Todos los distractores son atractivos y la respuesta correcta es electa por la mayoría de los estudiantes.
Ítem 20	59%	No	---	Sí (1)	Un distractor es poco atractivo y la respuesta correcta es electa por la mayoría de los estudiantes.
Ítem 21	17%	Sí	30%, 27%,23%	No	Todos los distractores son atractivos que la respuesta correcta.
Ítem 22	73%	No	---	Sí (2)	Dos distractores son poco atractivos y la respuesta correcta es electa por la mayoría de los estudiantes.
Ítem 23	10%	Sí	9.7%,20%,57%	No	Todos los distractores son atractivos que la respuesta correcta.
Ítem 24	7%	Sí	57%,11%,17%	No	Todos los distractores son atractivos que la respuesta correcta.
Ítem 25	34%	No	---	Sí (1)	Un distractor es poco atractivo y la respuesta correcta es electa por la mayoría de los estudiantes.
Ítem 26	63%	No	---	Sí (1)	Un distractor es poco atractivo y la respuesta correcta es electa por la mayoría de los estudiantes.

Fuente: Fuente: Elaboración propia según los resultados obtenidos de la prueba de diagnóstico años 2016, 2017 y 2019

4. Coeficiente de confiabilidad:

El cálculo del coeficiente de confiabilidad se realizó por medio del programa Jmetrik; que además estableció un intervalo con el 95% de confianza del valor. El coeficiente de Alpha de Cronbach mide la consistencia interna, se identifica que el valor 0.6965 el cual se aproxima a 0.70 y según la clasificación de Cortina (1993) en (Tarqui, 2017) se considera en un nivel aceptable.

Tabla 9: Análisis de confiablidad

Método	**Estimador**	**Intervalo de confianza al 95%**
Coefficient Alpha	0.6965	(0.6601 , 0.7119)

Fuente: Elaboración propia según los resultados obtenidos de la prueba de diagnóstico años 2016, 2017 y 2019

De igual manera en la Tabla 5 se calculó cada coeficiente de confiablidad borrando el ítem, para verificar si la confiablidad de la prueba aumentaba al ser eliminado el ítem. Entre los resultados se puede observar que al eliminar los ítems 9,19, 22 y el 26 la confiablidad de la prueba aumenta, aunque no de manera significativa.

Tabla 10: Confiablidad de la prueba borrando el ítem

Ítem	**Alpha**
p1	0.6648
p2	0.6538
p3	0.6766
p4	0.6660
p5	0.6652
p6	0.6709
p7	0.6502
p8	0.6477
p9	0.7304
p10	0.6827
p11	0.6827
p12	0.6684
p13	0.6670

Ítem	Alpha
p14	0.6541
p15	0.6653
p16	0.6608
p17	0.6473
p18	0.6919
p19	0.7140
p20	0.6881
p21	0.6800
p22	0.6975
p23	0.6836
p24	0.6808
p25	0.6927
p26	0.7143

Fuente: Elaboración propia según los resultados obtenidos de la prueba de diagnóstico años 2016, 2017 y 2019

B. Análisis de ítems Teoría de Respuesta al Ítem

El análisis de la Teoría de Respuesta al ítem fue realizado con el modelo de Rasch que permite describir la habilidad Theta (θ) y conocer el comportamiento de los ítems.

1. Dificultad:

En la Tabla 11 se representa el análisis de los ítems desarrollado en Rasch. Uno de los primeros elementos por observar es el denominado parámetro de la dificultad, la interpretación de este indica que los ítems con parámetro de dificultad negativo son clasificados como ítems fáciles y los ítems positivos como difíciles. Cuantos más cerca se encuentra el parámetro de la dificultad de cero, el ítem tiende a comportarse de una manera modera. Es importante mencionar que el análisis del parámetro de la dificultad expone resultados similares que el análisis del índice de dificultad de TC; descrito en la Tabla 6. Por consiguiente, las observaciones son exactamente las mismas.

2. WMS (Infit) y UMS (Oufit):

WMS (Infit) también es interpretado como medida de “ajuste cercano” propone de manera básica si los estudiantes contestan respecto a cómo la teoría establece que debieran de contestar según el comportamiento del modelo de Rasch. UMS (Oufit) también interpretado como la medida de “ajuste lejano” establece si los ítems miden respecto a cómo la teoría establece que deben medir según el modelo de Rasch. Ambas medidas permiten identificar eventos poco usuales tanto en personas como en ítems. Los valores considerados como aceptables para estas medidas oscilan entre 0.5 y 1.5. En la Tabla No. 11 se observa que el ítem 9 no se ajusta en cuanto a su ajuste de respuestas ni de medida, mientras que el ítem 19 en su ajuste para medir.

Tabla 11: Análisis de dificultad TRI

Item	Difficulty	Std. Error	WMS	Std. WMS	UMS	Std. UMS
p1	-0.60	0.06	0.89	-6.28	0.84	-4.97
p2	1.39	0.08	0.72	-5.89	0.58	-6.96
p3	1.29	0.08	1.00	0.00	0.98	-0.25
p4	-1.00	0.06	0.87	-6.17	0.79	-5.15
p5	-0.55	0.06	0.90	-5.92	0.84	-4.83
p6	-0.60	0.06	0.94	-3.61	0.92	-2.41
p7	0.94	0.07	0.75	-6.83	0.66	-7.21
p8	1.14	0.08	0.69	-7.55	0.58	-8.24
p9	-1.15	0.07	1.40	15.05	2.89	26.29
p10	-0.58	0.06	1.03	1.72	1.04	1.08
p11	-1.39	0.07	0.97	-1.09	1.02	0.40
p12	0.86	0.07	0.91	-2.24	0.92	-1.57
p13	-0.63	0.06	0.90	-5.60	0.86	-4.04
p14	0.54	0.07	0.80	-6.57	0.78	-5.81
p15	-0.08	0.06	0.90	-4.56	0.90	-3.14
p16	-0.44	0.06	0.86	-7.78	0.82	-6.00
p17	0.40	0.07	0.76	-9.05	0.71	-8.44
p18	-1.88	0.07	1.02	0.59	1.14	1.74
p19	-0.94	0.06	1.29	12.94	1.53	10.81
p20	-0.76	0.06	1.07	3.40	1.12	2.96
p21	1.42	0.08	1.05	1.04	0.98	-0.21
p22	-1.45	0.07	1.09	3.11	1.33	5.15
p23	2.15	0.10	1.02	0.30	1.12	1.10
p24	2.53	0.12	0.92	-0.88	1.01	0.10
p25	0.35	0.07	1.18	5.98	1.16	4.05
p26	-0.96	0.06	1.30	13.13	1.54	10.79

Fuente: Elaboración propia según Jmetrik con base a los resultados de la prueba de diagnóstico 2016,2017 y 2019

La Tabla No. 12 presenta un resumen de los puntajes alcanzados, es decir la cantidad de ítems contestados de manera correcta y la respectiva asignación de Theta, que representa

el valor ponderado relativo que tiene un estudiante al contestar una cantidad determinada de ítems. Un valor Theta negativo representa que un estudiante tiene habilidad de contestar correctamente hasta el 50% de los ítems de manera correcta. Mientras que un valor Theta positivo representa que un estudiante tiene una habilidad de contestar más del 50% de los ítems de manera correcta.

