Fecha de presentación: octubre, 2023, **Fecha de Aceptación**:noviembre, 2023, **Fecha de publicación**: enero, 2024

14

LA EDUCACIÓN SUPERIOR VIRTUAL Y LAS PRUEBAS ESTANDARIZADAS

VIRTUAL HIGHER EDUCATION AND STANDARDIZED TESTS

Hendry Luzardo Martínez[1]
E-mail: hendry.luzardo@ugm.cl
ORCID: https://orcid.org/0000-0001-5083-6074
Daniel Alfredo Meza Molina [2]
E-mail: dmezavech@gmail.com
ORCID: https://orcid.org/0009-0008-3735-2444
Teresa de Jesús Molina Gutiérrez [3]
E-mail: ui.teresamolina@uniandes.edu.ec
ORCID: https://orcid.org/0000-0002-5957-3482
Camila Francisca Godoy Tapia [1]
E-mail: virtual.camila.godoy@ugm.cl
ORCID: https://orcid.org/0000-0002-2522-0363
[1] Universidad Gabriela Mistral, Santiago de Chile. Chile.
[2] Empresa Binder Dijker Otte. Senior Contable. Ibarra.
[3] Universidad Regional Autónoma de Los Andes, Ibarra. Ecuador.

Cita sugerida (APA, séptima edición)

Luzardo Martínez, H., Meza Molina, D. A., Molina Gutiérrez, T. J. de. y Godoy Tapia, C. F. (2024). La Educación Superior Virtual y las pruebas estandarizadas. ***Revista Conrado, 20***(96), 148-156.

RESUMEN

La educación superior en función de la formación de profesionales adopta variadas modalidades de estudio y su evaluación, en la cual la virtual ocupa un lugar importante. La evaluación de forma virtual constituye la aplicación de las potencialidades que brinda la tecnología, que a su vez no solo tiene aspectos positivos, sino que recibe variadas críticas, lo que hace que se requiera su estudio con el fin de lograr un mayor perfeccionamiento. Se realiza un análisis de la validez y confiabilidad de las pruebas estandarizadas mediante la medición de atributos psicométricos. Para obtener los datos de las evaluaciones aplicadas en una carrera universitaria con modalidad en línea y describir los rasgos de las estadísticas de las preguntas, se emplea la metodología descriptiva con diseño de campo. Los resultados obtenidos muestran que las pruebas analizadas se consideran con una calidad técnica satisfactoria.

Palabras clave:

Evaluación, pruebas, educación, confiabilidad, validez.

ABSTRACT

Higher education based on the training of professionals adopts various study modalities and their evaluation, in which the virtual one occupies an important place. Virtual evaluation constitutes the application of the potential offered by technology, which in turn not only has positive aspects but also receives various criticisms, which requires its study in order to achieve greater improvement. An analysis of the validity and reliability of standardized tests is carried out by measuring psychometric attributes. To obtain the data from the evaluations applied in a university degree with an online modality and describe the characteristics of the statistics of the questions, the descriptive methodology with field design is used. The results obtained show that the analyzed tests are considered to have satisfactory technical quality.

Keywords:

Evaluation, testing, education, reliability, validity.

INTRODUCCIÓN

Las tendencias actuales apuntan a que las personas pretenden aprender y estudiar en entornos flexibles. Por ende, la educación virtual en la educación superior juega un rol fundamental para satisfacer esta tendencia (Durán et al., 2015). Los estudiantes deben ser formados para desarrollar competencias que los preparen para enfrentar un ambiente incierto, complejo y de posibilidades ilimitadas (Olivares et al., 2018). En la educación superior virtual, el uso de pruebas estandarizadas está siempre presente y sigue siendo cada vez más común.

Estas pruebas se utilizan para medir el conocimiento de los estudiantes sobre los contenidos estudiados y para evaluar la eficacia del proceso de enseñanza y aprendizaje en línea. Por lo tanto, es importante tener en cuenta que la calidad técnica de estas pruebas, en términos de su validez y confiabilidad, es perentorio para avalar que las decisiones basadas en sus resultados sean precisas y confiables. Sin embargo, existe una falta de información sobre la capacidad de los LMS o sistemas de gestión del aprendizaje para llevar a cabo análisis psicométricos de manera eficiente y efectiva, aspecto necesario al tener en cuenta que mientras se fomenta el pensamiento crítico, uno de sus efectos es que el estudiante se enfrenta a una vasta y compleja cantidad de información (Dominguez y Vega, 2020).

