

Validación de un instrumento mediante técnicas psicométricas modernas para la medición del razonamiento cuantitativo

Validation of an instrument using modern psychometric techniques for measuring quantitative reasoning

Juan Carlos Acosta Jiménez¹, Ingrid Selene Torres Rojas²

RESUMEN

Introducción: Las competencias en razonamiento cuantitativo son fundamentales en la educación superior, especialmente porque los resultados de pruebas estandarizadas, como Saber TyT, reflejan deficiencias en su desarrollo. Por lo tanto, es necesario contar con instrumentos válidos y confiables que permitan evaluar este proceso y facilitar la implementación de estrategias de mejora. **Objetivo:** Validar un instrumento de medición para evaluar el razonamiento cuantitativo en estudiantes de

tecnología en Colombia. **Métodos:** Se llevó a cabo un estudio cuantitativo que incluyó un análisis factorial exploratorio (AFE) y, posteriormente, un análisis factorial confirmatorio (AFC). La muestra estuvo compuesta por 165 participantes y se utilizaron modelos de ecuaciones estructurales (SEM) junto con una estimación DWLS para datos categóricos. Se diseñó un instrumento que contiene 35 reactivos alineados con las dimensiones del ICFES (interpretación, formulación y argumentación). Además, se realizó un pilotaje con 53 estudiantes. La validez de contenido fue evaluada mediante el juicio de expertos y el coeficiente V de Aiken, mientras que la confiabilidad se analizó con el coeficiente KR-20. **Resultados:** El instrumento mostró una confiabilidad de KR-20 = 0,846. La validez de contenido presentó coeficientes V de Aiken superiores a 0,90 en claridad, pertinencia y neutralidad. El AFE identificó tres factores que se alinean con el modelo teórico, y el AFC confirmó la estructura factorial propuesta, obteniendo índices de ajuste adecuados (CFI = 1,000, TLI = 1,008, RMSEA = 0,000, SRMR = 0,090), lo que indica un modelo bien ajustado. **Conclusiones:** Se logró desarrollar un instrumento con alta confiabilidad y validez para evaluar el razonamiento cuantitativo en la educación superior, garantizando su estabilidad estructural a través de la validación confirmatoria.

DOI: <https://doi.org/10.47307/GMC.2025.133.s1.23>

ORCID: 0000-0002-8783-2111¹

ORCID: 0000-0002-2939-1071^{2*}

¹Fundación Universitaria Tecnológico Comfenalco, Colombia.
E-mail: jacosta@tecnologicocomfenalco.edu.co

²Corporación Universitaria Autónoma del Cauca, Colombia.
E-mail: ingrid.torres.r@uniautonomo.edu.co, iselenetr@gmail.com

*Corresponding author: Ingrid Selene Torres-Rojas, ²Corporación Universitaria Autónoma del Cauca, Colombia. E-mail: ingrid.torres.r@uniautonomo.edu.co, iselenetr@gmail.com

Recibido: 13 de enero 2025

Aceptado: 18 de marzo 2025

Palabras clave: Validación psicométrica, razonamiento cuantitativo, análisis factorial confirmatorio (AFC), confiabilidad, validez de constructo.

SUMMARY

Introduction: *Competencies in quantitative reasoning are fundamental in higher education, especially because the results of standardized tests, such as Saber T&T, reflect deficiencies in their development. Therefore, it is necessary to have valid and reliable instruments to evaluate this process and facilitate the implementation of improvement strategies.*

Objective: *To validate a measurement instrument to assess quantitative reasoning in technology students in Colombia.* **Methods:** *A quantitative study was conducted, which included an exploratory factor analysis (EFA) and, subsequently, a confirmatory factor analysis (CFA). The sample consisted of 165 participants, and structural equation modeling (SEM) was employed in conjunction with the DWLS estimation method for categorical data. An instrument containing 35 items aligned with the ICFES dimensions (interpretation, formulation, and argumentation) was designed. In addition, a pilot test was conducted with 53 students. Content validity was evaluated using expert judgment and Aiken's V coefficient, while reliability was assessed with the KR-20 coefficient.* **Results:** *The instrument showed a reliability of KR-20 = 0.846. The content validity presented Aiken V coefficients above 0.90 in clarity, relevance, and neutrality. The AFE identified three factors that align with the theoretical model, and the AFC confirmed the proposed factor structure, obtaining adequate fit indices (CFI = 1.000, TLI = 1.008, RMSEA = 0.000, SRMR = 0.090), indicating a well-adjusted model.* **Conclusions:** *It was possible to develop an instrument with high reliability and validity for assessing quantitative reasoning in higher education, guaranteeing its structural stability through confirmatory validation.*

Keywords: *Psychometric validation, quantitative reasoning, confirmatory factor analysis (CFA), reliability, construct validity.*

INTRODUCCIÓN

En la actualidad, el desarrollo de estrategias pedagógicas innovadoras es crucial para fortalecer la comprensión de diversas áreas del conocimiento, especialmente en matemáticas, donde se han identificado deficiencias persistentes. La pandemia del COVID-19 acentuó estos desafíos, alterando los procesos de enseñanza-aprendizaje, resaltando la necesidad de integrar Tecnologías de la Información y Comunicación (TIC) para mitigar las limitaciones de la educación remota (1).

Uno de los aspectos críticos en la educación superior en Colombia es el rendimiento en razonamiento cuantitativo, una competencia evaluada en pruebas estandarizadas como Saber TyT (Tómate tu tiempo) y Saber Pro. Según el Instituto Colombiano para la Evaluación de la Educación Superior, esta habilidad permite a los estudiantes desenvolverse eficazmente en contextos sociales, políticos, económicos, educativos y laborales (2). Sin embargo, el bajo rendimiento en esta competencia puede generar retrasos en la formación académica, afectar la aprobación de asignaturas matemáticas y, en algunos casos, contribuir a la deserción universitaria (3).

El razonamiento cuantitativo se define como la capacidad para resolver problemas matemáticos mediante la interpretación de información numérica, la formulación de estrategias y la argumentación lógica (4,5). Además, su dominio se considera un indicador clave del rendimiento en disciplinas STEM (Ciencia, Tecnología, Ingeniería y Matemáticas) y en la preparación para desafíos académicos avanzados (6).

El desempeño en esta competencia se evalúa en tres dimensiones: interpretación y representación, formulación y ejecución, y argumentación (7). Su importancia también radica en que los resultados obtenidos en estas pruebas impactan los procesos de acreditación institucional y la calidad educativa en el país (8). Además, el razonamiento cuantitativo no solo mejora la comprensión de problemas matemáticos, sino que también potencia la capacidad de tomar decisiones informadas en la vida cotidiana (9).

A pesar de la relevancia de esta competencia, no existen instrumentos validados en el contexto colombiano para su medición precisa. La falta de una prueba estandarizada con propiedades psicométricas robustas dificulta la obtención de datos confiables y comparables en el tiempo, lo que limita el diseño de estrategias pedagógicas eficaces (10). Además, los informes del ICFES indican una disminución en los puntajes de razonamiento cuantitativo en los últimos años (11), lo que subraya la necesidad de desarrollar y validar un instrumento específico para su evaluación. Sin embargo, la ausencia de un instrumento validado también afecta la detección temprana de deficiencias en razonamiento cuantitativo,

dificultando la implementación de estrategias pedagógicas adecuadas. Sin herramientas confiables, las decisiones curriculares pueden basarse en datos imprecisos, lo que impacta la calidad del aprendizaje y el desempeño académico de los estudiantes (12).

Para garantizar la validez y fiabilidad de un instrumento de medición, es fundamental emplear metodologías rigurosas como el análisis factorial exploratorio y confirmatorio, junto con estimaciones de consistencia interna y estabilidad temporal. Estas técnicas permiten evaluar la estructura de la prueba y su capacidad para medir de manera precisa la competencia en diferentes poblaciones académicas. Es decir, la validación de un instrumento es fundamental porque garantiza su confiabilidad y validez en la medición del constructo que se pretende evaluar (13).