Tabla 12: Resumen de puntuaciones y Theta TRI

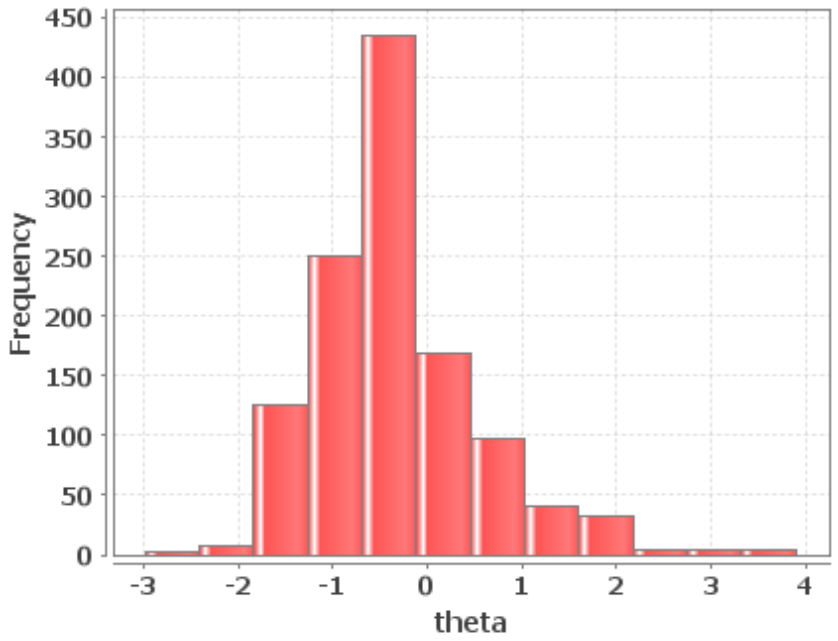
Score	Theta	Std. Err
0.00	-4.96	1.85
1.00	-3.71	1.04
2.00	-2.94	0.76
3.00	-2.46	0.64
4.00	-2.10	0.57
5.00	-1.79	0.53
6.00	-1.53	0.50
7.00	-1.29	0.48
8.00	-1.06	0.47
9.00	-0.85	0.46
10.00	-0.64	0.45
11.00	-0.44	0.45
12.00	-0.24	0.45
13.00	-0.05	0.45
14.00	0.15	0.45
15.00	0.36	0.45
16.00	0.57	0.46
17.00	0.78	0.47
18.00	1.01	0.48
19.00	1.25	0.50
20.00	1.52	0.52
21.00	1.81	0.55
22.00	2.14	0.60
23.00	2.53	0.66
24.00	3.05	0.78
25.00	3.85	1.05
26.00	5.13	1.86

Fuente: Elaboración propia según Jmetrik con base a los resultados de la prueba de diagnóstico 2016,2017 y 2019

La Gráfica 2 muestra un histograma que representa Theta alcanzada por los estudiantes y la frecuencia de estos en los datos reportados para este estudio. La concentración más

grande de estudiantes se ubica aproximadamente entre el rango de -1 a 0. Esto indica que la frecuencia más alta de estudiantes muestra la habilidad de contestar de manera correcta hasta el 50% de los ítems de manera correcta.

Gráfica 2: Frecuencia Theta de los estudiantes



Fuente: Elaboración propia según Jmetrik con base a los resultados de la prueba de diagnóstico 2016,2017 y 2019

En la Tabla 13 se presenta el resumen de la escala estadística propuesto por el análisis de Rasch, en este se observa que, respecto a las personas, en este caso estudiantes el índice de separación es de 1.5 y el número de estratos es de 2.45 lo cual indica que únicamente es posible separar la población en dos grupos. Ya que la diferencia para separar entre uno y otro no es muy grande. El valor de la confiablidad en referencia a las personas se comporta de una manera similar a la confiablidad medida con TCT en este caso es de 0.71. Mientras que la confiablidad de los ítems de 0.99 representa la capacidad que tienen los estudiantes para distinguir entre los ítems.

Tabla 13: Escala Estadística Cualitativa

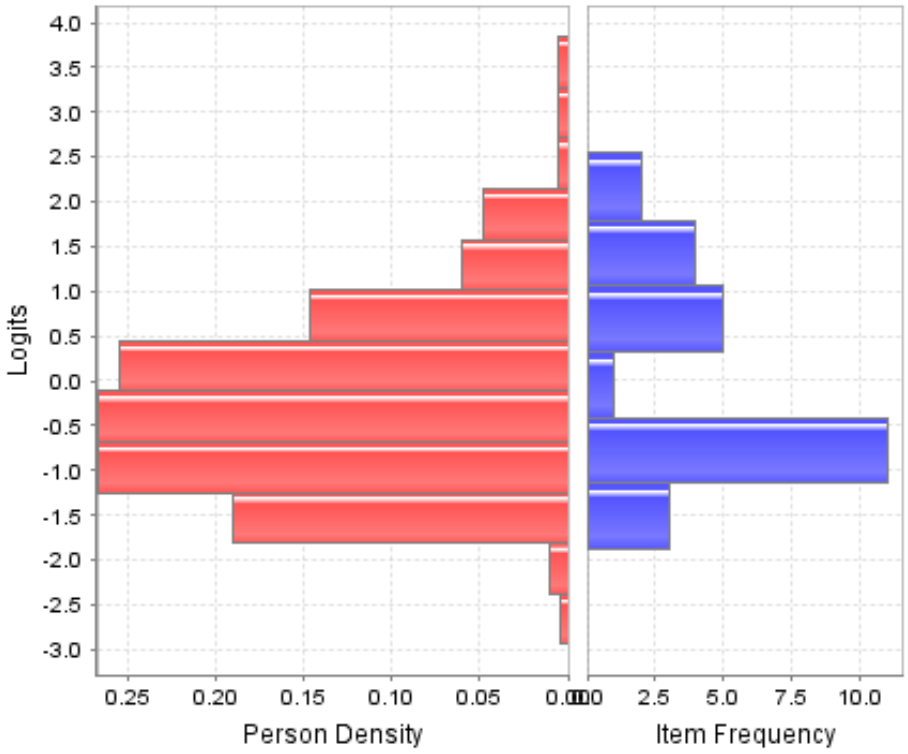
SCALE QUALITY STATISTICS

Statistic	Items	Persons
Observed Variance	1.3227	0.7863
Observed Std. Dev.	1.1501	0.8867
Mean Square Error	0.0053	0.2228
Root MSE	0.0728	0.4720
Adjusted Variance	1.3174	0.5635
Adjusted Std. Dev.	1.1478	0.7507
Separation Index	15.7631	1.5903
Number of Strata	21.3508	2.4537
Reliability	0.9960	0.7166