Las universidades tienen como reto impulsar la calidad de la enseñanza para lo que se hace necesaria la puesta en marcha de metodologías didácticas centradas en el alumnado (Rodríguez, 2020). Algunos estudios señalan como las principales limitaciones o dificultades que los estudiantes han detectado en esta forma virtual de educación son la falta de adaptabilidad, las notas no muy sobresalientes, el desconocimiento de las formas de evaluación (Rodriguez et al, 2022).

En este ámbito dentro de los antecedentes referenciales está Gutiérrez y Acuña (2021), quienes analizaron en la literatura el abordaje de la evaluación del aprendizaje desde el enfoque estandarizado, con el empleo de mediciones psicométricas, además de señalar que la evaluación del aprendizaje ha experimentado un gran desarrollo como resultado del uso de métodos y teorías de influencia europea. También, sostienen que no se cuenta con información suficiente que explique la capacidad de los softwares educativos evaluativos para efectuar análisis psicométricos. Por su parte Báez (2020) realizó un estudio comparativo sobre los resultados de las pruebas Saber 11° aplicadas en Colombia a los estudiantes de colegios Públicos y Privados (2019), sus hallazgos indican que es necesario cuestionarse y reflexionar al realizar comparaciones relativas a la calidad de la educación de acuerdo con el desempeño de estudiantes en las pruebas estandarizadas, ya que los resultados pueden variar a partir de la desigualdad que existe entre la educación pública y privada.

En el mismo sentido, Medina y Verdejo (2020) indagaron sobre las evidencias que determinan la validez y confiabilidad de las puntuaciones obtenidas mediante los instrumentos evaluativos aplicados en el contexto universitario. Sus resultados puntualizan que es fundamental considerar la evidencia relacionada con el contenido y la consistencia de las puntuaciones (evidencia de validez) al elaborar juicios y tomar decisiones que repercuten en los estudiantes. Por último, Backhoff (2018) se interesó en precisar el concepto de evaluaciones estandarizadas de aprendizaje, explicar sus usos y establecer las limitaciones y retos que enfrentan este tipo de instrumentos. Dan relevancia al aporte de las ciencias cognitivas para mejorar la validez y contenido de ese tipo de evaluaciones y consideran que falta mucha investigación acerca de los elementos cognitivos y su medición. Concluyen que los problemas generados al determinar consecuencias asociadas a las evaluaciones no son únicos de las pruebas estandarizadas, “son comunes a cualquier instrumento cuyos resultados tengan consecuencias positivas o negativas, ya sea para los estudiantes, los profesores” (s/p) o para la institución.

Acerca de los conceptos claves que fundamentan la investigación, se destaca que, “La validez se refiere a la medida en que una prueba mide lo que pretende medir y cómo lo mide” (American Educational Research Association [AERA], American Psychological Association [APA], & National Council on Measurement in Education [NCME], 2014, p. 11); en consecuencia, si una prueba no es válida, los resultados pueden ser engañosos y no reflejan el conocimiento o la habilidad real del estudiante. Por otra parte, “La confiabilidad se refiere a la consistencia o estabilidad de las mediciones a través de diferentes ocasiones, observadores, elementos de prueba y otras condiciones” (McMillan & Schumacher, 2019, p. 230), que determina que una prueba confiable debe producir resultados consistentes cuando se administra a diferentes grupos de estudiantes o en diferentes momentos.

El objetivo de este artículo es analizar la calidad técnica (validez y confiabilidad) de las pruebas estandarizadas utilizadas en la educación superior virtual, mediante la medición de atributos psicométricos, con las estadísticas de las preguntas. Para lograr esto, el estudio examina la calidad técnica de 21 pruebas estandarizadas aplicadas a estudiantes del primer trimestre de una carrera universitaria en la modalidad virtual en términos de su validez

y confiabilidad. Además, se analizan los métodos y técnicas utilizadas para medirlas, que constituye una valiosa contribución al campo de la evaluación de aprendizajes en la educación superior virtual, ya que proporciona una revisión crítica de la calidad técnica de las pruebas estandarizadas utilizadas en este contexto. Los hallazgos y recomendaciones presentados en este artículo son de gran ayuda para mejorar la calidad de las pruebas y, por lo tanto, para garantizar la validez y confiabilidad de los resultados de la evaluación del aprendizaje en línea, permite una mayor comprensión de los ajustes requeridos para lograr la calidad de las pruebas utilizadas en la evaluación del aprendizaje en línea.