Este estudio tiene como propósito validar una prueba que permita evaluar el razonamiento cuantitativo con precisión y confiabilidad en el contexto colombiano. Se espera que este instrumento sea utilizado por instituciones de educación superior, investigadores y docentes para medir el desarrollo de esta competencia en distintos niveles académicos. Además, la validación contribuirá a la generación de datos empíricos robustos, facilitando el diseño de estrategias de mejora en la enseñanza de las matemáticas y el razonamiento cuantitativo.

Dado lo anterior, la carencia de una prueba validada representa una brecha en la evaluación del razonamiento cuantitativo en Colombia. Este estudio busca subsanar dicha deficiencia mediante la aplicación de técnicas psicométricas avanzadas que garanticen la fiabilidad y validez de la prueba propuesta, por lo cual la generación de un instrumento estandarizado no solo mejorará la medición de esta competencia, sino que también fortalecerá los procesos educativos en todos los niveles académicos del país.

MÉTODO

Dentro del paradigma positivista, esta investigación tuvo un enfoque netamente cuantitativo, con un alcance de tipo descriptivo cuyo objetivo es la validación de un instrumento de medición del razonamiento cuantitativo. En cuanto al diseño, este estudio fue de tipo no experimental, puesto que no se manipularon variables, sino que se recopilaban datos en un solo momento para el análisis de su estructura y confiabilidad. Es decir, se enmarcó dentro de un diseño instrumental (14), dado que su propósito principal es desarrollar y evaluar las propiedades psicométricas de una prueba.

Los participantes de este estudio fueron estudiantes universitarios de una facultad de ingeniería en una institución de educación superior en Cartagena, Colombia. Para el pilotaje del instrumento, se seleccionaron estudiantes matriculados en el período 2024-1, mientras que, para su validación, la muestra incluyó estudiantes inscritos en los períodos 2024-2 y 2025-1. Todos los participantes eran mayores de edad y se encontraban cursando programas de nivel tecnológico y universitario.

En el contexto de esta investigación, fue fundamental definir un método de muestreo que permitiera determinar un tamaño de muestra representativo, asegurando así que las conclusiones obtenidas fueran estadísticamente significativas (15). Dado que no existía en la literatura un instrumento previamente validado para la medición del razonamiento cuantitativo, se procedió a la elaboración de un cuestionario piloto. Para su aplicación, se optó por un muestreo aleatorio simple, adecuado para una proporción en una población finita, cuyo tamaño muestral se calculó mediante la siguiente fórmula:

$$n = \frac{NZ_{\alpha/2}^2 p(1-p)}{Z_{\alpha/2}^2 p(1-p) + (N-1)e^2} = \frac{157(1,96)^2 (0,5)(1-0,5)}{(1,96)^2 (0,5)(1-0,5) + (157-1)(0,114)^2} \approx 53$$

En esta fórmula, $Z_{(\alpha/2)}$, representa el punto crítico asociado al nivel de confianza, es la proporción de éxito esperada, corresponde al margen de error o precisión, y denota el tamaño total de la población. En este estudio, el tamaño poblacional fue de $N=157$, correspondiente a la cantidad de estudiantes matriculados en el período 2024-1.

Se estableció un nivel de confianza del 95 %, y debido a la ausencia de estudios previos que sirvieran como referencia, se asumió una proporción de éxito $p=50$ %, lo cual es una práctica común en investigaciones exploratorias. Con estos parámetros, se determinó un margen de error de $e=11,04$ %. Al sustituir estos valores en la fórmula, el cálculo arrojó un tamaño de muestra de 53 estudiantes, los cuales fueron seleccionados para participar en el pilotaje del instrumento.

Procedimiento

La investigación se desarrolló en tres fases principales: el pilotaje inicial en donde se aplicó el instrumento a una muestra de 53 estudiantes con el fin de evaluar la comprensión de los ítems y realizar ajustes preliminares. Luego, se realizó la validación del contenido mediante juicio de expertos, calculándose el coeficiente V de Aiken. Para el constructo se trabajó un análisis factorial exploratorio con la comprobación de sus respectivos supuestos, y para confiabilidad con el cálculo del Coeficiente KR-20. Finalmente, se administró la versión ajustada a una muestra representativa de 166 participantes para determinar la estructura factorial con el fin de verificar la estabilidad del modelo de medición.

Instrumento

En toda investigación se busca la eficiencia de esta, y para esto se requiere el cumplimiento de los objetivos trazados y que el proceso se lleve a cabo con el menor número de dificultades. En dicho proceso es frecuente y necesaria la recopilación de información mediante instrumentos de medición, los cuales deben garantizar que los datos encontrados sean significativos e insesgados.

Para ello, es necesario validar las encuestas, formularios y cuestionarios utilizados; un primer paso es realizar una prueba piloto que permita cerciorar la consistencia de las preguntas y su respectiva confiabilidad (16).

Se construyó un cuestionario inicial con 35 reactivos diseñados con base en los componentes de la competencia de razonamiento cuantitativo evaluada por el ICFES, y se llevó a cabo utilizando un formulario en la plataforma de Google, dado que la institución en estudio opera bajo la suite de esta herramienta. En el contexto de esta investigación, el objetivo principal fue evaluar el nivel de comprensión de la competencia de razonamiento cuantitativo. Para ello, se abordaron tres habilidades fundamentales establecidas en las pruebas Saber TyT y Saber Pro: interpretación de la información, resolución de problemas y argumentación de afirmaciones (10). Con base en lo expuesto, el instrumento empleado se estructuró en cuatro secciones: un apartado sociodemográfico compuesto por 11 preguntas, seguido de tres componentes específicos de la competencia de razonamiento cuantitativo: interpretación y representación con 14 preguntas, formulación y ejecución con 14 preguntas, y argumentación con 7 preguntas.

Análisis de datos

Se utilizó un enfoque psicométrico basado en las siguientes técnicas: Análisis de fiabilidad mediante el cálculo del coeficiente V de Aiken para evaluar la consistencia interna del instrumento. Análisis factorial exploratorio (AFE) en la fase de pilotaje, con el fin de identificar la estructura subyacente de la prueba y Análisis factorial confirmatorio (AFC), utilizando modelado de ecuaciones estructurales, para comprobar la validez del constructo y la estabilidad de los ítems.

En cuanto a la manipulación de los datos, se utilizaron hojas de Microsoft Excel para la consolidación de las bases de datos. Para en análisis del constructo AFE y AFC, se utilizó el software libre de Jamovi. Sin embargo, debido a la estructura dicotómica del instrumento y las dificultades en el análisis de este tipo de variables, se empleó el software libre JASP para el manejo de correlaciones Tetracóricas.

RESULTADOS

Pilotaje de instrumento

Las respuestas proporcionadas por los 53 estudiantes que participaron en el pilotaje fueron recopiladas y organizadas en una base de datos utilizando Microsoft Excel. A partir de este análisis, se destacan algunos de los hallazgos

más relevantes. En la Figura 1, se observa una distribución de datos con sesgo positivo, lo que indica que ciertos estudiantes alcanzaron puntajes significativamente superiores al promedio. La media obtenida fue de 63 puntos sobre un total de 200, lo que equivale a una efectividad del 31,5 % ($63/200=0,315$), un resultado que evidencia un desempeño bajo y plantea preocupaciones sobre el nivel de dominio de la competencia evaluada.

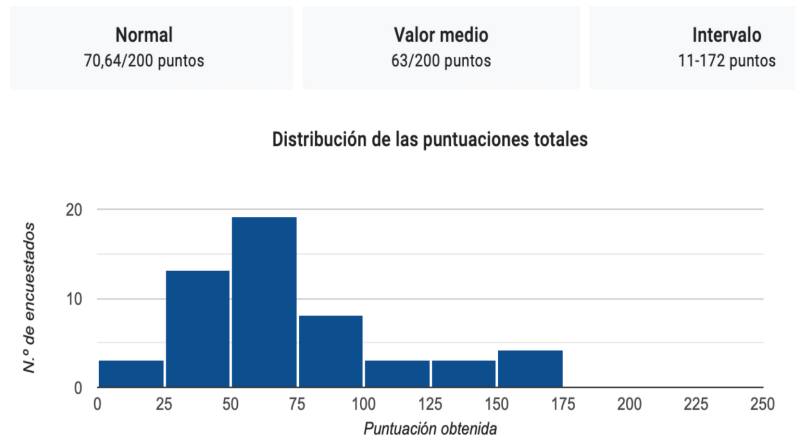


Figura 1. Histograma de los puntajes obtenidos en la prueba diagnóstica.