Fuente: Elaboración propia según Jmetrik con base a los resultados de la prueba de diagnóstico 2016,2017 y 2019

En la Figura 3 se representa de manera gráfica de la frecuencia de estudiantes según Theta (Logits) y la frecuencia de ítems también según Theta, de una manera comparativa. Se puede interpretar que la prueba no cuenta con ítems “muy fáciles” ni con suficientes ítems “muy difíciles”, pero sí hay estudiantes que se ubican en los niveles de habilidades extremas, por lo tanto, no hay manera de medir su habilidad porque no hay ítems correspondientes a esta habilidad. Esto conlleva a interpretar que la puntuación de estos estudiantes no se ajusta al modelo. Si se considera que el punto de corte que limita la habilidad para contestar más de 50% de los ítems se ubica en Theta=0.15, se puede identificar que según los estudiantes hay una densidad relativamente alta, pero no hay suficientes ítems para medir este punto de corte. Lo que conlleva a pensar que se puede ubicar a las personas en una habilidad que no le corresponde.

Gráfica 3: Relación frecuencia estudiantes- frecuencia ítems según Theta



Fuente: Elaboración propia según en Jmetrik con base a los resultados de la prueba de diagnóstico 2016,2017 y 2019

C. Análisis Metodología Bookmark

En relación a la metodología de Bookmark, se trabajó el juicio de expertos con 5 catedráticos del Departamento de Matemática que fueron convocados y que sus horarios lograron coincidir con la actividad. El perfil de los catedráticos participantes en la metodología describe que cuentan con amplia experiencia en el curso de Pensamiento Cuantitativo, 2 de ellas son las actuales coordinadoras del curso; otra fue la anterior directora del Departamento de Matemática y los otros dos catedráticos destacan por su amplia experiencia en el área y trabajan todos los semestres con varios grupos en dicho curso.

Los catedráticos expertos participaron con el fin de establecer los “separadores” que permitan diferenciar niveles de desempeño alcanzado por los alumnos según sus resultados en la prueba de diagnóstico. Para desarrollar el taller se le entregó a cada experto un “Cuadernillo de Bookmark”, que representa los mismos ítems de la prueba ordenados de fácil a difícil y con el índice de dificultad. La primera percepción fue de demostrar desacuerdo en el orden del cuadernillo según el índice de dificultad alcanzado por los estudiantes. Los expertos desarrollaron una amplia discusión que permitió comparar los contenidos e identificaron que había ítems que a su percepción eran más difíciles que otros, pero que un porcentaje significativo de los estudiantes contestaban de manera correcta; y a su vez ítems que consideraban más fáciles pero que un porcentaje bajo de estudiantes habían contestado de manera correcta. Esto llevó a generar la idea de ser innecesaria separar la prueba.

Sin embargo, luego de varias discusiones de la importancia de tratar de interpretar los resultados de los alumnos más que en una puntuación se logró considerar la separación y al momento de realizar la primera ronda la coincidencia de los separadores fue unánime. Todos los catedráticos coincidieron en colocar los puntos de corte en los mismos ítems. Esto se debe a la discusión previa que desarrollaron y que fueron comparando los índices en donde se mencionaba por ejemplo todo se mencionó “que todo sea mas de 0.50 en índice de dificultad debe estar que ser el nivel más bajo”. Y por ende sus propuestas de separadores fueron iguales.

Por esta razón no hubo necesidad de realizar otra ronda, debido a que en la primera luego de la discusión y un acuerdo común se lograron establecer los puntos de corte. Finalmente se hace la propuesta de clasificar a la población en tres grupos, de acuerdo con la Tabla 14: “Debajo de lo esperado” que representa a los estudiantes que contestan correctamente hasta 13 ítems; “Esperado” son estudiantes que contestan correctamente entre 14 y 20 ítems; “Arriba de lo esperado” estudiantes que contestan de 21 a 26 ítems correctos.

En relación con los denominados puntos de corte y si se considera la información de la Tabla No. 12 el primer punto de corte que separa los niveles “Debajo de lo esperado” y “Esperado” hace correspondencia a los estudiantes que demuestran una habilidad $Theta = 0.15$. El segundo punto de corte que separa los niveles “Esperado” y “Arriba de lo esperado” corresponde a los estudiantes con una habilidad $Theta = 1.81$, medido en unidades “Logits” que utiliza el modelo de Rasch como unidad de medida.

Tabla 14: Niveles propuestos de la prueba según modelo Rasch

No. De posición del ítem	Nivel propuesto
1	Debajo de lo esperado
2	
3	
4	
5	
6	
7	
8	
9	
10	
11	
12	
13	
14	Esperado
15	
16	
17	
18	
19	
20	
21	Arriba de lo esperado
22	
23	
24	
25	
26	

Fuente: Elaboración propia según los resultados del juicio de expertos (2019).

De igual manera se trabajó con algunos de los expertos del taller los descriptores del desempeño de esos niveles y de acuerdo con lo observado en los ítems del cuadernillo de Bookmark y a su experiencia en el curso, se logró determinar cada descriptor de desempeño descrito en el Cuadro No. 1.

