

Capítulo 7. Validación de un cuestionario con métodos mixtos: una guía metodológica para investigar el uso de la inteligencia artificial y la tutoría inteligente en la enseñanza de Cálculo en ingeniería



ROLANDO RODRÍGUEZ GONZÁLEZ*

RUBÉN SUÁREZ ESCALONA**

JESÚS ALEJANDRO QUIROZ AGUILAR***

DOI: <https://doi.org/10.52501/cc.472.07>

Resumen

Las tasas de reprobación en Cálculo en la FIME (UANL) oscilaron entre 40% y 60% entre 2020 y 2023, un problema que se agravó con la llegada del plan de estudios 440. Ante este escenario, no basta con proponer soluciones tecnológicas: se necesitan instrumentos rigurosos que evalúen si realmente funcionan. Este capítulo documenta la construcción y validación del cuestionario TAM/UTAUT-Cálculo, una herramienta original —no traducida ni adaptada— que integra cuatro marcos teóricos (TAM, UTAUT, SDT y CLT) para comparar ChatGPT-4 con Khan Academy como apoyos tecnológicos en la enseñanza de Cálculo. Bajo un diseño de métodos mixtos explicativo secuencial, se desarrolló la validación en tres fases: traducción de los marcos teóricos a reactivos, juicio de cinco expertos y aplicación piloto con dos grupos de estudiantes. Los resultados respaldan el instrumento: un cvc de 0.96 y una confiabilidad excelente en ambas versiones ($\alpha = 0.957$ en Khan Academy, $\alpha = 0.938$ en ChatGPT-4), aunque con matices relevantes en Carga cognitiva. Más que cerrar el tema, el capítulo ofrece una hoja de ruta

* Doctorante en Filosofía. Profesor de tiempo completo en la Universidad Autónoma de Nuevo León, México. ORCID: <https://orcid.org/0009-0001-1832-1991>; correo electrónico rrodriguezgz@uanl.edu.mx

** Doctor en Filosofía. Profesor-investigador en la Universidad Autónoma de Nuevo León, México. ORCID: <https://orcid.org/0000-0002-1563-3666>

*** Doctor en Psicología con orientación en Psicología y Educación por la Universidad Autónoma de Nuevo León, México. ORCID: <https://orcid.org/0000-0002-1044-0283>

replicable para quienes deban validar instrumentos con muestras reducidas en sus primeras etapas, dejando el análisis factorial confirmatorio para la fase definitiva.

Palabras clave: *validación de instrumentos, ChatGPT-4, Khan Academy, enseñanza de Cálculo, ingeniería.*

Introducción

Las tasas de reprobación en Cálculo Diferencial e Integral estuvieron entre el 40% y el 60% en la Facultad de Ingeniería Mecánica y Eléctrica (FIME-UANL), durante el periodo comprendido entre 2020 y 2023. Este indicador muestra una dificultad grave. El semestre de agosto a diciembre de 2024 se vuelve más complejo con la implementación del plan de estudios 440. Este plan exige que los alumnos terminen todas las asignaturas del primer ciclo antes de pasar al segundo, lo cual impacta directamente a los estudiantes porque ambos cálculos se enseñan en el primer ciclo.

Una cifra que llama la atención en la FIME es la que reportan Lara Pinales y Neira Rosales (2023): 36.59% de eficiencia terminal a los siete años y medio de estar en el programa de Ingeniería en Administración y Sistemas. Este dato demuestra que el progreso académico del alumno se ve interrumpido si reprueba Cálculo. Por ende, no es suficiente con sugerir soluciones; es imprescindible elaborar instrumentos rigurosos que evalúen si las intervenciones tecnológicas realmente amortiguan esta situación rezagada.

El hecho de que alguien repruebe Cálculo no se debe a una única causa. Artigue (2002) señala que se presentan obstáculos cognitivos específicos al hacer la transición del pensamiento aritmético al algebraico, así como cuando se trata de conceptos abstractos como integral, derivada y límite.