Nota: Elaboración automática generada en el aplicativo de formularios de Google. La escala de los puntajes es de 0 a 200.

Otro aspecto clave en este pilotaje fue identificar las preguntas con menor porcentaje de aciertos, ya que esto permitió reconocer las áreas temáticas donde los estudiantes presentaron mayores dificultades. Este análisis sugiere que ciertos conceptos fundamentales no han sido completamente asimilados, en particular aquellos relacionados con la resolución de problemas complejos que requieren el manejo de múltiples variables. Asimismo, posibilita evaluar si la redacción de los ítems es clara o si contiene elementos que pueden generar confusión, lo que indicaría la necesidad de mejorar la formulación de las preguntas antes de la aplicación del pretest.

La pregunta con el menor número de respuestas correctas fue la número 15, correspondiente al componente de formulación y ejecución, como se muestra en la Figura 2. Según los resultados obtenidos, solo 7 de los 53 estudiantes (13,2 %) respondieron correctamente, lo que implica que el 86,8 % de los evaluados no lograron resolver adecuadamente este reactivo. Este tipo de

análisis se replicó en cada una de las 35 preguntas evaluadas, lo que permitió establecer las áreas de mayor dificultad y orientar la intervención pedagógica basada en la estrategia propuesta. Una vez recopilada y organizada esta información en una base de datos, se compartió con un grupo de expertos temáticos para su respectiva evaluación mediante el juicio de expertos.

Validez de contenido

La validez de contenido fue un aspecto fundamental en este estudio, ya que garantizó que el instrumento evaluara de manera precisa el concepto teórico para el cual fue diseñado. Este proceso aseguró que las variables observadas representaran de manera adecuada las dimensiones del constructo, proporcionando coherencia y precisión en la medición (17). Para definir el constructo de la investigación, se partió de los conceptos asociados a la variable de interés, estableciendo la relación entre la implementación

VALIDACIÓN DE UN INSTRUMENTO MEDIANTE TÉCNICAS PSICOMÉTRICAS MODERNAS

Pregunta 15 *

Para transportar mango y banano desde un pueblo cercano a dos ciudades, W y Z, un comerciante utiliza tres (3) camiones con capacidad de 5 toneladas cada uno; por cada camión contrata dos trabajadores en cada viaje. El comerciante compra a \$400.000 la tonelada de banano y a \$500.000, la de mango. En la tabla se muestra el precio de venta por tonelada de cada producto y los gastos de transporte y de trabajadores para cada ciudad.

Ciudad	Precio de venta tonelada de banano	Precio de venta tonelada de mango	Costo transporte por camión	Pago por trabajador por viaje
W	\$1.000.000	\$1.300.000	\$150.000	\$180.000
Z	\$1.200.000	\$1.350.000	\$180.000	\$200.000

Tabla

Durante enero, el comerciante vendió 100 toneladas de mango y 50 de banano, y contrató 10 trabajadores. Con esta información es posible conocer

- A. la ganancia de los productores.
- B. el pago que recibirá cada trabajador en enero.
- C. los costos totales del comerciante.
- D. el número mínimo de viajes que se realizaron desde el pueblo.

Pregunta 15

7 de 53 respuestas correctas

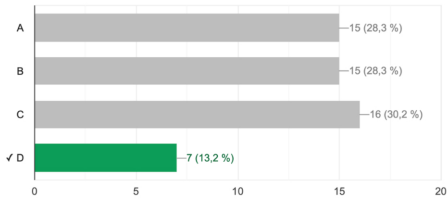


Figura 2. Reactivo con menos aciertos en pilotaje.
Nota: Elaboración automática generada en el aplicativo de formularios de Google.
La pregunta 15 es reactivo que hace parte del listado de preguntas liberadas por el ICFES.

de la metodología del aula invertida y su impacto en el rendimiento académico en la competencia de razonamiento cuantitativo. En consecuencia, el constructo se definió como la “Efectividad del Aula Invertida en el aprendizaje del razonamiento cuantitativo”.

Para validar el contenido, se llevó a cabo un juicio de expertos utilizando el método de agregados individuales. En este procedimiento, el instrumento empleado en el pilotaje fue

enviado a nueve expertos en el área, de los cuales solo siete respondieron oportunamente, tal como se detalla en el Cuadro 1. A cada uno de los expertos se les solicitó evaluar los ítems del instrumento en función de cinco criterios específicos. Posteriormente, las respuestas fueron analizadas mediante el cálculo del V de Aiken, con el propósito de medir el nivel de concordancia entre los expertos y determinar si los ítems eran adecuados y representaban con fidelidad el constructo estudiado (18).

Cuadro 1. Información de expertos temáticos

Experto	Años de experiencia	Posgrado
E_AL	19	Doctor en Ciencias de la Educación
E_JB	17	Doctor en Ciencias mención Matemáticas
E_LV	22	Doctor en Ciencias de la Educación
E_EA	17	Doctor en Ciencias especialidad de Matemáticas
E_AY	20	Magíster en Matemáticas
E_IP	12	Magíster en Estadística Aplicada
E_AT	20	Doctora en Educación Matemática

Nota: Elaboración propia.

En el proceso de validación de los 35 reactivos propuestos, se establecieron cinco criterios fundamentales: Claridad, Pertinencia, Neutralidad, Lenguaje y Validez. Cada uno de estos aspectos fue evaluado mediante dos opciones de respuesta: “Sí” y “No”, asignando un valor de 1 a las respuestas afirmativas y un valor de 0 a las negativas. Esta metodología se fundamentó en la necesidad de calcular el coeficiente V de Aiken, el cual produce valores comprendidos entre 0 y 1, donde un resultado más alto indica una mayor evidencia de validez de contenido. Para determinar este coeficiente, se empleó la fórmula clásica propuesta por Escurra (19) para valores dicotómicos, la cual es completamente igual al compararla con la tradicional razón de validez de contenido (CVR).

$$V = \frac{\Sigma}{n(k-1)}$$

Donde, Σ se define como la suma de los valores “Sí” (valor asignado por cada jurado como 1), n es la cantidad de jurados y k es definido como el número de valores de la escala de valoración. En el Cuadro 2, se presentan los promedios obtenidos tras realizar los cálculos en el software estadístico R. En relación con la claridad en la redacción de los ítems, el coeficiente V de Aiken mostró valores entre 0,67 y 1, evidenciando un alto nivel de concordancia entre los expertos. La mayoría de los ítems alcanzó valores cercanos a 1, lo que indica que fueron percibidos como comprensibles y bien formulados. No obstante, algunos reactivos, como los ítems 1, 17 y 30, registraron valores inferiores (0,67), lo que sugiere la necesidad de realizar ajustes para mejorar su claridad. En términos generales, el promedio obtenido de 0,905 refleja una percepción positiva sobre la claridad del instrumento.

En relación con el segundo criterio, pertinencia, el análisis reflejó un promedio general de 0,952, lo que indicó una percepción favorable sobre la relevancia del instrumento en su conjunto. No obstante, algunos ítems presentaron valores más bajos (0,67), lo que sugirió que ciertos expertos consideraron que estos reactivos podrían no ser tan pertinentes en comparación con los demás.

Cuadro 2. Coeficientes de V de Aiken por criterio

Criterio	V de Aiken
Claridad en la redacción	0,905
Pertinencia	0,952
Neutralidad en la redacción	0,976
Lenguaje adecuado con el nivel del estudiante	0,948
Mide lo que pretende	0,948

Nota: Elaboración propia.