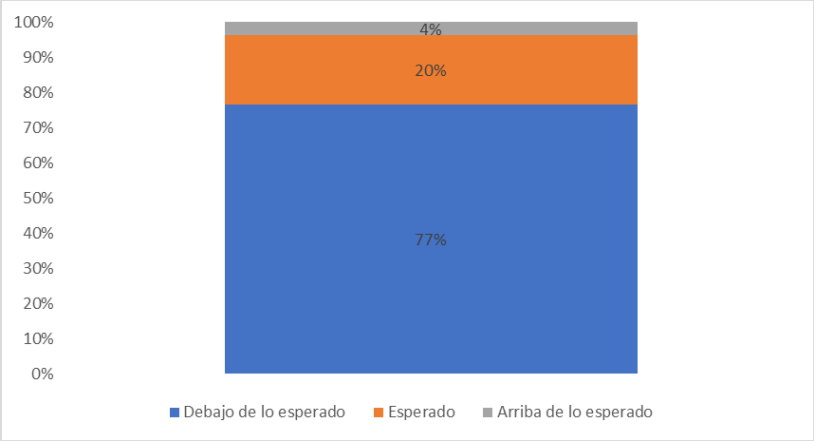
Cuadro 1: Criterios de desempeño

Criterios de desempeño	**Descriptores de desempeño**
Debajo de lo esperado	El estudiante demuestra habilidad operatoria relacionada a simplificación algebraica y aritmética.
Esperado	El estudiante cuenta las habilidades descritas en los niveles anteriores y, además resuelve operaciones complejas.
Arriba de lo esperado	El estudiante demuestra dominio competente en temas de álgebra y aritmética.

Fuente: Elaboración propia según los resultados del juicio de expertos (2019).

Con los resultados de los 1172 estudiantes de primer ingreso considerados para este estudio se establece la distribución porcentual en la Gráfica 4 de estos según el nivel de desempeño alcanzado. En donde es representativo observar que el 77% de los alumnos se ubica en el nivel “Debajo de lo esperado” y únicamente el 4% de los estudiantes de primer ingreso de los años 2016,107 y 2019 se posiciona en el nivel “Arriba de lo esperado”.

Gráfica 4: Distribución según desempeño alcanzado



Fuente: Elaboración propia según los resultados del juicio de expertos (2019).

VIII. CONCLUSIONES

- Según el análisis del índice de dificultad, bajo el modelo de TCT e interpretado con la clasificación de Vallejo (2008) se puede concluir que el ítem 23 y 24 de la prueba de diagnóstico de matemática que utiliza el Departamento de Matemática son categorizados como “muy difíciles”; estos hacen referencia al tema de simplificación de expresiones logarítmicas y exponenciales. El 92% de los ítems de la prueba se encuentran clasificados entre las categorías difícil, moderado y fácil; de una manera equitativa. La prueba no cuenta con ítems denominados “Muy fáciles”. Los contenidos en donde los alumnos demuestran mayor dominio en los ítems son: la simplificación algebraica, aplicación de reglas de los radicales y resolución de desigualdades lineales.

- Con el análisis del índice de discriminación del modelo de TCT, interpretado con la escala propuesta por Ebel (1965) en Muñiz, Fidalgo, García-Cueto, Martínez, & Moreno, (2005) se puede concluir que el 50% de los ítems de la prueba de diagnóstico de matemática del Departamento de Matemática discriminan “bien” y “muy bien”. Los ítems 3,21 y 24 “discriminan poco”; mientras que los ítems 10,11,20,23 se ubican al límite del índice de discriminación y los ítems 9,18,19,22,25,26 “carecen de utilidad para discriminar”.

- Según el análisis de distractores los ítems 1,6,120,15,16,16 demuestran en su análisis de distractores que las opciones son atractivas y que ningún distractor tiene un porcentaje de elección más alto que la respuesta. Los ítems 21, 23 y 24 tienen distractores más atractivos que las respuestas correctas. Los ítems 2,7,8,12,17,4,18,22 tienen por lo menos un distractor que resulta poco atractivo para los estudiantes.

- El cálculo del coeficiente de confiabilidad Alpha de Cronbach es de 0.6965, lo cual según Cortina (1993) en Tarqui (2017) se considera un nivel de confiablidad aceptable de la prueba. También se concluye que al eliminar los ítems 9,19,22 y 26 la confiablidad de la prueba aumenta.

- Según el análisis de Rasch y en referencia a las medidas de WMS (Infit) y UMS (Oufit) el ítem 9 no tiene ajuste de las respuestas al modelo teórico, en este caso Modelo Rasch y el ítem tampoco miden respecto a cómo teóricamente debería de medir.

- Según el análisis de Rasch el número de estratos es de 2.45 y el índice de separación de 1.59 establece que únicamente se puede clasificar a la población en dos grupos, no hay una separación tan grande como para diferenciar entre un nivel y otro.

- Como conclusión general del análisis de Rasch, se establece que la prueba no cuenta con suficientes ítems clasificados como "muy difíciles" ni "muy fáciles", pero si hay estudiantes que ubican su habilidad en los extremos, lo que conlleva a describir que no hay ítems que mida sus habilidades en específico. Para el punto de corte $Theta = 0.15$ que diferencia de los estudiantes que demuestran una habilidad para contestar más del 50% de los ítems de manera correcta, la prueba no cuenta con suficientes ítems de tipo "moderado" que permitan distinguir la habilidad cerca del punto de corte lo que podría ubicar a estudiantes en un nivel que no le corresponde ya que no hay suficientes ítems que determinen de manera precisa la ubicación de su habilidad.

- Por medio de la metodología de Bookmark se establecen los separadores y puntos de corte que delimitan los niveles de desempeño alcanzados por los estudiantes, según el análisis de Rasch el primer punto de corte demuestran una habilidad $Theta = 0.15$ medido en unidades Logits. El segundo punto de corte se encuentra ubicado en $Theta = 1.81$ unidades Logits. Lo cual ubica al setenta y siete por ciento de la población en el nivel "Debajo de lo esperado"; veinte por ciento en "Esperado" y el cuatro por ciento en el nivel "Arriba de lo esperado".

- La Teoría de Respuesta al Ítem además de proporcionar el comportamiento de los ítems en relación a los elementos psicométricos. Relaciona en la misma escala los ítems y las habilidades de las personas de tal manera que se pueda visualizar de forma significativa si la prueba cuenta con ítems que miden todos los niveles de habilidad que las personas

presentan. Los indicios de realizar el análisis psicométrico de la prueba de diagnóstico que utilizó el Departamento de Matemática con TCT y con TRI enriquecen la visión de los resultados de los alumnos y de los ítems. Que permiten hacer modificaciones para futuros procesos y mejoras en las siguientes mediciones.

IX. RECOMENDACIONES

- Se recomienda trabajar con profundidad en la explicación y ejercitación del tema: simplificación de expresiones logarítmicas y exponenciales, debido a que el 90% de los alumnos de primer ingreso no demuestran dominio en esos ítems de la prueba de diagnóstico.