MATERIALES Y MÉTODOS

Se emplea la metodología descriptiva, apoyada en el diseño de campo, lo cual facilita obtener los datos de las evaluaciones aplicadas en una carrera universitaria (modalidad en línea) para describir los rasgos de las estadísticas (atributos psicométricos) de las preguntas como: Índice de facilidad promedio, Desviación estándar, Índice de discriminación promedio y Promedio de eficiencia discriminativa (validez y confiabilidad). La relación teoría-datos se obtuvo mediante el método analítico-sintético y los datos fueron recolectados de los registros almacenados en la plataforma de aprendizaje Moodle y se organizaron en matrices estadísticas (método del nivel empírico).

La muestra se selecciona mediante muestreo intencional (primer trimestre de la carrera PT23, una sección, período Octubre 2020-diciembre 2022, Universidad LGMC), las pruebas escogidas son 21, codificadas de manera consecutiva (T1A1P1 - 21), el número de preguntas por prueba oscila entre 4 y 88, al sumarlas se obtuvo un total de 1047 ítems. Los datos se analizan con técnicas de estadística descriptiva (porcentajes, frecuencias y desviación estándar).

RESULTADOS Y DISCUSIÓN

Los datos en análisis se componen de una muestra estadística de 21 pruebas con 49.8 preguntas en promedio, sobre las cuales se determinaron una serie de índices de evaluación que denotan la calidad técnica de las pruebas, se entiende a esta calidad técnica como la capacidad de las pruebas para captar y medir el conocimiento de los estudiantes que rinden estas pruebas.

Los índices que se utilizan denotan las siguientes relaciones:

1. Índice de facilidad promedio: denota la puntuación promedio de los estudiantes en la prueba y se estratifica de la siguiente forma:

Tabla 1. Rango de valores índice de facilidad

Índice de Facilidad	Interpretación
5% o menos	Extremadamente difícil, o algo está mal con la pregunta.
6% - 10%	Muy difícil.
11% - 20%	Difícil.
21% - 34%	Moderadamente difícil.
35% - 65%	Correcta para el estudiante promedio.
66% - 80%	Bastante fácil.
81% - 89%	Fácil.
90% - 94%	Muy fácil.
95% - 100%	Extremadamente fácil.

Fuente: Elaboración propia

2. Desviación estándar: La desviación estándar es una medida de la variabilidad de una distribución de puntuaciones, lo que indica cuánto se alejan los datos individuales de la media (Weiner & Greene, 2017, p. 116). En el contexto de las pruebas, es entendida como una medida de dispersión de los datos de calificaciones con respecto al valor medio de la muestra y, por tanto, una medida básica de la magnitud con que la pregunta puede discriminar o discernir entre los estudiantes más preparados y menos preparados.
3. Índice de discriminación promedio: El índice de discriminación promedio es una medida del poder de discriminación de una prueba y se define como la diferencia media entre las puntuaciones de los estudiantes que obtienen una puntuación alta en la prueba y aquellos que obtienen una puntuación baja (Gutierrez y Gamboa, 2022). En otras palabras, es la correlación existente entre las calificaciones ponderadas en la pregunta y las del resto de la prueba. Indica que tan efectiva es la pregunta para clasificar, separar o discernir a los estudiantes más capaces de los menos capaces y se estratifica de la siguiente forma, según Moodle (S.F.): Tabla 2

Tabla 2. Rango de valores índice discriminación promedio

Índex	Interpretación
50% y superior	Muy buena discriminación.
30% – 50%	Adecuada discriminación.
20% - 29%	Débil discriminación.
0 - 19%	Muy débil discriminación.
valores negativos	La pregunta probablemente sea inválida.