Por su parte, en el tercer criterio, neutralidad en la redacción, los coeficientes V de Aiken se situaron por encima de 0,83, con un promedio de 0,976, evidenciando un alto nivel de consenso entre los expertos sobre la imparcialidad de los ítems. En cuanto al cuarto criterio, evaluó si el lenguaje del instrumento era adecuado para el nivel de los estudiantes, en donde se obtuvo un promedio de 0,948, lo que indicó un fuerte acuerdo entre los evaluadores. Finalmente, el quinto criterio, “Mide lo que pretende”, reflejó nuevamente un alto grado de consenso, con un promedio de 0,948 en el coeficiente V de Aiken, lo que sugirió que los reactivos eran apropiados para evaluar el constructo propuesto y cumplían con los objetivos de medición.

Dentro de las observaciones generales recopiladas, los expertos recomendaron reorganizar los ítems de manera aleatoria para evitar que aquellos pertenecientes a competencias similares estuvieran agrupados. Asimismo, aunque la mayoría de los ítems fueron considerados adecuados para recopilar la información requerida, se identificó la necesidad de revisar algunos de ellos con el fin de mejorar su formulación y optimizar su efectividad dentro del cuestionario. Otra sugerencia clave consistió en adaptar las preguntas para que no fueran idénticas a las utilizadas en los exámenes estandarizados, permitiendo así una mayor originalidad en su formulación.

Posteriormente, se identificaron y consolidaron las preguntas que recibieron un mayor número de observaciones por parte de los expertos. Como resultado de este análisis, se eliminaron ciertos reactivos del instrumento final, reduciendo el total de 35 a 30 ítems, cantidad considerada suficiente para medir las tres dimensiones de la variable dependiente.

Además, se realizaron ajustes en la presentación y formulación de algunos reactivos, atendiendo las recomendaciones proporcionadas por los expertos temáticos. Entre las modificaciones implementadas se incluyeron mejoras visuales, como la incorporación de color en las figuras y la estructuración de las tablas conforme a las normas APA. Estas modificaciones fueron de carácter formal y no conceptual, por lo que se llevaron a cabo en su totalidad sin afectar el contenido de los ítems evaluados.

La distribución final de los reactivos quedó organizada de la siguiente manera: 11 preguntas para el componente de interpretación y representación, 11 preguntas para el componente de formulación y ejecución, y 8 preguntas para el componente de argumentación. Esta distribución mantiene las mismas ponderaciones establecidas en las pruebas Saber TyT y Saber Pro (37 %, 37 % y 26 %, respectivamente), las cuales son aplicadas por el ICFES como requisito obligatorio para la obtención del título a nivel tecnológico o profesional. Con estos ajustes, se consolidó la estructura final del instrumento, asegurando su alineación con los estándares establecidos para la evaluación de la competencia en razonamiento cuantitativo.

Confiabilidad del instrumento

El instrumento diseñado evalúa conocimientos específicos en cada uno de sus reactivos, los cuales siguen una estructura uniforme. Cada ítem inicia con un contexto, seguido de un esquema que puede presentarse en forma de tabla o figura, y finaliza con un enunciado que plantea una interrogante con cuatro opciones de respuesta identificadas como A, B, C o D. Como resultado, los datos obtenidos serán de naturaleza dicotómica, asignando un valor de “1” a las respuestas correctas y “0” a las tres opciones incorrectas en la base de datos correspondiente.

Para evaluar la consistencia interna del instrumento sometido a la prueba piloto, se utilizó el Coeficiente KR-20 (Kuder-Richardson 20). Esta medida permite verificar la fiabilidad del cuestionario, asegurando que las respuestas de los estudiantes sean consistentes. En otras palabras, si el mismo cuestionario fuera aplicado nuevamente bajo condiciones similares, se esperaría obtener resultados equivalentes, garantizando así la estabilidad y precisión del instrumento de evaluación (Ramírez y Badillo, 2024).

Seguido a esto, se calculó dicho coeficiente obteniéndose,

$$KR_{20} = \frac{k}{k-1} \left(1 - \frac{\sum_{i=1}^{52} p_i q_i}{S_i^2} \right) = \frac{35}{35-1} \left(1 - \frac{8,08}{42,9961} \right) \approx 0,836$$

en donde k es la cantidad de reactivos, p_i es la razón de personas que responde correctamente cada ítem. q_i es la razón de personas que no respondieron bien, es definido como la varianza total del puntaje obtenido en la prueba.

Los valores del Coeficiente KR-20 entre 0,7 y 0,9 suelen considerarse adecuados en estudios educativos y psicológicos, ya que reflejan una buena confiabilidad del instrumento. En este caso, el coeficiente obtenido de 0,836 indicó que el 83,6 % de la variabilidad en las respuestas se debía a diferencias reales entre los participantes en relación con el constructo evaluado. En contraste, el 16,4 % restante pudo atribuirse a errores

aleatorios o a factores externos no directamente relacionados con la medición del constructo.

Tras la eliminación de los cinco reactivos identificados en el proceso de validación, el coeficiente KR-20 se ajustó a un nuevo valor de 0,846, lo que representó un indicador sólido de confiabilidad para una prueba de naturaleza dicotómica.

Validez de constructo

El Análisis Factorial Exploratorio (AFE) es una de las metodologías más utilizadas en

la adaptación y validación de instrumentos de medición en el ámbito de las ciencias sociales (21). Por esta razón, la fase de validación concluyó con su implementación. Inicialmente, se verificaron los supuestos del modelo, para lo cual se aplicó el Índice Kaiser-Meyer-Olkin (KMO), que evalúa la adecuación de la muestra, considerando valores superiores a 0,6 como aceptables. Asimismo, se llevó a cabo la prueba de esfericidad de Bartlett, cuya significancia estadística ($p < 0,05$) es un requisito esencial para confirmar que las correlaciones entre los ítems son distintas de cero y que la aplicación del AFE es pertinente (22).

La Figura 3 muestra los resultados de la *Measure of Sampling Adequacy* (MSA), métrica que evalúa la adecuación muestral de cada variable para su inclusión en un Análisis Factorial Exploratorio (AFE). En este caso, el valor global obtenido fue de 0,742, lo que indicó una adecuación satisfactoria para la aplicación del AFE. Además, se analizaron los valores MSA de cada uno de los 30 reactivos del instrumento, identificándose un único valor bajo en la pregunta 6 de la competencia de interpretación (Int.6).

Por otro lado, la prueba de esfericidad de Bartlett arrojó un p-valor altamente significativo ($<0,001$), lo que permitió rechazar la hipótesis nula H_0 (que establece que la matriz de correlación es una matriz identidad). Este resultado confirmó que las correlaciones entre las variables eran significativamente distintas de cero, reforzando la pertinencia de realizar el Análisis Factorial Exploratorio (AFE).

Al calcular los factores a consolidar, se identificaron tres componentes principales, y se procedió a discriminar los reactivos correspondientes a cada uno. En conjunto, estos tres componentes explicaron el 65,7 % de la varianza total de los datos. No obstante, en lugar de utilizar una matriz de correlaciones de Pearson, se empleó la matriz de correlaciones Tetracóricas, ya que es la más adecuada para datos con distribución dicotómica, como los asociados a los reactivos del instrumento (23). Además, el análisis se llevó a cabo utilizando el software libre JASP, dado que programas como SPSS y Jamovi no incluyen esta funcionalidad.

En la Figura 4, se presentan los resultados del cálculo de las cargas factoriales para los

Resumen				Prueba de Esfericidad de Bartlett		
Componente	SC Cargas	% de la Varianza	% Acumulado	χ^2	gl	p
1	7.33	24.4	24.4	1288	435	<.001
2	7.14	23.8	48.3			
3	5.23	17.4	65.7			

Cargas de los Componentes				Medida de Idoneidad del Muestreo KMO	
	Componente			MSA	
	1	2	3	Unicidad	
Form. 4	0.884			0.1913	0.742
Form. 11	0.877			0.2302	0.656
Form. 6	0.865			0.2494	0.847
Form. 8	0.850			0.2688	0.784
Form. 2	0.846			0.2797	0.847
Form. 3	0.839			0.2934	0.760
Form. 9	0.789			0.3514	0.370
Form. 10	0.781			0.3797	0.773
Form. 5	0.749			0.4234	0.886
Form. 1	0.718			0.4786	0.810
Form. 7	0.652			0.5634	0.716
Int. 8		0.972		0.0408	0.595
Int. 9		0.910		0.1629	0.832
Int. 7		0.892		0.1870	0.763
Int. 10		0.843		0.2644	0.755
Int. 3		0.841		0.2587	0.799
Int. 4		0.819		0.2520	0.837
Int. 1		0.805		0.3398	0.678
Int. 5		0.755		0.4022	0.813
Int. 11		0.716		0.4685	0.694
Int. 2		0.667		0.5130	0.695
Int. 6		0.465		0.7417	0.816
Arg. 2			0.856	0.2538	0.566
Arg. 7			0.850	0.2601	0.663
Arg. 5			0.841	0.2868	0.819
Arg. 4			0.794	0.3296	0.605
Arg. 1			0.782	0.3819	0.851
Arg. 3			0.755	0.4207	0.755
Arg. 8			0.693	0.4798	0.757
Arg. 6			0.668	0.5444	0.693

Nota. Se utilizó la rotación 'varimax'

Figura 3. Resultados del Análisis factorial exploratorio (AFE).