- Se recomienda en general colocar más ítems, del banco de ítems ya existentes creados y piloteados en las pruebas antecedentes. Ubicar específicamente en los niveles "muy fáciles", "moderados" y "muy difíciles" de tal manera que la prueba este diseñada para medir todos los niveles de desempeño incluyendo los extremos que son alumnos que demuestran muy poco dominio y alumnos con un alto dominio. Y de igual manera clasificar a los alumnos que demuestran habilidad cerca del punto de corte.

- Según los resultados del índice de discriminación se recomienda hacer una revisión para mejorar los ítems 10,11,20,23 debido a que se encuentran en el límite del índice de discriminación, lo cual implica que una cantidad representativa de alumnos que tienen puntuaciones altas fallan en estos.

- Según los resultados del índice de discriminación se recomienda cambiar o quitar los ítems 9,18,19,22,25,26; debido que la interpretación de Ebel (1965) en Muñiz, Fidalgo, García-Cueto, Martínez, & Moreno (2005) los clasifica en la categoría "carecen de utilidad para discriminar".

- Se recomienda hacer una revisión de los ítems 21, 23 y 24 ya que en estos tres casos la respuesta correcta obtuvo el menor porcentaje de elección, lo cual implica que todos los distractores son más atractivos que la respuesta correcta.

- Según el análisis de distractores los ítems 2,7,8,12,17,4,18 y 22 contienen por lo menos un distractor que resultado poco atractivo para los estudiantes, por lo que se sugiere hacer una revisión de esos distractores y de esta manera mejorar la calidad del ítem.

- Se recomienda quitar o cambiar los ítems 9,19,22 y 26 para que aumente el coeficiente de confiabilidad Alpha de Cronbach.

- Según el análisis de Rasch y en referencia a las medidas de WMS (Infit) y UMS (Oufit) se recomienda quitar el ítem 9 ya que no tiene ajuste de las respuestas al modelo teórico, en este caso Modelo Rasch y el ítem tampoco miden respecto a cómo teóricamente debería de medir.

- Según el valor de "Números de estratos" determinado en el análisis de Rasch se recomienda clasificar a la población únicamente en dos grupos. Además, el índice de separación de los resultados no permite hacer diferencias tan representativas entre un nivel y otro.

- Se recomienda trabajar futuros análisis psicométricos de los resultados de las pruebas de diagnóstico de los alumnos de primer ingreso con la TCT y la TRI para visualizar sobre una misma escala las habilidades alcanzadas y las habilidades medidas.

- Debido a que la TRI permite trabajar pruebas con menor cantidad de ítems y con ítems de otros tipos que ya no son únicamente de opción múltiple se recomienda diseñar pruebas con estas denominaciones que son muy pertinentes en el área de matemática.

X. REFERENCIAS BIBLIOGRÁFICAS

Aiken, L. R. (2003). *Test Psicologicos y Evaluación* . México: Pearson.

Aliaga Tovar, J. (2011). Psicometría. En J. A. Tovar, *Psicometría* (págs. 86-108). Fondo Editorial UIGV.

Aravena, M., Kimelman, E., Micheli, B., Torrealba, R., & Zúñiga, J. (2006). *Investigación Educativa 1.* Chile: Convenio Interinstitucional.

Argibay, J. C. (2016). Técnicas psicométricas. Cuestiones de validez y confiabilidad. Subjetividad y Procesos Cognitivos. *Revista de UCES*, 15-33. Obtenido de http://dspace.uces.edu.ar:8180/dspace/bitstream/handle/123456789/765/T%C3%A9cnicas_psicom%C3%A9tricas.pdf?sequence=1

Bogoya, D., Barragán, S., Contento, M., & Ocaña, A. (2014). Calibración de instrumentos de evaluación - clasificación en matemáticas en la Universidad Jorge. *Revista Complutense de Educación Vol. 25 Núm. 2*, 501-519.

Bolado, R., Ibáñez, J., & Lantarón, A. (1998). *Página Web CSN (Consejo de Seguridad Nuclear)* . Obtenido de https://www.csn.es/documents/10182/1012054/ODE-04-08+El+juicio+de+expertos

Bombelli, E. C. (2011). *Importancia de la Evaluación Diagnóstica en Asignaturas de Nivel Superior con Conocimiento Preuniversitario.* Buenos Aires: Dialnet. Obtenido de https://dialnet.unirioja.es/descarga/articulo/4125245.pdf

Borsotti, C. (2007). *Temas de la metodología de la investigación en ciencias sociales empíricas.* Madrid: Miño y Dávila.

Carvajal Álzate, D. E., Méndez Sánchez, H., & Torres Angulo, M. B. (2016). *Pagina web Fundación Universitaria Los Libertadores.* Obtenido de https://repository.libertadores.edu.co/handle/11371/620

Cerdas Núñez, D., & Montero Rojas, E. (2017). Uso del modelo de Rasch para la construcción de tablas de especificaciones: Propuesta metodológica aplicada a una prueba de selección universitaria. *Revista Electrónica "Actualidades Investigativas en Educación"*, 1-16.

College Board. (2019). *Página web The College Board.* Obtenido de https://latam.collegeboard.org/paa/que-es-la-paa/

Digeduca. (8 de Mayo de 2018). *Página web Ministerio de Educación-Digeduca.* Obtenido de http://www.mineduc.gob.gt/digeduca/documents/resultados/Trifoliar_Digeduca_Graduandos_2017.pdf

Dirección Dirección General de Evaluación e Investigación Educativa, Ministerio de Educación. (Enero de 2010). *Página web Ministerio de Educación, DIGEDUCA.* Obtenido de http://www.mineduc.gob.gt/DIGEDUCA/documents/informes/pruebaLiberadabookmark_perito2009.pdf

Escobar-Pérez, J., & Cuervo-Martínez, A. (2008). VALIDEZ DE CONTENIDO Y JUICIO DE EXPERTOS: UNA APROXIMACIÓN A SU UTILIZACIÓN. *Revista Avances en Medición , 6*, 27-36. Obtenido de http://www.humanas.unal.edu.co/psicometria/files/7113/8574/5708/Articulo3_Juicio_de_expertos_27-36.pdf

Española, R. A. (2019). *Página web de Real Academia Española*. Obtenido de https://dle.rae.es/medici%C3%B3n

Eyzaguirre, B. (2003). *Exigencias para la construcción de una prueba de selección a la universidad.* Estudios Públicos. Obtenido de https://www.cepchile.cl/cep/site/artic/20160304/asocfile/20160304093107/rev90_beyzaguirre.pdf

Hernández Sampieri, R. (2010). *Metodología de la Investigación.* México: Mc Graw Hill.

Jiménez Alfaro, K., & Montero Rojas, E. (2013). Aplicación del modelo de Rasch, en el análisis psicométrico de una prueba de diagnóstico en matemática. *Revista digital Matemática, Educación e Internet*, 1-24.