Fuente: Elaboración propia

4. Promedio de eficiencia discriminativa: Estima que tan bueno es el índice de discriminación en relación con la dificultad de la pregunta. Tabla 3

Tabla 3. Rango de valores promedio de eficiencia discriminativa

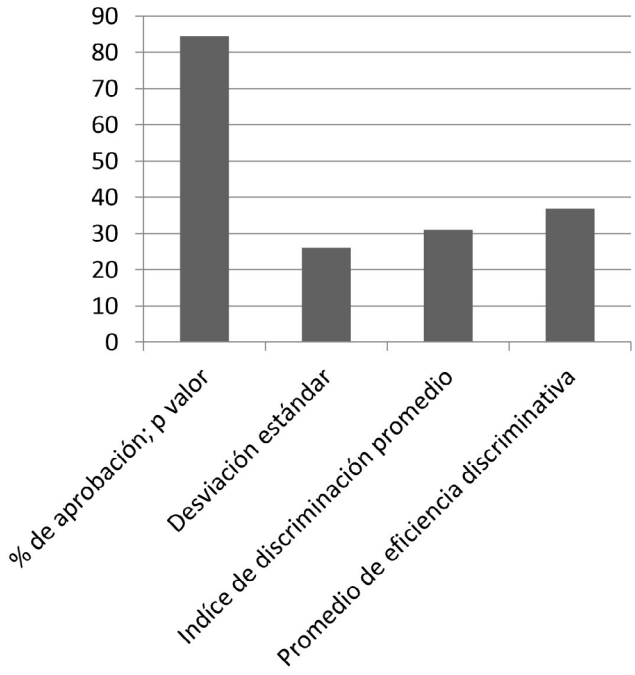
Índex	Interpretación
50% y superior	Sí son preguntas particularmente buenas
49% y Menor	No son preguntas particularmente buenas (requieren revisión).

Fuente: Elaboración propia

El análisis de los datos generales de la muestra estadística, evidencia un alto comportamiento en el porcentaje de aprobación y el promedio de la eficiencia discriminativa. Resultan inferiores el índice de discriminación promedio y la desviación estándar entre las 21 pruebas en estudio (Figura 1).

Fig. 1: Datos generales de la muestra estadística.

Fuente: Elaboración propia.

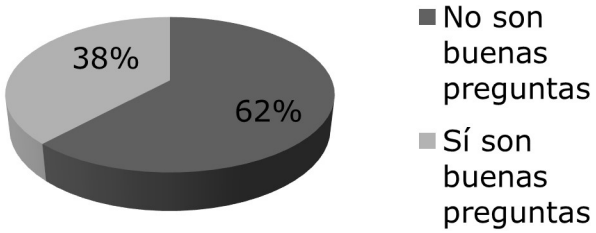


En términos generales los índices derivados de la muestra estadística denotan que, en promedio, la facilidad de las pruebas se ubica en un 84.6%, lo cual sitúa a la batería de pruebas en el rango de fácil, esto implica que, en términos generales, los estudiantes han logrado aprobar las pruebas de manera satisfactoria. Por otra parte, en términos generales, el valor medio de la desviación estándar muestral se sitúa en 26.1%, un valor entendido como aceptable por parte de la institución, esto implica una variabilidad de los resultados que está ajustada al valor esperado normal, lo cual, a su vez, indica que los resultados obtenidos por los estudiantes son satisfactorios y la calidad técnica de la prueba es buena.

El índice de discriminación promedio arroja un valor medio total de 31.1%, lo que sitúa este resultado en el rango de adecuada discriminación, la significación de este dato, en términos generales, denota que las preguntas desarrolladas en las pruebas tienen la capacidad real para discriminar entre los estudiantes capaces de los no capaces.

Ahora bien, el promedio de eficiencia discriminativa plantea un resultado que va en contra de la corriente de los índices anteriores, el valor medio muestral de este índice se sitúa en 39.6%, lo cual, para los estándares aceptados por la institución, se cataloga como que no son preguntas particularmente buenas. Dando entender que, si bien los estudiantes han logrado aprobar las pruebas y en términos general, las medidas de dispersión se encuentran dentro de los parámetros aceptados, la forma en la que se redactan o estructuran las preguntas pueden ser mejoradas para cerrar la brecha entre la intención del evaluador y el significado que les dan a las preguntas los estudiantes (Figura 2).