Nota: Elaboración propia en el software Jamovi.

componentes principales (PC1, PC2 y PC3), mostrando la asociación de cada reactivo con su respectivo componente. Las cargas factoriales altas, cercanas a 1, indicaron una fuerte relación entre el reactivo y su componente, mientras que las cargas bajas reflejaron una asociación débil. En este sentido, los reactivos “Form. 4”, “Int. 8” y “Arg. 5” registraron cargas significativamente altas en PC1, PC2 y PC3, respectivamente, lo que sugirió que estaban bien representados por esos componentes.

Tras aplicar una rotación Varimax, los autovalores ajustados indicaron que el primer componente explicó el 29,42 % de la varianza, el segundo componente el 28 %, y el tercero el 21,6 %, alcanzando un 79,02 % de varianza acumulada. Este ajuste redistribuyó la varianza de manera más equitativa entre los tres componentes, lo que facilitó la interpretación de la estructura factorial del instrumento (24).

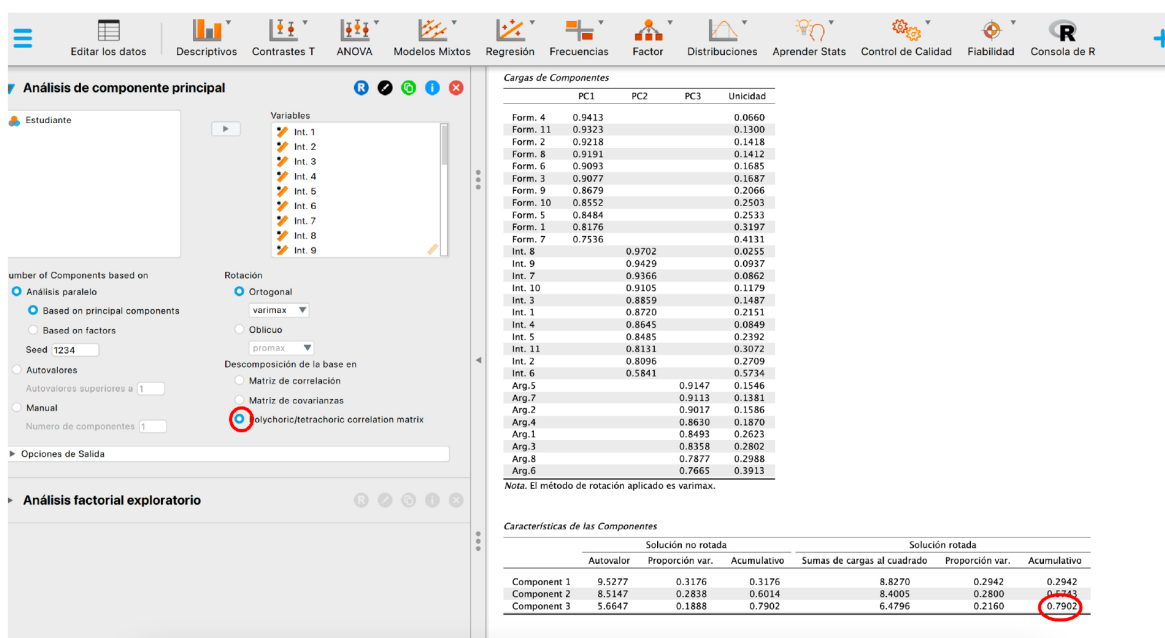


Figura 4. Solución rotada de AFE con matriz Tetracórica.
Nota: Elaboración propia en el software JASP.

El Análisis Factorial Confirmatorio (AFC) es un procedimiento esencial dentro de la validación de constructo, ya que permite comprobar si la estructura factorial identificada en el Análisis Factorial Exploratorio (AFE) se mantiene en una nueva muestra. Es decir, este proceso es crucial porque garantiza que los ítems realmente reflejan el constructo que se pretende medir, asegurando coherencia y precisión en la medición. Para ello, el AFC emplea índices de ajuste como CFI, TLI, RMSEA y SRMR, que permiten evaluar la calidad del modelo y confirmar la adecuación de la estructura factorial propuesta (25). De esta manera, la validación del constructo mediante el AFC garantiza que el instrumento desarrollado tiene una base teórica y estadística sólida, facilitando su aplicación en futuras investigaciones y asegurando que mide correctamente la competencia evaluada. Por lo tanto, la inclusión del AFC en esta etapa permitió no solo validar la estructura factorial de la prueba, sino también proporcionar una herramienta estandarizada y confiable para la medición del razonamiento cuantitativo en contextos educativos.

Por otro lado, el AFC utiliza modelos de Ecuaciones Estructurales (SEM) para evaluar el ajuste del modelo, permitiendo verificar si los reactivos se agrupan correctamente en los factores definidos. Mientras que el Análisis Factorial Exploratorio (AFE) explora la estructura factorial sin imponer un modelo previo, el AFC confirma si la estructura identificada en el AFE es válida en una nueva muestra, en este caso conformada por 165 participantes diferentes a los que participaron en el pilotaje (26).

En la Figura 5 se plasman los resultados obtenidos al iniciar el modelo de AFC, en donde se plantean 3 factores: los dos primeros fueron Interpretación y Formulación con 11 reactivos cada uno y como tercer factor se definió Argumentación con 8 reactivos. La estimación se realizó mediante el método Diagonally Weighted Least Squares (DWLS), crucial en el AFC para datos categóricos, incluidas preguntas dicotómicas, ya que permite obtener estimaciones precisas sin asumir normalidad multivariada. Su importancia radica en que utiliza correlaciones Tetracóricas o Policóricas, en lugar de covarianzas

de Pearson, lo que mejora la estimación de los parámetros cuando las variables son binarias u ordinales (27).

Models Info	
Estimation Method	DWLS
Optimization Method	NLMINB
Number of observations	165
Free parameters	63
Standard errors	Robust
Scaled test	Mean adjusted scaled and shifted
Converged	TRUE
Iterations	53
Model	Interpretation =~ I1 + I2 + I3 + I4 + I5 + I6 + I7 + I8 + I9 + I10 + I11 Formulacion =~ F1 + F2 + F3 + F4 + F5 + F6 + F7 + F8 + F9 + F10 + F11 Argumentation =~ A1 + A2 + A3 + A4 + A5 + A6 + A7 + A8
Nota. Variable (I1,I2,I3,I4,I5,I6,I7,I8,I9,I10,I11,F1,F2,F3,F4,F5,F6,F7,F8,F9,F10,F11,A1,A2,A3,A4,A5,A6,A7,A8) has been coerced to ordered type.	

Model tests			
Label	X²	df	p
User Model	588	402	<.001
Baseline Model	3396	435	<.001
Scaled User	581	402	<.001
Scaled Baseline	1580	435	<.001

Fit indices					
			95% Confidence Intervals		
Type	SRMR	RMSEA	Lower	Upper	RMSEA p
Classical	0.133	0.053	0.044	0.062	0.288
Robust	0.129				
Scaled	0.129	0.052	0.042	0.061	0.351

Figura 5. Modelo inicial de AFC

Nota: Elaboración en Jamovi.