Martínez Arias, M. R., Hernández Lloreda, M. V., & Hernández Lloreda, M. J. (2006). *Psicometría.* Madrid: Editorial Alianza.

Muñiz, J., Fidalgo, Á., García-Cueto, E., Martínez, R., & Moreno, R. (2005). *Análisis de los ítems.* Madrid: LA MURALLA S.A.

Ramirez, G., & Barquero , J. A. (2011). Análisis de las pruebas de Diagnóstico de Matemática del Instituto Tecnologico de Costa Rica. *Revista Digital de Matemática, Educación e internet*. Obtenido de http://revistas.tec.ac.cr/index.php/matematica/article/view/1957/1779

Sánchez Restrepo, H., & Espinosa Rodríguez, J. D. (Julio de 2012). *Academia.edu.* Obtenido de https://www.academia.edu/17927874/0_Construcci%C3%B3n_de_%C3%8Dtems_de_opci%C3%B3n_m%C3%BAltiple_para_pruebas_objetivas

Secreteria de Evaluación Educativa, Ministerio de Argentina. (2016). *Aprender 2016 Bookmark Establecimiento de Puntos de Corte.* Obtenido de Ministerio de Educación de Argentina: https://www.argentina.gob.ar/sites/default/files/manual-bookmark-595bd361cf4e7.pdf

Tarqui, A. A. (2017). *EVALUACIÓN DE LA CIUDAD DE PUNO COMO DESTINO TURÍSTICO - PERÚ.* Obtenido de http://www.scielo.org.pe/pdf/comunica/v8n2/a05v8n2.pdf

Universidad Del Valle de Guatemala. (2017). *Página web Universidad Del Valle de Guatemala.* Obtenido de Código de Ética: http://uvg.edu.gt/nosotros/doc/Codigo_de_Etica_GEDV.pdf

Universidad Del Valle de Guatemala. (2019). *Página web Universidad Del Valle de Guatemala*. Obtenido de https://www.uvg.edu.gt/admisiones/preguntas-frecuentes/

Vallejo, P. M. (2008). *Estadística Aplicada a las Ciencias Sociales.* Madrid: Universidad Pontificia Comillas.

Zamora Araya, J. A. (2013). Análisis de la confiabilidad de los resultados de la Prueba de diagnóstico Matemática en la Universidad Nacional de Costa Rica utilizando el modelo de Rasch. *Portal de Revistas Academicas*, 154-164. doi:http://dx.doi.org/10.15517/ap.v29i119.18693

XI. ANEXOS

Tabla 15: Análisis de ítem 1

XII. Ítem --- p1	Option (Score) --- Overall	Difficulty --- 0.5520	Std. Dev. --- 0.4975	Discrimin. --- 0.3708
	a(1.0)	0.5520	0.4975	0.3708
	b(0.0)	0.0503	0.2187	-0.1804
	c(0.0)	0.2730	0.4457	-0.3654
	d(0.0)	0.0742	0.2623	-0.2480
	e(0.0)	0.0179	0.1327	-0.1371
	f(0.0)	0.0316	0.1749	-0.2109

Fuente: Elaboración propia en Jmetrik según los resultados de la prueba de diagnóstico 2016,2017 y 2019

Tabla 16: Análisis de ítem 2

Ítem --- p2	Option (Score) --- Overall	Difficulty --- 0.1724	Std. Dev. --- 0.3778	Discrimin. --- 0.5664
	a(0.0)	0.7022	0.4575	-0.6297
	b(0.0)	0.0333	0.1794	-0.0227
	c(1.0)	0.1724	0.3778	0.5664
	d(0.0)	0.0631	0.2433	-0.0165
	e(0.0)	0.0256	0.1580	-0.0411
	f(0.0)	0.0034	0.0583	0.0114

Fuente: Elaboración propia en Jmetrik según los resultados de la prueba de diagnóstico 2016,2017 y 2019

Tabla 17: Análisis de ítem 3

Ítem --- p3	Option (Score) --- Overall	Difficulty --- 0.1860	Std. Dev. --- 0.3893	Discrimin. --- 0.2514
	a(1.0)	0.1860	0.3893	0.2514
	b(0.0)	0.2841	0.4512	0.1078
	c(0.0)	0.3968	0.4894	-0.4155
	d(0.0)	0.0802	0.2717	-0.2652
	e(0.0)	0.0119	0.1087	-0.0134
	f(0.0)	0.0410	0.1983	-0.1943

Fuente: Elaboración propia en Jmetrik según los resultados de la prueba de diagnóstico 2016,2017 y 2019

Tabla 18: Análisis de ítem 4

Ítem	Option (Score)	Difficulty	Std. Dev.	Discrimin.
---	---	---	---	---
p4	Overall	0.6416	0.4797	0.3622
	a(1.0)	0.6416	0.4797	0.3622
	b(0.0)	0.2602	0.4390	-0.4256
	c(0.0)	0.0375	0.1902	-0.1906
	d(0.0)	0.0333	0.1794	-0.1828
	e(0.0)	0.0145	0.1196	-0.1469
	f(0.0)	0.0119	0.1087	-0.1252

Fuente: Elaboración propia en Jmetrik según los resultados de la prueba de diagnóstico 2016,2017 y 2019

Tabla 19: Análisis de ítem 5

Ítem	Option (Score)	Difficulty	Std. Dev.	Discrimin.
---	---	---	---	---
p5	Overall	0.5410	0.4985	0.3662
	a(0.0)	0.2312	0.4218	-0.3052
	b(1.0)	0.5410	0.4985	0.3662
	c(0.0)	0.2116	0.4086	-0.4236
	d(0.0)	0.0034	0.0583	-0.0259
	e(0.0)	0.0026	0.0506	-0.0622
	f(0.0)	0.0094	0.0965	-0.1087

Fuente: Elaboración propia en Jmetrik según los resultados de la prueba de diagnóstico 2016,2017 y 2019

Tabla 20: Análisis de ítem 6

Ítem	Option (Score)	Difficulty	Std. Dev.	Discrimin.
---	---	---	---	---
p6	Overall	0.5512	0.4976	0.3073
	a(0.0)	0.0614	0.2402	-0.2139
	b(0.0)	0.3046	0.4604	-0.4280
	c(1.0)	0.5512	0.4976	0.3073
	d(0.0)	0.0648	0.2464	-0.0707
	e(0.0)	0.0043	0.0652	-0.0481
	f(0.0)	0.0137	0.1161	-0.1654

Fuente: Elaboración propia en Jmetrik según los resultados de la prueba de diagnóstico 2016,2017 y 2019