Las plataformas revisadas usan enfoques distintos, lo que afecta la forma en que medimos su efectividad. Mientras que un LMS básico solo clasifica los archivos y Khan Academy ofrece un tutor personalizado con rutas preestablecidas, ChatGPT-4 es una LLM. Esto significa que genera respuestas según patrones probabilísticos; es adaptable, aunque no cuenta con un validador matemático interno (Kasneci et al., 2023).

Esto cambia la manera en que valoramos las capacidades de nuestros alumnos. El ambiente cerrado de Khan Academy estandariza la experiencia y posibilita que se midan los constructos clásicos del TAM/UTAUT. Por otro lado, hablar con ChatGPT depende completamente de la capacidad del estudiante: plantear correctamente las instrucciones y detectar las “alucinaciones” del sistema genera nuevas exigencias cognitivas para el trabajo, lo cual no ocurre en aulas convencionales. Por lo tanto, si el alumno trabaja con un tutor inteligente o con una inteligencia artificial, la puntuación psicométrica de la misma variable puede cambiar considerablemente. Lo que nuestras cifras preliminares del piloto indican es que esto es cierto. Es muy difícil entender cómo un alumno de ingeniería hace uso de una tecnología para adquirir conocimientos sobre cálculo. Una única teoría conceptual no es suficiente para abarcar la abstracción de este tema. Cuatro principios fundamentales constituyen la base de esta investigación.

Para medir la interacción tecnológica, se empleó el Modelo de Aceptación Tecnológica (TAM; Davis, 1989) junto al modelo UTAUT (Venkatesh et al., 2003). Ambos modelos permiten comprobar si la herramienta se percibe como útil, sencilla de emplear y cuánto influye el contexto institucional. Pero la sencillez no es lo único importante. La teoría de la autodeterminación (Deci y Ryan, 2000) permite determinar si el estudiante mantiene su motivación autónoma al enfrentarse a derivadas o integrales. Por último, añadimos la teoría de la carga cognitiva (Sweller et al., 2011). El cálculo requiere mucha atención mental debido a su simbolismo; existe el riesgo real de que la misma interfaz tecnológica acabe absorbiendo al estudiante con exigencias innecesarias.

Con el objetivo de implementar estos cuatro marcos, se desarrolló una herramienta original: el cuestionario TAM/UTAUT-Cálculo. Se trata de una creación nueva, no una traducción ni una copia. Los marcos teóricos (TAM, UTAUT, SDT y CLT) fueron aplicados en un lenguaje adaptado a la realidad del estudiante de ingeniería para la formulación de los reactivos. Porque era una herramienta que no existía hasta ese momento, se tuvo que realizar una evaluación rigurosa. Es evidente que la literatura mexicana carece de recursos. No existen instrumentos validados para medir los efectos cognitivos y motivacionales que causan diferentes tecnologías tales como un tutor inteligente y un

LLM (Modelo Grande de Lenguaje). Por lo tanto, el objetivo principal de la investigación es recopilar los datos del primer estudio piloto y evaluar la consistencia de estas. El procedimiento debe ser transparente en este punto. Se documenta la fase inicial y se difiere el análisis factorial confirmatorio hasta contar con la muestra final ($n \approx 105-120$). Decisión que se explicará más adelante.

Descripción de enfoques y metodologías para el estudio

Enfoque de la investigación, tipo y diseño

La investigación descrita en este capítulo, en el cual se especifica la fase instrumental, se llevó a cabo en la Facultad de Ingeniería Mecánica y Eléctrica de la Universidad Autónoma de Nuevo León, se desarrolló bajo un método mixto explicativo secuencial (Creswell y Plano Clark, 2018), con diseño cuasiexperimental, grupos de control y mediciones pre-post. Para la obtención de resultados, se validó el cuestionario TAM/UTAUT-Cálculo, de lo contrario, la precisión de la evaluación de las intervenciones sería errónea. Dicha validación se implementó mediante tres fases, la primera, traduciendo los marcos teóricos de los ítems que se evaluarán; segunda fase, la validación de expertos para garantizar la validez de contenido (Hernández-Nieto, 2002; Lawshe, 1975); y, por último, la aplicación del instrumento en una prueba piloto con dos grupos de estudiantes, lo que permitió identificar la confiabilidad y ajustar términos para la comprensión adecuada de las variables. Las impresiones cualitativas recogidas en esta fase se analizarán bajo el método temático de Braun y Clarke (2006).