El User Model representa el modelo evaluado, donde la prueba de Chi-Cuadrado ($X^2 = 588$, $df = 402$, $p < 0,001$) resultó significativa, indicando que el modelo difiere significativamente de los datos reales. En cuanto a los índices de ajuste, el SRMR obtuvo valores de 0,129 – 0,133, lo que supera el umbral recomendado de $\leq 0,08$, indicando que el modelo aún presenta margen de mejora. Por otro lado, el RMSEA registró valores entre 0,052 y 0,053, con un intervalo de confianza de 0,042 a 0,062 y p-valores superiores a 0,05, lo que sugiere un ajuste aceptable. Dado que el SRMR es alto, pero el RMSEA está dentro de los valores recomendados, se concluye que el modelo es

poco adecuado. Por ende, se procedió a eliminar reactivos revisando sus cargas factoriales, el criterio usado fue que, si estimación era inferior a 0,3, se eliminaba el reactivo.

El resultado final de la depuración se sintetiza en la Figura 6, en donde luego de seis iteraciones, se logró mejorar los índices del modelo, obteniéndose una estructura factorial de solo 14 reactivos discriminados por I2, I4, I7, I9 y I10 para la competencia de Interpretación. En la competencia de Formulación también quedaron cinco preguntas F4, F5, F6, F8 y F11. Por su parte, en la competencia de Argumentación solo quedaron los cuatro reactivos A2, A3, A4 y A7.

Measurement model							Fit indices			
	Latent	Observed	Estimate	SE	β	z	p	Type	SRMR	RMSEA
Interpretacion	I2	1.000	0.0000		0.376					
	I4	2.020	0.6452		0.759	3.13	0.002			
	I7	2.050	0.6240		0.770	3.28	0.001			
	I9	2.301	0.7003		0.864	3.29	0.001			
	I10	2.255	0.6715		0.847	3.36	<.001			
Formulacion	F4	1.000	0.0000		0.982					
	F5	0.547	0.0993		0.537	5.50	<.001			
	F6	0.755	0.0844		0.741	8.94	<.001			
	F8	0.762	0.0953		0.748	7.99	<.001			
	F11	0.790	0.0800		0.775	9.87	<.001			
Argumentacion	A2	1.000	0.0000		0.784					
	A3	0.939	0.1743		0.736	5.39	<.001			
	A4	0.772	0.1587		0.605	4.86	<.001			
	A7	0.706	0.1528		0.553	4.62	<.001			
							Model tests			
	Label	X ²	df	p						
	User Model	65.4	74	0.752						
	Baseline Model	1374.6	91	<.001						
	Scaled User	98.1	74	0.032						
	Scaled Baseline	797.3	91	<.001						
							User model versus baseline model			
							Model			
	Comparative Fit Index (CFI)						1.000			
	Tucker-Lewis Index (TLI)						1.008			
	Bentler-Bonett Non-normed Fit Index (NNFI)						1.008			
	Relative Noncentrality Index (RNI)						1.007			
	Bentler-Bonett Normed Fit Index (NFI)						0.952			
	Bollen's Relative Fit Index (RFI)						0.941			
	Bollen's Incremental Fit Index (IFI)						1.007			
	Parsimony Normed Fit Index (PNFI)						0.774			

Figura 6. Modelo final de AFC. Nota: Elaboración en Jamovi.

Los valores resaltados en verde representan las cargas de las estimaciones, las cuales todas están por encima de 0,3, en donde la más baja fue de 0,376 asociada la pregunta I2; la más alta fue de 0,982 correspondiente a la pregunta F4. Los otros valores resaltados en morado indican métricas clave para evaluar la calidad del modelo en el Análisis Factorial Confirmatorio (AFC). A saber, el índice de ajuste SRMR de 0,090 es cercano al límite recomendado de $\leq 0,08$, lo que indica un ajuste moderado, mientras que el RMSEA de 0,0001 sugiere un ajuste perfecto, ya que se considera adecuado cuando es $\leq 0,05$. Además, el p-valor del modelo ($p=0,752$) indica que la diferencia entre la matriz observada y la estimada no es significativa ($p>0,05$ es lo ideal), lo que respalda un buen ajuste de dicho modelo.

Finalmente, los valores resaltados con azul representan el *Comparative Fit Index* (CFI) de 1,000 y el *Tucker-Lewis Index* (TLI) de 1,008, los cuales muestran un ajuste excelente del modelo. Es decir, el CFI compara el modelo estimado con

un modelo base sin relaciones entre variables, y su valor ideal es $\geq 0,95$, lo que en este caso muestra que el modelo se ajusta perfectamente a los datos. Por su parte, el TLI evalúa el ajuste penalizando modelos excesivamente complejos, y su valor recomendado también es $\geq 0,95$. Un TLI de 1,008 sugiere que el modelo no solo se ajusta bien, sino que es eficiente en términos de parsimonia, es decir, logra un buen ajuste sin añadir parámetros innecesarios.

Por último, se presenta visualmente en la Figura 7 el modelo de cierre por medio de un SEM, dicho gráfico estructura es obtenida a partir de un Análisis Factorial Confirmatorio (AFC), donde se visualizan las relaciones entre las variables latentes (Interpretación, Formulación y Argumentación) y sus respectivos ítems observados. En este diagrama de senderos, los círculos representan las variables latentes, mientras que los óvalos corresponden a los ítems de la prueba.

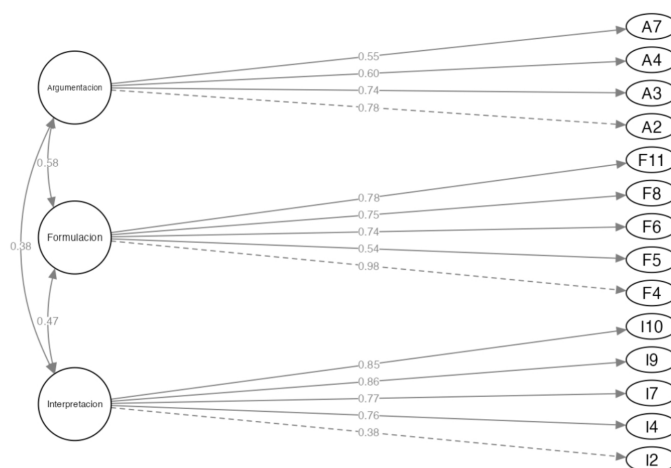


Figura 7. Modelo final de AFC con ecuaciones estructurales. Nota: Elaboración en Jamovi.

Las flechas indican las cargas factoriales, que muestran qué tan fuertemente cada ítem está asociado con su factor subyacente. Valores cercanos a 1,0 representan asociaciones fuertes, mientras que valores más bajos sugieren una relación más débil. En este caso, los ítems I10 e I9 presentan cargas altas con la dimensión de

Interpretación, mientras que F4 y F8 tienen una fuerte relación con Formulación.

Además, las correlaciones entre los factores también son visibles; en particular, los valores 0,38, 0,47 y 0,58 representan las correlaciones entre las variables latentes del modelo, indicando el grado de relación entre los factores Interpretación,

Formulación y Argumentación. La correlación de 0,38 entre Interpretación y Formulación sugiere una relación moderada-baja, lo que significa que, aunque estos dos factores están relacionados, no son completamente dependientes entre sí. En términos simples, un estudiante que tiene un alto desempeño en la competencia de Interpretación no necesariamente tendrá un desempeño igual de alto en Formulación, aunque existe cierta conexión entre ambas habilidades.

Por otro lado, la correlación de 0,47 entre Interpretación y Argumentación indica una relación moderada, lo que sugiere que hay una mayor dependencia entre estas dos competencias. Esto significa que los estudiantes que son buenos interpretando información tienen más probabilidades de desarrollar habilidades de argumentación, aunque sigue habiendo espacio para variaciones individuales. Un ejemplo práctico sería un estudiante que puede leer un gráfico (Interpretación) y, con base en ello, construir una justificación lógica para sus respuestas (Argumentación), aunque no todos los estudiantes que interpretan bien logran argumentar con la misma eficacia.