Fig. 2: Nivel de adecuación de las preguntas.



Fuente: Elaboración propia.

El hecho que los resultados generales de la muestra estadística permiten determinar además que la mayoría de las preguntas resultan no ser particularmente buenas, denota que se puede mejorar el cuestionario para continuar con el desarrollo del estudio en nuevas variantes y objetos relacionados con la temática. Se puede lograr una mayor discriminación de forma adecuada que permite adecuar además el número de preguntas en cada una. Sin embargo, más allá de esta observación referida

al promedio de eficiencia discriminativa, los resultados obtenidos en términos generales indican que las pruebas si poseen una calidad técnica satisfactoria.

Observando individualmente la desviación estándar de cada prueba, se puede indicar que existen elementos que se salen de los parámetros de aprobación definidos por la institución, al superar el techo de aceptación del 30% como medida límite de la desviación estándar. Es importante destacar que las pruebas que superan este techo de aceptación son también las pruebas que tienen el menor valor de índice de facilidad, inicialmente esto indica que la percepción de los estudiantes con respecto a estas pruebas es que son más difíciles que las otras y en un segundo plano las preguntas formuladas en estas pruebas pierden poder de discriminación, lo cual afecta su potencial para medir válidamente el conocimiento de los estudiantes (tabla 4). En consecuencia, hay debilidades en los atributos: índice de dificultad del ítem, índice de discriminación y coeficiente de discriminación, con los cuales, según Gutiérrez y Acuña (2021) se determina la calidad técnica de una prueba. Tabla 4

Tabla 4. Pruebas con desviación estándar superior al límite aceptado.

Código de prueba	N° de preguntas	% de aprobación; p valor	Desviación estándar
T1A1P1	40	80%	33%
T1A2P9	78	80%	37%
T1A3P11	40	82%	30%
T1A3P14	32	73%	38%
T1A3P16	78	73%	34%
T1A4P17	60	72%	41%
T1A4P20	60	81%	35%

Fuente: Elaboración propia.

De igual forma, en los casos en los que desviación estándar se encuentra por debajo del límite aceptado, el índice de facilidad es superior, lo que podría interpretarse a primera vista como una señal de que estas pruebas captaron correctamente el conocimiento de los estudiantes y a su vez, los estudiantes tuvieron la percepción de que las preguntas eran adecuadas y, por ende, tuvieron mejores resultados. Tabla 5

Tabla 5. Pruebas con desviación estándar inferior al límite aceptado.

Código de prueba	N° de preguntas	% de aprobación; p valor	Desviación estándar
T1A1P2	61	87%	28%
T1A1P3	30	91%	20%
T1A1P4	60	94%	16%
T1A1P5	30	90%	20%
T1A1P6	90	83%	29%
T1A2P7	4	97%	7%
T1A2P8	4	98%	4%
T1A3P10	40	84%	23%
T1A3P12	40	86%	23%
T1A3P13	60	84%	28%
T1A3P15	40	86%	23%
T1A4P18	60	85%	28%
T1A4P19	60	86%	28%
T1A4P21	80	85%	24%

Fuente: Elaboración propia.

Al respecto, es indudable el importante rol que tienen las pruebas estandarizadas en la evaluación, pues como lo señala Guevara (2017) permiten clasificaciones, comparaciones, rankings, sin embargo, es fundamental buscar perfeccionarlas ya que, en algunas ocasiones sus resultados deciden el destino de las instituciones educativas obviando criterios evaluativos fundamentales.

Ahora bien, si se agrega un piso adicional de análisis y cruce de los índices propuestos, las conclusiones pueden cambiar, tomando en consideración el índice de discriminación promedio, el cual indica qué tan efectiva es una pregunta para clasificar o discernir entre los estudiantes más capaces de los menos capaces. Se puede evidenciar ciertos patrones que denotan una relación entre un índice de facilidad en promedio bajo con una desviación estándar alta, generan una discriminación promedio débil, lo cual indica que puede existir un problema con la forma de redacción de las preguntas de la prueba y puede malograr la capacidad técnica esperada del test. Al respecto, Moreira y otros (2022), dan relevancia al hecho de que hay instrumentos que pueden mostrar grados aceptables de validez para fines y poblaciones específicas, pero no para otros; de modo, que "la validez no es un rasgo dicotómico, al contrario, se trata de una cuestión de grado en que la prueba exhibe un grado aceptable de validez para ciertos usos específicos y con ciertas poblaciones" (pág.22).