Técnicas y métodos para el estudio

Los datos se obtuvieron mediante la aplicación del cuestionario TAM/UTAUT-Cálculo, instrumento creado específicamente para esta investigación. La escala cuantitativa comprende 25 ítems organizados en 30 reactivos con escala tipo Likert, y en cuanto al método cualitativo, 5 preguntas abiertas.

La parte cuantitativa está organizada en cinco dimensiones basadas en el estado del arte; del modelo TAM, se obtuvo la Utilidad percibida (seis ítems), la Facilidad de uso (cinco ítems). Asimismo, del modelo UTAUT se eligieron los ítems sobre Satisfacción y actitud. En el apartado de Impacto motivacional, se añadieron seis ítems inspirados en la SDT, para finalmente usar del modelo CLT los reactivos sobre Carga cognitiva.

Debido a que el instrumento es nuevo, la validación es indispensable. En primer punto, el contenido se llevó con expertos para calcular el coeficiente de validez de contenido (CVC), mediante la ecuación $CVC_i = \frac{M_x}{V_{\max}}$. Se adoptó un criterio escrito en cuanto a claridad, pertinencia y relevancia: $CVC \geq 0.80$ (Hernández-Nieto, 2002). Respecto a la confiabilidad, se basó en el alfa de Cronbach, tomando en cuenta los ítems con correlación ítem-total (rit) ≥ 0.30 .

Al tratarse de una fase inicial, el análisis se orientó a identificar tendencias descriptivas más que los efectos inferenciales, es por ello que para la parte cualitativa, se analizó con el enfoque temático de Braun y Clarke (2006); en la cuestión estadística, la muestra piloto no permitió asumir normalidad, por lo que se decidió llevar pruebas no paramétricas como la de U de Mann-Whitney y la prueba exacta de Fisher.

De acuerdo con Álvarez-Aguilar et al. (2024), la tecnología inmersa en la formación de los estudiantes universitarios exige no solo innovación, sino también contar con instrumentos linealmente validados estadísticamente, para que los resultados obtenidos garanticen la pertinencia pedagógica. En ese mismo sentido, en este capítulo se exhiben los pasos metodológicos para que las herramientas necesarias evidencien una sólida validación del instrumento de investigación.

Población y muestra

La muestra fue seleccionada por conveniencia e incluyó 120 estudiantes de distintos programas educativos de la dependencia, con un nivel de confianza del 95% para detectar diferencias de efecto mediano ($d \geq 0.50$). Anterior a ello, se realizó una prueba empírica del instrumento, de ahí surgió la muestra actual. Se reunieron dos grupos de estudiantes con experiencia en el uso de plataformas digitales. El equipo de ChatGPT-4 aportó

$n=14$ encuestas válidas. El de Khan Academy inició con 12, pero se eliminó un caso por falta de variabilidad en las respuestas (*straight-lining*), quedando en $n=11$. Las pruebas confirmaron que ambos grupos son demográficamente comparables ($p>0.05$). La única diferencia significativa fue la experiencia previa ($p=0.007$), variable que se controla en la interpretación de los resultados.

Para la validación de expertos, se invitó a cinco especialistas del tema, tres con más de una década de experiencia en tecnología educativa, así como dos con experiencia en impartición de la materia de cálculo, de esta manera se cubren los aspectos teóricos-investigativos y la intervención aplicada.