Además, la correlación de 0,58 entre Formulación y Argumentación indica la relación más fuerte entre los tres factores, lo que sugiere que estas dos competencias están más estrechamente vinculadas. Esto implica que los estudiantes que tienen un buen desempeño en Formulación y Ejecución de problemas matemáticos también tienden a desarrollar argumentaciones más sólidas. Un ejemplo cotidiano sería un estudiante que, al resolver un problema matemático, no solo aplica correctamente las fórmulas necesarias (Formulación), sino que también explica con claridad por qué eligió dicho procedimiento y justifica su respuesta (Argumentación). En conclusión, aunque las tres competencias están relacionadas, algunas conexiones son más fuertes que otras, lo que refuerza la idea de que el razonamiento cuantitativo es un proceso multidimensional.

DISCUSIÓN

El desarrollo de un instrumento psicométricamente válido y confiable para la medición del razonamiento cuantitativo es una necesidad en la

educación superior, dado que esta competencia es fundamental para el desempeño académico y profesional en diversas disciplinas. La validación realizada en este estudio, utilizando análisis factorial exploratorio (AFE) y análisis factorial confirmatorio (AFC), permite garantizar la solidez estructural del instrumento y su aplicabilidad en contextos educativos.

Diversos estudios previos han enfatizado la importancia del diseño de instrumentos que midan de manera precisa competencias específicas en el ámbito académico. En este sentido, la metodología utilizada en este estudio es coherente con investigaciones previas que han aplicado modelos de ecuaciones estructurales (SEM) para validar escalas de medición en diversas áreas del conocimiento. Por ejemplo, Contreras Páez y col. (28), en su estudio en el año 2024 utilizaron AFC para evaluar la validez de un instrumento diseñado para medir ambientes de aprendizaje en educación médica, encontrando resultados similares en cuanto a la solidez del modelo.

De manera análoga, Flores-Cohaila y col. (29), en ese mismo año validaron la estructura factorial de la Escala de Identidad Profesional de Macleod Clark, aplicando análisis psicométricos que confirmaron la multidimensionalidad del constructo. Los resultados obtenidos en este estudio también son consistentes con los hallazgos de Montoro-Pérez y col., quienes validaron el Eating Attitudes Test-26 (EAT-26) en escolares españoles y encontraron que la reducción de ítems en un instrumento, cuando se realiza de manera controlada, no compromete su capacidad de medición (26). En este sentido, la optimización del número de reactivos en este estudio, sin afectar la confiabilidad ($KR-20 = 0,846$), respalda la eficiencia del instrumento y su utilidad práctica en entornos educativos.

En cuanto a la discusión de la validez y confiabilidad, el instrumento desarrollado en esta investigación mostró un alto nivel de validez de contenido, con coeficientes V de Aiken superiores a 0,90, lo que indica que los ítems fueron bien diseñados en términos de claridad, pertinencia y neutralidad. Este hallazgo es congruente con la validación de la escala BARRIERS en el contexto español, donde González-de la Torre y col. (25) demostraron que la aplicación del coeficiente V de Aiken es un método robusto para establecer la validez de constructo en escalas de medición.

El análisis factorial exploratorio (AFE) reveló la existencia de tres factores fundamentales: interpretación, formulación y argumentación, alineados con el modelo teórico subyacente. Esta estructura factorial es coherente con la literatura sobre evaluación del razonamiento cuantitativo en la educación superior. Rojas y col. (2018) han señalado que estas dimensiones son clave para medir el desempeño académico en disciplinas STEM y evaluar competencias matemáticas en estudiantes universitarios (6).

Desde la perspectiva de la fiabilidad, el coeficiente KR-20 obtenido en este estudio (0,846) supera el umbral mínimo recomendado de 0,70, lo que confirma la consistencia interna del instrumento. Este resultado es comparable a lo reportado en la validación del instrumento *Johns Hopkins Learning Environment Scale* (JHLES) en estudiantes de medicina en Colombia, donde se encontró un coeficiente de confiabilidad Cronbach $\alpha = 0,91$, lo que respalda la estabilidad y precisión de los instrumentos validados bajo enfoques psicométricos rigurosos (28).

Por otro lado, con respecto a las implicaciones en la educación superior es menester mencionar que la importancia del razonamiento cuantitativo en la educación superior ha sido ampliamente documentada en la literatura. Su desarrollo influye directamente en la capacidad de los estudiantes para resolver problemas matemáticos, analizar datos y formular argumentos lógicos, competencias fundamentales para disciplinas como la ingeniería, la economía y las ciencias naturales. En este sentido, la validación de un instrumento especializado en la medición de esta competencia representa un avance significativo en la evaluación educativa.

Los resultados obtenidos en este estudio pueden contribuir a la mejora de las estrategias pedagógicas en cursos de matemáticas y estadística. La posibilidad de aplicar un instrumento validado permite a docentes y administradores educativos identificar con mayor precisión las áreas en las que los estudiantes presentan dificultades y diseñar intervenciones didácticas más efectivas. Investigaciones previas, como la de Valencia-Naranjo y col. (30), han demostrado que una evaluación precisa de las habilidades cognitivas en etapas tempranas del aprendizaje permite desarrollar estrategias

personalizadas para mejorar el rendimiento académico.

Además, el impacto del razonamiento cuantitativo en la educación superior no se limita únicamente al ámbito académico, sino que también tiene implicaciones en el mundo laboral. La capacidad para interpretar datos y realizar análisis cuantitativos es cada vez más valorada en el mercado profesional, lo que hace que instrumentos como el desarrollado en este estudio sean de gran utilidad para evaluar la preparación de los estudiantes para enfrentar desafíos del mundo real.

Por último, en cuanto a las limitaciones y futuras líneas de investigación, se puede manifestar que, si bien los resultados de este estudio son prometedores, es importante considerar ciertas limitaciones. En primer lugar, la muestra utilizada en el AFC ($n=165$) es suficiente para los análisis psicométricos, pero futuras investigaciones podrían ampliar el tamaño muestral para mejorar la generalización de los hallazgos. Además, sería pertinente evaluar la estabilidad del instrumento en diferentes contextos educativos y poblaciones con características demográficas diversas.

Otra posible línea de investigación es la comparación de este instrumento con otros modelos de evaluación del razonamiento cuantitativo, a fin de analizar la convergencia y discriminación entre diferentes herramientas de medición. Adicionalmente, se recomienda realizar estudios longitudinales para evaluar la evolución del razonamiento cuantitativo en los estudiantes a lo largo de su formación universitaria y explorar la relación entre el desempeño en esta competencia y otras variables académicas, como la deserción o el rendimiento en asignaturas específicas.

CONCLUSIONES

Este estudio validó un instrumento con alta fiabilidad y validez para medir el razonamiento cuantitativo en educación superior. La estructura factorial confirmada mediante AFC respaldó la adecuación de la prueba, con índices de ajuste que superan los estándares recomendados en la

literatura psicométrica. La reducción del número de ítems optimizó la aplicabilidad del instrumento sin comprometer su capacidad de evaluación, garantizando una medición eficiente y precisa.

Desde una perspectiva metodológica, este estudio aporta evidencia empírica sobre la eficacia del AFE y el AFC, en combinación con la estimación DWLS, para la validación de instrumentos con datos categóricos. La incorporación de correlaciones Tetracóricas permitió mejorar la precisión del modelo, asegurando que los ítems seleccionados reflejen adecuadamente las dimensiones del razonamiento cuantitativo. Estos hallazgos refuerzan la utilidad de los modelos de ecuaciones estructurales (SEM) en la validación de pruebas educativas, lo que representa una contribución metodológica relevante para futuras investigaciones en este ámbito.

En términos prácticos, este instrumento representa una herramienta innovadora para instituciones de educación superior, al facilitar la evaluación de competencias cuantitativas de manera objetiva y estandarizada. Su aplicación permite identificar con precisión las áreas en las que los estudiantes presentan mayores dificultades, lo que posibilita el diseño de estrategias pedagógicas focalizadas en mejorar su desempeño. Además, su estructura alineada con los estándares de evaluación del razonamiento cuantitativo garantiza su aplicabilidad en distintos programas académicos y contextos educativos.