Como un ejemplo específico de esta situación está la prueba T1A1P1, la cual tiene un índice de facilidad del 80%, se sitúa entre las cuatros pruebas con menor calificación, a su vez tiene una desviación estándar del 33% que supera el techo de aceptación fijado por la institución para este índice, esto da como resultado un índice de discriminación promedio del 27% que la sitúa en un estándar de débil discriminación, lo cual se interpreta como que la prueba no es muy apta para discernir entre los estudiantes más aptos y menos aptos Tabla 6.

Tabla 6. Prueba T1A1P1.

Prueba	Índice facilidad	Desv. estándar	Índice de discriminación promedio	Interpretación	Promedio de eficiencia discriminativa	Interpretación
T1A1P1	80%	33%	27%	Débil discriminación.	41%	No son preguntas particularmente buenas

Fuente: Elaboración propia.

Como complemento al índice de discriminación promedio se incluye el promedio de eficiencia discriminativa, el cual se concibe como un estimador que relaciona el índice de discriminación promedio con la dificultad de la pregunta. Continuando con el ejemplo de la prueba T1A1P1 se observa que este promedio de eficiencia discriminativa se sitúa en 41%, lo cual evidencia que, para los estándares de clasificación definidos por la institución, indica que las preguntas no son particularmente buenas y que la prueba como tal tiene un problema para medir el conocimiento de los estudiantes, lo que a su vez afecta la eficiencia técnica de la prueba.

Ahora bien, un caso límite de las interpretaciones anteriormente planteadas, es la prueba T1A2P8 Tabla 7, la cual presenta características muy particulares en la que el primer rasgo importante a destacar es que solo se compone de 4 preguntas, significativamente menor que el resto de las pruebas, el índice de facilidad alcanzado por esta prueba es del 98%, lo cual podría interpretarse como muy fácil dentro del baremo proporcionado, la desviación estándar fue solo del 4%, muy por debajo de la media general de este índice, a primera vista podría interpretarse como un resultado positivo y sólido. Sin embargo, al evaluar el índice de discriminación promedio y el promedio de eficiencia discriminativa se puede constatar que los mismos se van a terreno negativo, indicando que las preguntas son inválidas y tienen poca o nula relación con lo que se pretende evaluar, adicionalmente las preguntas no son particularmente buenas.

Tabla 7. Prueba T1A2P8.

Prueba	Índice facilidad	Desv. estándar	Índice de discriminación promedio	Interpretación	Promedio de eficiencia discriminativa	Interpretación
T1A2P8	98%	4%	-8%	Débil discriminación.	-11%	No son preguntas particularmente buenas

Fuente: Elaboración propia.

A nivel general esto puede interpretarse como que los estudiantes tienen la sensación de que la prueba fue muy fácil y tienen la percepción de que sus resultados fueron óptimos, sin embargo, se evidencia que el planteamiento de las preguntas de la prueba no fue bien comprendido por los estudiantes y por ende, las respuestas dadas no están relacionadas con la materia que se pretende evaluar y existe una confusión generalizada por parte de los estudiantes, esto derriba por completo la eficiencia técnica de la prueba.

Una apreciación a nivel general que debe ser tomada en cuenta es que, al evaluar el promedio de eficiencia discriminativa la mayoría de las pruebas dan como resultado que las preguntas no son particularmente buenas, lo cual indica que existe una oportunidad de mejora generalizada en la forma de redacción o planteamiento de las preguntas aplicadas, sin embargo, esto no menoscaba el hecho de que la mayoría de las pruebas tienen un índice de discriminación promedio con una adecuada discriminación, lo que indica que las pruebas sí tienen la capacidad de discernir entre los estudiantes más aptos y menos aptos, la situación acá es que los instrumentos de evaluación se pueden mejorar para afinar este ratio de evaluación Tabla 8, aspectos que evidencian es necesario mejorar la calidad de la educación desde la evaluación, para lo cual se debe conceptualizar la evaluación como proceso complejo donde además de trabajar por optimizar los instrumentos de evaluación, se debe tener en cuenta los aspectos restantes que hacen la totalidad del sistema educativo.