Análisis de resultados

Análisis del pilotaje y confiabilidad de los instrumentos propuestos

Los datos que se reportan a continuación constituyen productos parciales del proceso de validación instrumental que se integran a la tesis doctoral del primer autor (“Estrategias y herramientas tecnológicas para aumentar el índice de aprobación en las unidades de aprendizaje de Cálculo: caso de estudio en la FIME de la UANL”, doctorado en Filosofía con acentuación en Estudios de la Educación, UANL, Dr. Rubén Suárez Escalona, dir.).

La validez del coeficiente de validez de contenido mostró una solidez adecuada, promediando un 0.96, superando el mínimo establecido de 0.80 determinado por Hernández-Nieto (2002). En las cinco dimensiones teóricas, se obtuvieron valores de 0.95, lo que refleja una precisión sobre la evaluación de los ítems, de acuerdo con lo que pretendía medir. Por parte de los expertos, la pertinencia (0.99), la relevancia (0.97) y la claridad (0.93) alcanzaron prácticamente un resultado sólido.

La tabla 7.1 presenta el detalle de esta validación por constructo.

Tabla 7.1. *Validez de contenido del cuestionario TAM/UTAUT-Cálculo por dimensión teórica*

<i>Dimensión teórica</i>	<i>Fundamento</i>	<i>N.º ítems</i>	<i>CVC Claridad</i>	<i>CVC Pertinencia</i>	<i>CVC Relevancia</i>	<i>CVC Promedio</i>
Utilidad percibida	TAM (Davis, 1989)	6	0.96	0.98	0.96	0.96
Facilidad de uso	TAM (Davis, 1989)	5	0.96	1.00	0.98	0.98
Satisfacción y actitud	TAM/UTAUT	5	0.99	0.97	0.97	0.98
Impacto motivacional	SDT (Deci y Ryan, 2000)	6	0.93	1.00	0.91	0.95
Carga cognitiva	CLT (Sweller, 2011)	3	0.88	1.00	1.00	0.96
Preguntas abiertas	Análisis cualitativo	5	0.84	1.00	1.00	0.95
Total instrumento	—	30	0.93	0.99	0.97	0.96

Nota; CVC = Coeficiente de Validez de Contenido. Criterio de aceptación: $CVC \geq 0.80$ (Hernández-Nieto, 2002). TAM = Technology Acceptance Model; UTAUT = Unified Theory of Acceptance and Use of Technology; SDT = Self-Determination Theory; CLT = Cognitive Load Theory.

Fuente: elaboración propia con base en Davis (1989), Deci y Ryan (2000), Hernández-Nieto (2002) y Sweller et al. (2011).

Como se muestra en la tabla 7.1, se detectó que existe área de oportunidad en dos reactivos específicos, en los cuales se obtuvieron valores por debajo del umbral de 0.80; para redimirlos, se tomaron los datos numéricos con notas de los expertos.

La tabla 7.2 exhibe los ítems con resultado de cvc menor al 0.80, mencionando las modificaciones realizadas; esto confirma el valor real del cvc, que no solo es un filtro estadístico, funciona como un parámetro para mejorar y afinar el instrumento.

Tabla 7.2. *Ítems con CVC inferior a 0.80 en un criterio: análisis individual y decisión metodológica*

<i>Ítem</i>	<i>Dimensión</i>	<i>Criterio</i>	<i>CVC</i>	<i>Observación del panel</i>	<i>Decisión adoptada</i>
28	Impacto motivacional	Relevancia	0.70	Dos jueces argumentaron que el constructo "reducción de ansiedad matemática" se ubica más en el dominio psicológico que en el motivacional.	Conservar el ítem: su CVC promedio fue 0.87 y la SDT vincula la ansiedad con la motivación intrínseca (Deci y Ryan, 2000).
33	Preguntas abiertas	Claridad	0.70	Un experto identificó terminología ambigua en la pregunta original.	Reformular el ítem incorporando ejemplos concretos de dificultades técnicas, pedagógicas y personales.