Tabla 21: Análisis de ítem 7

Ítem	Option (Score)	Difficulty	Std. Dev.	Discrimin.
p7	Overall	0.2372	0.4255	0.5680
	a(0.0)	0.6348	0.4817	-0.6248
	b(0.0)	0.0529	0.2239	0.0029
	c(1.0)	0.2372	0.4255	0.5680
	d(0.0)	0.0435	0.2041	-0.1312
	e(0.0)	0.0282	0.1655	-0.0502
	f(0.0)	0.0034	0.0583	-0.0892

Fuente: Elaboración propia en Jmetrik según los resultados de la prueba de diagnóstico 2016,2017 y 2019

Tabla 22: Análisis de ítem 8

Ítem	Option (Score)	Difficulty	Std. Dev.	Discrimin.
p8	Overall	0.2056	0.4043	0.6175
	a(1.0)	0.2056	0.4043	0.6175
	b(0.0)	0.6843	0.4650	-0.6828
	c(0.0)	0.1007	0.3010	-0.0165
	d(0.0)	0.0043	0.0652	-0.0648
	e(0.0)	0.0026	0.0506	-0.0020
	f(0.0)	0.0026	0.0506	-0.0794

Fuente: Elaboración propia en Jmetrik según los resultados de la prueba de diagnóstico 2016,2017 y 2019

Tabla 23: Análisis de ítem 9

Ítem	Option (Score)	Difficulty	Std. Dev.	Discrimin.
p9	Overall	0.6724	0.4696	-0.4144
	a(0.0)	0.0546	0.2273	-0.0714
	b(0.0)	0.1869	0.3900	0.3764
	c(1.0)	0.6724	0.4696	-0.4144
	d(0.0)	0.0776	0.2677	-0.1806
	e(0.0)	0.0060	0.0771	0.0385
	f(0.0)	0.0026	0.0506	-0.0794

Fuente: Elaboración propia en Jmetrik según los resultados de la prueba de diagnóstico 2016,2017 y 2019

Tabla 24: Análisis de ítem 10

Ítem	Option (Score)	Difficulty	Std. Dev.	Discrimin.
p10	Overall	0.5478	0.4979	0.1811
	a(0.0)	0.1152	0.3194	-0.1888
	b(0.0)	0.0768	0.2664	-0.1677
	c(0.0)	0.1681	0.3741	-0.2075
	d(1.0)	0.5478	0.4979	0.1811
	e(0.0)	0.0410	0.1983	-0.0181
	f(0.0)	0.0503	0.2187	-0.2834

Fuente: Elaboración propia en Jmetrik según los resultados de la prueba de diagnóstico 2016,2017 y 2019

Tabla 25: Análisis de ítem 11

Ítem	Option (Score)	Difficulty	Std. Dev.	Discrimin.
p11	Overall	0.7201	0.4491	0.1732
	a(0.0)	0.1553	0.3623	-0.2842
	b(1.0)	0.7201	0.4491	0.1732
	c(0.0)	0.0572	0.2323	-0.1313
	d(0.0)	0.0478	0.2134	-0.1818
	e(0.0)	0.0128	0.1125	-0.0833
	f(0.0)	0.0068	0.0824	-0.0788

Fuente: Elaboración propia en Jmetrik según los resultados de la prueba de diagnóstico 2016,2017 y 2019

Tabla 26: Análisis de ítem 12

Ítem	Option (Score)	Difficulty	Std. Dev.	Discrimin.
p12	Overall	0.2491	0.4327	0.3474
	a(0.0)	0.3780	0.4851	-0.3319
	b(0.0)	0.2295	0.4207	-0.1825
	c(1.0)	0.2491	0.4327	0.3474
	d(0.0)	0.0444	0.2060	-0.1721
	e(0.0)	0.0538	0.2256	0.0113
	f(0.0)	0.0452	0.2079	-0.2561

Fuente: Elaboración propia en Jmetrik según los resultados de la prueba de diagnóstico 2016,2017 y 2019

Tabla 27: Análisis de ítem 13

Ítem	Option (Score)	Difficulty	Std. Dev.	Discrimin.
p13	Overall	0.5580	0.4968	0.3477
	a(0.0)	0.1382	0.3453	-0.3090
	b(0.0)	0.2005	0.4006	-0.3156
	c(1.0)	0.5580	0.4968	0.3477
	d(0.0)	0.0307	0.1726	-0.0321
	e(0.0)	0.0154	0.1230	-0.1296
	f(0.0)	0.0572	0.2323	-0.2661

Fuente: Elaboración propia en Jmetrik según los resultados de la prueba de diagnóstico 2016,2017 y 2019

Tabla 28: Análisis de ítem 14

Ítem	Option (Score)	Difficulty	Std. Dev.	Discrimin.
p14	Overall	0.3072	0.4615	0.4994
	a(1.0)	0.3072	0.4615	0.4994
	b(0.0)	0.1365	0.3435	-0.1861
	c(0.0)	0.4548	0.4982	-0.4684
	d(0.0)	0.0589	0.2355	-0.1765
	e(0.0)	0.0222	0.1473	-0.0577
	f(0.0)	0.0196	0.1388	-0.1853

Fuente: Elaboración propia en Jmetrik según los resultados de la prueba de diagnóstico 2016,2017 y 2019

Tabla 29: Análisis de ítem 15

Ítem	Option (Score)	Difficulty	Std. Dev.	Discrimin.
p15	Overall	0.4343	0.4959	0.3657
	a(0.0)	0.3328	0.4714	-0.3836
	b(0.0)	0.1015	0.3022	-0.2731
	c(1.0)	0.4343	0.4959	0.3657
	d(0.0)	0.0887	0.2845	-0.0934
	e(0.0)	0.0094	0.0965	-0.1154
	f(0.0)	0.0324	0.1772	-0.2039

Fuente: Elaboración propia en Jmetrik según los resultados de la prueba de diagnóstico 2016,2017 y 2019

Tabla 30: Análisis de ítem 16

Ítem	Option (Score)	Difficulty	Std. Dev.	Discrimin.
p16	Overall	0.5162	0.5000	0.4111
	a(1.0)	0.5162	0.5000	0.4111
	b(0.0)	0.2150	0.4110	-0.2812
	c(0.0)	0.1945	0.3960	-0.4148
	d(0.0)	0.0683	0.2523	-0.2431
	e(0.0)	0.0043	0.0652	-0.0615
	f(0.0)	0.0017	0.0413	-0.0684

Fuente: Elaboración propia en Jmetrik según los resultados de la prueba de diagnóstico 2016,2017 y 2019