No obstante, futuras investigaciones deberían ampliar la validación del instrumento en diferentes poblaciones y niveles educativos, lo que permitirá evaluar su estabilidad y generalización en diversos entornos. Asimismo, sería pertinente comparar este instrumento con otras pruebas de evaluación del razonamiento cuantitativo, para analizar su convergencia y capacidad discriminativa respecto a otras herramientas existentes. Otra línea de investigación relevante sería la aplicación del instrumento en estudios longitudinales, para examinar cómo evoluciona esta competencia a lo largo de la formación universitaria y su impacto en el rendimiento académico y la permanencia estudiantil.

Finalmente, este estudio enfatiza la importancia de desarrollar instrumentos psicométricamente robustos, que contribuyan a una evaluación

rigurosa del razonamiento cuantitativo, una competencia esencial para el éxito académico y profesional. La validación de esta prueba no solo mejora la medición de esta habilidad en educación superior, sino que también sienta las bases para futuras investigaciones en el campo de la evaluación educativa y el desarrollo de herramientas de medición más especializadas y adaptadas a las necesidades del siglo XXI.

REFERENCIAS

1. Aldazabal Melgar OF, Vertiz-Osores RI, Zorrilla Tarazona E, Aldazabal Melgar LH, Guevara Duarez MF. Software GeoGebra en la mejora de capacidades resolutivas de problemas de figuras geométricas bidimensionales en universitarios. *Propósitos y Represent.* 2021;9(1):1040.
2. ICFES. Guía de orientación del Examen Saber TyT. Bogotá; 2025. Disponible en: https://www.icfes.gov.co/wp-content/uploads/2025/01/02-Diciembre_Guia-de-Orientacion-Modulos-Genericos-Saber-TyT-2025-1.pdf
3. Acosta JC, Baldiris I. Revisión de los métodos de medición del rendimiento académico en los estudiantes universitarios. *Rev Tek.* 2020;1:10-16.
4. Barrera D, Salazar-Torres J, Vera M, Gelvez-Almeida E. Problem-solving strategy in the teaching and learning processes of quantitative reasoning. *J Phys Conf Ser.* 2019;1408(1).
5. Vergara JJ, Fontalvo J, Muñoz A, Valbuena S. Estrategia didáctica para el fortalecimiento del razonamiento cuantitativo mediante el uso de las TIC. *MATUA - Rev Progr Matem.* 2015;2:71-80.
6. Rojas Torres L, Mora Badilla M, Ordóñez Gutierrez G. Asociación del razonamiento cuantitativo con el rendimiento académico en cursos introductorios de matemática de carreras STEM. *Rev Digit Matem Educ Internet.* 2018;19(1):1-13.
7. ICFES. Marco de Referencia para la Evaluación. Módulo de Razonamiento Cuantitativo Saber 11 y Saber Pro. 2015:1-46. Disponible en: <https://www.icfes.gov.co/instituciones-educativas-y-secretarias/saber-pro/documentos>
8. Oviedo AI, Jiménez J. Minería de datos educativos: Análisis del desempeño de estudiantes de ingeniería en las pruebas SABER-PRO. *Rev Politécnica.* 2019;15(29):128-140.
9. Borda M, Latorre G. La gamificación mediada por TICs para fortalecer el razonamiento cuantitativo de los estudiantes de grado tercero de primaria del Colegio Gimnasio Moderno de Neiva María. *Fundación Universitaria Los Libertadores;* 2021.

10. ICFES. Balance de la presentación de las pruebas Saber 3° y 5° 2023. 2023. Disponible en: https://www.icfes.gov.co/wp-content/uploads/2024/11/Matematicas-3_3579_V2.pdf
11. Sáenz-Castro P, Vlachopoulos D, Fàbregues S. Exploring the relationship between saber pro test outcomes and student teacher characteristics in Colombia: Recommendations for improving bachelor's degree education. *Educ Sci*. 2021;11(9).
12. Maldonado-Suárez N, Santoyo-Telles F. Validez de contenido por juicio de expertos: Integración cuantitativa y cualitativa en la construcción de instrumentos de medición. *REIRE Rev d'Innovació i Recer en Educ*. 2024;17(2):1-19.
13. Sánchez-Cabrero R, Sandoval-Mena M, Saez-Suanes GP, López-De-Arana Prado E. Design and validation of the classroom climate for an inclusive education questionnaire (CCIEQ). *Environ Soc Psychol*. 2024;9(5):1-17.
14. Montero I, León OG. A guide for naming research studies in Psychology. *Int J Clin Heal Psychol*. 2007;7(3):847-862.
15. Pérez C. Muestreo estadístico. Prentice Hall; 2005. Disponible en: https://books.google.com.co/books/about/Muestreo_estadístico_conceptos_y_proble.html?id=OP7xAAAACAAJ&redir_esc=y
16. Hernández R, Fernández C, Baptista P. Metodología de la Investigación. 6ª edición. Mexico: Mc Graw Hill; 2014.
17. Jiménez I, Vázquez G, Juárez L, Bracamontes E. Identificación de habilidades socioemocionales y salud mental en profesores de Educación Superior: Validez de constructo. *Rev San Gregorio*. 2023;53:144-166.
18. Merino-Soto C. Coeficientes V de Aiken: diferencias en los juicios de validez de contenido. *MHSalud Rev en Ciencias del Mov Hum Salud*. 2023;20(1):1-10.
19. Ecurra LM. Cuantificación de la validez de contenido por criterio de jueces. *Rev Psicol*. 1988;6(1-2):103-111.
20. Ramirez M, Badillo, Lady. Estrategias didácticas visuales para mejorar el rendimiento académico en la asignatura emprendimiento y gestión. *Rev Sci*. 2024;(9):78-95.
21. Lloret-Segura S, Ferreres-Traver A, Hernández-Baeza A, Tomás-Marco I. El análisis factorial exploratorio de los ítems: Una guía práctica, revisada y actualizada. *An Psicol*. 2014;30(3):1151-1169.
22. Pizarro Romero K, Martínez Mora O. Análisis factorial exploratorio mediante el uso de las medidas de adecuación muestral kmo y esfericidad de Bartlett para determinar factores principales. *J Sci Res*. 2020;5:21.
23. Freiberg Hoffmann A, Stover JB, De la Iglesia G, Fernández Liporace M. Correlaciones Policóricas y Tetracóricas en Estudios Factoriales Exploratorios Y Confirmatorios. *Cienc Psicol*. 2013;21(2):151-164.
24. Ramírez Jiménez AL, Ramírez Guzmán ME, González Estrada E. Validación de la Escala Mexicana de Seguridad Alimentaria a través de Correlación Tetracórica con base en la Distribución Normal Asimétrica Bivariada. *Agrociencia*. 2021;55(1):55-71.
25. González-de la Torre H, Díaz-Pérez D, Rodríguez-Suárez CA, Pinto-Plasencia RJ, Verdú-Soriano J, Cidoncha-Moreno MÁ. Construct validity and reliability of the BARRIERS scale in the Spanish context. *Enferm Clin*. 2024;34(4):259-270.
26. Montoro-Pérez N, Montejano-Lozoya R, Martín-Baena D, Talavera-Ortega M, Rosario Gómez-Romero M. Psychometric properties of the Eating Attitudes Test-26 in Spanish school children. *An Pediatría*. 2024;100(4):241-250.
27. Marsden J, Saunders L, Harman JJ. Pilot study of parental alienation items in the adverse childhood experiences scale. *J Affect Disord*. 2024;367(June):715-744.
28. Contreras Páez S, Domínguez LC, Restrepo JA, Sanabria Á. Evaluating the learning environment in medicine: Validity and reliability of the Johns Hopkins Learning Environment Scale in a population of Colombian students. *Educ Med*. 2024;25(3):100897.
29. Flores-Cohaila JA, Bustamante-Ordoñez MA, Patiño-Villena AF, Pajuelo-Vasquez R, Grau-Monge J, Runzer-Colmenares F. Exploratory and confirmatory factor analysis of the professional identity scale of Macleod Clark in Peruvian medical students. *Educ Med*. 2025;26(1):100971.
30. Valencia-Naranjo N, Calet N, Robles-Bello MA, Gutiérrez-Palma N. On the role of non-linguistic rhythm skills in the early stages of formal learning to read. *Rev Psicodidact*. 2024;29(2):166-175.