Nota: CVC = Coeficiente de Validez de Contenido. SDT = Self-Determination Theory. Solo se incluyen los ítems cuyo CVC en algún criterio fue inferior al umbral de 0.80 (Hernández-Nieto, 2002); todos los demás ítems superaron el criterio en los tres aspectos evaluados.

Fuente: elaboración propia con base en Hernández-Nieto (2002) y Deci y Ryan (2000).

A través del coeficiente alfa de Cronbach se calculó la confiabilidad del instrumento. En general, se alcanzaron valores de 0.957 para el grupo que usó Khan Academy, mientras que el grupo que usó ChatGPT-4 obtuvo un 0.938. Sin embargo, los resultados obtenidos por dimensiones reflejan matices distintos. El instrumento arrojó mejores resultados con Khan Academy, con diferencias significativas ($\Delta\alpha>0.10$) en las dimensiones de Facilidad de uso, Impacto motivacional y Carga cognitiva. Particularmente, la última dimensión señalada (Carga cognitiva) alcanzó un nivel aceptable ($\alpha=0.745$), mientras que esa misma dimensión, con el grupo de ChatGPT-4, obtuvo un resultado cuestionable ($\alpha=0.641$).

La tabla 7.3 presenta el detalle de esta comparación psicométrica.

Tabla 7.3. Comparación de confiabilidad (alfa de Cronbach) entre versiones del cuestionario

<i>Dimensión teórica</i>	<i>Khan Academy (n = 11)</i>	<i>ChatGPT-4 (n = 14)</i>	<i>$\Delta\alpha$</i>	<i>Interpretación</i>
Utilidad percibida (6 ítems)	$\alpha = 0.889$	$\alpha = 0.857$	+0.032	Buena en ambos
Facilidad de uso (5 ítems)	$\alpha = 0.892$	$\alpha = 0.792$	+0.100	Khan Academy superior
Satisfacción y actitud (5 ítems)	$\alpha = 0.913$	$\alpha = 0.886$	+0.027	Excelente / Buena
Impacto motivacional (6 ítems)	$\alpha = 0.852$	$\alpha = 0.743$	+0.109	Khan Academy superior
Carga cognitiva (3 ítems)	$\alpha = 0.745$	$\alpha = 0.641$	+0.104	Khan superior; ChatGPT cuestionable
Escala total (25 ítems)	$\alpha = 0.957$	$\alpha = 0.938$	+0.019	Excelente en ambos

Nota: Criterios de interpretación según George y Mallery (2020): $\alpha \geq 0.90$ = Excelente; 0.80–0.89 = Bueno; 0.70–0.79 = Aceptable; < 0.70 = Cuestionable. Los valores deben interpretarse como estimaciones preliminares dado el tamaño muestral del piloto ($n = 11-14$); la estabilización del alfa se espera con $n \geq 30$. Fuente: elaboración propia a partir de los datos del pilotaje y con base en los criterios de interpretación de George y Mallery (2020).

Solamente un ítem tuvo problemas en el análisis ítem-total. En Khan Academy, los 25 ítems superaron el punto de corte ($\text{rit} \geq 0.30$), lo cual funcionó adecuadamente. Para ChatGPT-4, 24 ítems cumplieron con la norma; no obstante, el reactivo 31 no lo hizo porque obtuvo un $\text{rit} = 0.29$. Desde un punto de vista estadístico, eliminarlo habría elevado el alfa a 0.942. Sin embargo, por razones metodológicas, optamos por conservarlo y cambiar

su redacción. El problema era la formulación: sustituimos la frase ambigua “Usar ChatGPT-4 requirió poco esfuerzo mental adicional” por una más precisa y contundente: “Emplear ChatGPT-4 no me causó sobrecarga mental al estudiar cálculo” (Sweller et al., 2011).