Tabla 31: Análisis de ítem 17

Ítem	Option (Score)	Difficulty	Std. Dev.	Discrimin.
p17	Overall	0.3345	0.4720	0.5652
	a(0.0)	0.0256	0.1580	-0.1315
	b(1.0)	0.3345	0.4720	0.5652
	c(0.0)	0.5486	0.4978	-0.5760
	d(0.0)	0.0776	0.2677	-0.2298
	e(0.0)	0.0102	0.1007	-0.0189
	f(0.0)	0.0026	0.0506	-0.0751

Fuente: Elaboración propia en Jmetrik según los resultados de la prueba de diagnóstico 2016,2017 y 2019

Tabla 32: Análisis de ítem 18

Ítem	Option (Score)	Difficulty	Std. Dev.	Discrimin.
p18	Overall	0.8046	0.3967	0.0407
	a(1.0)	0.8046	0.3967	0.0407
	b(0.0)	0.0845	0.2782	-0.2209
	c(0.0)	0.0367	0.1881	-0.0847
	d(0.0)	0.0188	0.1358	-0.0550
	e(0.0)	0.0469	0.2116	-0.0137
	f(0.0)	0.0085	0.0920	-0.1575

Fuente: Elaboración propia en Jmetrik según los resultados de la prueba de diagnóstico 2016,2017 y 2019

Tabla 33: Análisis de ítem 19

Ítem	Option (Score)	Difficulty	Std. Dev.	Discrimin.
p19	Overall	0.6288	0.4833	-0.1894
	a(0.0)	0.1408	0.3479	0.0834
	b(1.0)	0.6288	0.4833	-0.1894
	c(0.0)	0.1365	0.3435	-0.0521
	d(0.0)	0.0640	0.2448	-0.1025
	e(0.0)	0.0060	0.0771	-0.0490
	f(0.0)	0.0230	0.1501	-0.2068

Fuente: Elaboración propia en Jmetrik según los resultados de la prueba de diagnóstico 2016,2017 y 2019

Tabla 34: Análisis de ítem 20

Ítem	Option (Score)	Difficulty	Std. Dev.	Discrimin.
p20	Overall	0.5879	0.4924	0.1198
	a(1.0)	0.5879	0.4924	0.1198
	b(0.0)	0.1451	0.3523	-0.1845
	c(0.0)	0.1015	0.3022	-0.1887
	d(0.0)	0.0469	0.2116	-0.0363
	e(0.0)	0.0503	0.2187	-0.0287
	f(0.0)	0.0666	0.2494	-0.2970

Fuente: Elaboración propia en Jmetrik según los resultados de la prueba de diagnóstico 2016,2017 y 2019

Tabla 35: Análisis de ítem 21

Ítem	Option (Score)	Difficulty	Std. Dev.	Discrimin.
p21	Overall	0.1681	0.3741	0.2051
	a(0.0)	0.2961	0.4567	-0.2096
	b(0.0)	0.2705	0.4444	-0.2085
	c(1.0)	0.1681	0.3741	0.2051
	d(0.0)	0.2312	0.4218	-0.1505
	e(0.0)	0.0256	0.1580	0.0223
	f(0.0)	0.0077	0.0873	-0.1623

Fuente: Elaboración propia en Jmetrik según los resultados de la prueba de diagnóstico 2016,2017 y 2019

Tabla 36: Análisis de ítem 22

Ítem	Option (Score)	Difficulty	Std. Dev.	Discrimin.
p22	Overall	0.7321	0.4431	-0.0127
	a(0.0)	0.1613	0.3679	-0.1626
	b(0.0)	0.0333	0.1794	-0.0904
	c(0.0)	0.0333	0.1794	-0.0252
	d(1.0)	0.7321	0.4431	-0.0127
	e(0.0)	0.0290	0.1679	-0.0306
	f(0.0)	0.0111	0.1048	-0.1804

Fuente: Elaboración propia en Jmetrik según los resultados de la prueba de diagnóstico 2016,2017 y 2019

Tabla 37: Análisis de ítem 23

Ítem	Option (Score)	Difficulty	Std. Dev.	Discrimin.
p23	Overall	0.0956	0.2941	0.1487
	a(0.0)	0.0973	0.2965	0.0528
	b(0.0)	0.2039	0.4031	0.1016
	c(0.0)	0.5742	0.4947	-0.4556
	d(1.0)	0.0956	0.2941	0.1487
	e(0.0)	0.0145	0.1196	0.0145
	f(0.0)	0.0145	0.1196	-0.1614

Fuente: Elaboración propia en Jmetrik según los resultados de la prueba de diagnóstico 2016,2017 y 2019

Tabla 38: Análisis de ítem 24

Ítem	Option (Score)	Difficulty	Std. Dev.	Discrimin.
p24	Overall	0.0691	0.2538	0.2119
	a(0.0)	0.5887	0.4923	-0.3813
	b(0.0)	0.1084	0.3110	-0.1074
	c(0.0)	0.1715	0.3771	0.0685
	d(1.0)	0.0691	0.2538	0.2119
	e(0.0)	0.0495	0.2170	0.0722
	f(0.0)	0.0128	0.1125	-0.1718

Fuente: Elaboración propia en Jmetrik según los resultados de la prueba de diagnóstico 2016,2017 y 2019

Tabla 39: Análisis de ítem 25

Ítem	Option (Score)	Difficulty	Std. Dev.	Discrimin.
p25	Overall	0.3439	0.4752	0.0613
	a(1.0)	0.3439	0.4752	0.0613
	b(0.0)	0.2935	0.4556	-0.2231
	c(0.0)	0.2312	0.4218	-0.2504
	d(0.0)	0.0367	0.1881	0.0367
	e(0.0)	0.0785	0.2691	0.0458
	f(0.0)	0.0162	0.1263	-0.1868

Fuente: Elaboración propia en Jmetrik según los resultados de la prueba de diagnóstico 2016,2017 y 2019

Tabla 40: Análisis de ítem 26

Ítem	Option (Score)	Difficulty	Std. Dev.	Discrimin.
p26	Overall	0.6314	0.4826	-0.1935
	a(0.0)	0.1988	0.3993	0.0505
	b(0.0)	0.0213	0.1445	-0.0100
	c(0.0)	0.0973	0.2965	-0.1720
	d(1.0)	0.6314	0.4826	-0.1935
	e(0.0)	0.0324	0.1772	0.0625
	f(0.0)	0.0188	0.1358	-0.1728

Fuente: Elaboración propia en Jmetrik según los resultados de la prueba de diagnóstico 2016,2017 y 2019