Ese rango diverso de respuestas evidenció un hallazgo relevante: la interacción con la IA puede resultar cognitivamente demandante, porque el estudiante tiene que crear el prompt, depurar errores y verificar la información con sus clases. Se analizaron las preguntas abiertas para complementar estos descubrimientos. La tabla 7.4 sintetiza lo que los alumnos dijeron sobre la utilidad, los desafíos y sus recomendaciones utilizando el método temático de Braun y Clarke (2006).

Tabla 7.4. *Temas emergentes en las respuestas abiertas, por herramienta*

<i>Eje</i>	<i>Tema emergente</i>	<i>Khan Academy (n = 12)</i>	<i>ChatGPT-4 (n = 14)</i>	<i>Lectura comparativa</i>
Utilidad	Ejercicios estructurados	58% (7)	7% (1)	Fortaleza Khan
	Videos explicativos	50% (6)	0% (0)	Exclusivo de Khan
	Explicación paso a paso	8% (1)	64% (9)	Fortaleza ChatGPT-4
	Disponibilidad 24/7	0% (0)	14% (2)	Ventaja ChatGPT-4
Dificultades	Errores en cálculos	0% (0)	36% (5)	Limitación ChatGPT-4
	Formular preguntas	0% (0)	21% (3)	Barrera ChatGPT-4
	Navegación / temas	17% (2)	0% (0)	UX Khan mejorable
Sugerencias	Mejorar precisión cálculos	0% (0)	29% (4)	Limitación de IA
	Capacitación en prompts	0% (0)	14% (2)	Taller necesario
	Sin cambios	33% (4)	21% (3)	Mayor satisfacción Khan

Nota: los porcentajes derivan del conteo manual de respuestas textuales; un mismo estudiante pudo aportar más de un tema.

Fuente: elaboración propia a partir del análisis de las respuestas abiertas del pilotaje, mediante el método temático de Braun y Clarke (2006).

El análisis cualitativo hizo posible diferenciar dos perfiles de uso muy claros. La estructura de la propuesta de Khan Academy convenció a los alumnos: un 58% apreció sus ejercicios y un 50% resaltó los videos (E9: “Hay aspectos que solo se aprenden al explicarlos y verlos”). ChatGPT-4, sin embargo, captó el interés del 64% de los estudiantes debido a su habilidad para descomponer problemas paso a paso y por estar disponible en todo momento: “24/7”, como indicó E2.

Cuando se trató de los obstáculos, las percepciones estaban claramente divididas. El mayor inconveniente en Khan Academy fue la interfaz. El estudiante E7 observó que le resultaba difícil vincular los nombres de las materias con el buscador. El problema relacionado con ChatGPT-4 se centró en su precisión: un 36% detectó errores en las matemáticas (E3: “cometió errores con frecuencia”). Como resultado de este hallazgo, se realizaron cambios en la estructura original. Las intervenciones futuras implicarán un taller de enseñanza matemática orientado hacia la redacción de los prompts de ChatGPT-4. Finalmente, la tabla 7.5 resume las metas alcanzadas y las restantes a cumplir. La validez y la fiabilidad globales del cuestionario son validadas; cabe señalar que la demanda cognitiva que implique la inteligencia artificial resultará un aspecto fundamental en el análisis para toda la muestra.

Tabla 7.5. Síntesis de evidencias de validación y decisiones derivadas

Tipo de validez	Método	Khan Academy	ChatGPT-4	Decisión
Contenido	Juicio de expertos (CVC)	0.96	0.96	Excelente; ítems 28 y 33 reformulados
Consistencia interna	Alfa de Cronbach (total)	0.957	0.938	Excelente preliminar (n piloto)
Consistencia interna	Alfa por dimensión	0.745–0.913	0.641–0.886	Carga cognitiva ChatGPT cuestionable
Análisis ítem-total	$r_{it} \geq 0.30$	Cumple los 25	24 de 25 (ítem 31)	Reformulación del ítem 31
Validez ecológica	Tiempo y comprensión	≈ 16 min	≈ 17 min	Aceptable
Cualitativa	Análisis temático	3 ejes	3 ejes	Capacitación en prompts (ChatGPT-4)
Constructo (AFC)	AFC multigrupo	Pendiente	Pendiente	Programado para $n \approx 105-120$

Nota: Pendiente = no procede en pilotaje por insuficiencia de potencia estadística (Kline, 2016). Las alfas se reportan como estimaciones preliminares dado el tamaño muestral ($n = 11-14$).
Fuente: elaboración propia a partir de los resultados del pilotaje y con base en Hernández-Nieto (2002), George y Mallery (2020), Braun y Clarke (2006) y Kline (2016).

Plan analítico para la muestra definitiva

El ciclo de enero a junio de 2026 corresponde al periodo programado para la intervención principal con estudiantes de cálculo en la FIME. La intervención incluirá una participación conjunta con tres grupos integrales, lo que ofrece una muestra aproximada de $n \approx 105-120$. Los resultados definitivos se comunicarán a través de un informe posterior, sin embargo, el análisis estadístico está descrito en este documento. Cabe señalar que los datos descriptivos mostrados en el apartado 3.2 no tienen alcance inferencial; cualquier juicio sobre la eficacia del proyecto deberá formularse una vez disponible la base de datos completa.

En el plano psicométrico, la muestra amplia permitirá recalcular el alfa de Cronbach con mayor precisión, incorporando intervalos de confianza al 95% (Feldt et al., 1987). La literatura señala que al superar un $n \geq 30$, el coeficiente tiende a estabilizarse (George y Mallery, 2020). También avanzaremos hacia la validación estructural mediante un análisis factorial confirmatorio (AFC) multigrupo. El objetivo será comprobar el modelo de cinco dimensiones apoyándonos en los índices de Kline (2016): χ^2/gl , CFI, TLI, RMSEA y SRMR. En paralelo, evaluaremos la invarianza factorial; sin este paso, sería metodológicamente incorrecto comparar a los usuarios de ChatGPT-4 con los de Khan Academy.

En la fase cuasi-experimental, vamos a tratar los datos desde tres perspectivas. En cuanto a la parte académica, utilizaremos un ANCOVA (teniendo en cuenta el examen diagnóstico) para comparar el rendimiento de los grupos y determinaremos el tamaño del efecto utilizando η^2_p y d de Cohen. En la dimensión perceptual, emplearemos regresiones múltiples para relacionar las calificaciones con las respuestas del cuestionario, de modo que podamos comprobar su validez a nivel de criterio. Por último, llevaremos a cabo una triangulación explicativa: los resultados cualitativos se utilizarán para interpretar la variación cuantitativa. Al concluir este ciclo, se contará con las respuestas definitivas del estudio.

Conclusiones

Desde su diseño original, el cuestionario TAM/UTAUT-Cálculo requirió una validación estadística rigurosa; el presente capítulo establece precisamente esa base. La metodología presentada, más que mostrar dos versiones paralelas, sirve como una hoja de ruta que se puede replicar para investigaciones que afrontan el desafío de trabajar con muestras pequeñas en sus etapas iniciales. Los resultados respaldan el diseño: la evaluación de cinco expertos arrojó un cvc de 0.96 y la consistencia interna alcanzó niveles óptimos ($\alpha=0.957$ en Khan Academy y $\alpha=0.938$ en ChatGPT-4).

Por supuesto, la prudencia metodológica obliga a tratar estos coeficientes como estimaciones preliminares, dado el tamaño limitado del piloto. Aun así, la tendencia de fiabilidad es clara. Presentar índices sin potencia estadística sería cuestionable; por ello, se optó por la transparencia y se difirió explícitamente el análisis factorial confirmatorio hasta contar con la muestra definitiva.

Las diferencias psicométricas y los resultados cualitativos también sugieren directrices adecuadas para el diseño de la enseñanza. Las dos plataformas no son intercambiables bajo un único enfoque educativo. ChatGPT-4 demanda que el estudiante domine la formulación de instrucciones precisas y requiere supervisión continua para detectar errores algebraicos del sistema. Khan Academy, en cambio, opera mediante rutas predefinidas que estructuran el seguimiento de las tareas. El alcance de este trabajo trasciende la evaluación instrumental. Reducir los índices de reprobación en Cálculo en la FIME no se logrará únicamente mediante la integración tecnológica. Las decisiones institucionales requieren evidencia empírica que justifique el uso reflexivo y crítico de las herramientas.

Referencias

Álvarez-Aguilar, N. T., Treviño-Cubero, A., y Castillo-Elizondo, J. A. (2024). Autovaloración de estudiantes universitarios sobre aplicación y potencialidades de la Inteligencia

- Artificial para su formación. *Atenas*, (62), e10798, 1-14. <https://atenas.umcc.cu/index.php/atenas/article/view/976>
- Artigue, M. (2002). Learning mathematics in a CAS environment: The genesis of a reflection about instrumentation and the dialectics between technical and conceptual work. *International Journal of Computers for Mathematical Learning*, 7(3), 245-274. <https://doi.org/10.1023/A:1022103903080>
- Braun, V., y Clarke, V. (2006). Using thematic analysis in psychology. *Qualitative Research in Psychology*, 3(2), 77-101. <https://doi.org/10.1191/1478088706qp0630a>
- Creswell, J. W., y Plano Clark, V. L. (2018). *Designing and conducting mixed methods research* (3.ª ed.). SAGE.
- Davis, F. D. (1989). Perceived usefulness, perceived ease of use, and user acceptance of information technology. *MIS Quarterly*, 13(3), 319-340. <https://doi.org/10.2307/249008>
- Deci, E. L., y Ryan, R. M. (2000). The “what” and “why” of goal pursuits: Human needs and the self-determination of behavior. *Psychological Inquiry*, 11(4), 227-268. https://doi.org/10.1207/S15327965PLI1104_01
- Feldt, L. S., Woodruff, D. J., y Salih, F. A. (1987). Statistical inference for coefficient alpha. *Applied Psychological Measurement*, 11(1), 93-103. <https://doi.org/10.1177/014662168701100107>
- George, D., y Mallery, P. (2020). *IBM SPSS Statistics 26 step by step: A simple guide and reference* (16.ª ed.). Routledge. <https://doi.org/10.4324/9780429056765>
- Hernández-Nieto, R. A. (2002). *Contributions to statistical analysis*. Universidad de Los Andes.
- Kasneci, E., Seßler, K., Küchemann, S., Bannert, M., Dementieva, D., Fischer, F., Gehrke, M., Krusche, S., Kutyniok, G., Michaeli, T., Nerdel, C., Pfeffer, J., Poquet, O., Sailer, M., Schmidt, A., Seidel, T., Stadler, M., Weller, J., Widmer, J., y Kasneci, G. (2023). ChatGPT for good? On opportunities and challenges of large language models for education. *Learning and Individual Differences*, 103, 102274. <https://doi.org/10.1016/j.lindif.2023.102274>
- Kline, R. B. (2016). *Principles and practice of structural equation modeling* (4.ª ed.). Guilford Press.
- Lara Pinales, O., y Neira Rosales, A. M. (2023). Reporte de aprobación y reprobación por unidades de aprendizaje [Conjunto de datos]. Facultad de Ingeniería Mecánica y Eléctrica, Universidad Autónoma de Nuevo León.
- Lawshe, C. H. (1975). A quantitative approach to content validity. *Personnel Psychology*, 28(4), 563-575. <https://doi.org/10.1111/j.1744-6570.1975.tb01393.x>
- Sweller, J., Ayres, P., y Kalyuga, S. (2011). *Cognitive Load Theory*. Springer. <https://doi.org/10.1007/978-1-4419-8126-4>
- Venkatesh, V., Morris, M. G., Davis, G. B., y Davis, F. D. (2003). User acceptance of information technology: Toward a unified view. *MIS Quarterly*, 27(3), 425-478. <https://doi.org/10.2307/30036540>