Tabla 8. Promedio de eficiencia discriminativa.

Código de prueba	Índice facilidad	Desv. estándar	Índice de discriminación promedio	Interpretación	Promedio de eficiencia discriminativa	Interpretación
T1A1P1	80%	33%	27%	Débil discriminación.	41%	No son preguntas particularmente buenas
T1A1P2	87%	28%	32%	Adecuada discriminación.	44%	No son preguntas particularmente buenas
T1A1P3	91%	20%	30%	Adecuada discriminación.	47%	No son preguntas particularmente buenas
T1A1P4	94%	16%	40%	Adecuada discriminación.	51%	Si son preguntas particularmente buenas
T1A1P5	90%	20%	35%	Adecuada discriminación.	53%	Si son preguntas particularmente buenas
T1A1P6	83%	29%	36%	Adecuada discriminación.	54%	Si son preguntas particularmente buenas
T1A2P7	97%	7%	48%	Adecuada discriminación.	6%	No son preguntas particularmente buenas
T1A2P8	98%	4%	-8%	La pregunta probablemente sea inválida.	-11%	No son preguntas particularmente buenas
T1A2P9	80%	37%	32%	Adecuada discriminación.	41%	No son preguntas particularmente buenas
T1A3P10	84%	23%	25%	Débil discriminación.	32%	No son preguntas particularmente buenas
T1A3P11	82%	30%	38%	Adecuada discriminación.	50%	Si son preguntas particularmente buenas
T1A3P12	86%	23%	34%	Adecuada discriminación.	41%	No son preguntas particularmente buenas
T1A3P13	84%	28%	41%	Adecuada discriminación.	57%	Si son preguntas particularmente buenas
T1A3P14	73%	38%	27%	Débil discriminación.	38%	No son preguntas particularmente buenas
T1A3P15	86%	23%	34%	Adecuada discriminación.	41%	No son preguntas particularmente buenas
T1A3P16	73%	34%	11%	Muy débil discriminación.	15%	No son preguntas particularmente buenas

T1A4P17	72%	41%	45%	Adecuada discriminación.	58%	Si son preguntas particularmente buenas
T1A4P18	85%	28%	27%	Débil discriminación.	39%	No son preguntas particularmente buenas
T1A4P19	86%	28%	41%	Adecuada discriminación.	55%	Si son preguntas particularmente buenas
T1A4P20	81%	35%	41%	Adecuada discriminación.	55%	Si son preguntas particularmente buenas
T1A4P21	85%	24%	17%	Muy débil discriminación.	25%	No son preguntas particularmente buenas

Fuente: Elaboración propia.

Un caso modelo a ser estudiando es el caso de la prueba T1A4P19, la cual obtuvo resultados aceptables en todos los índices de evaluación, la misma consta de 60 preguntas donde el índice de facilidad alcanzó un 86%, definiéndola como fácil, a su vez la desviación estándar es de 28%, situándose por debajo del techo aceptado por la institución, el índice de discriminación promedio es del 41%, lo cual la sitúa como una prueba con adecuada discriminación y el ratio promedio de eficiencia discriminativa llega al 55%, lo cual indica que las preguntas son particularmente buenas para evaluar la materia del test Tabla 9.

Tabla 9. Prueba T1A2P8.

Prueba	Índice facilidad	Desv. estándar	Índice de discriminación promedio	Interpretación	Promedio de eficiencia discriminativa	Interpretación
T1A4P19	86%	28%	41%	Adecuada discriminación	55%	Sí son preguntas particularmente buenas

Fuente: Elaboración propia.

En opinión de Tristán y Pedraza (2017) la calidad técnica de las pruebas estandarizadas involucra que la misma sea aplicable "a todas las personas, en todos los ambientes y condiciones, obteniendo medidas libres de otras características ajenas al objeto" (pág. 22).

CONCLUSIONES

En conclusión, la evaluación del aprendizaje en la educación superior en línea es un tema de gran relevancia y las pruebas estandarizadas son una herramienta fundamental que permite medir el desempeño de los estudiantes y evaluar la eficacia del proceso de enseñanza y aprendizaje en línea. En este artículo, se realiza un análisis de la calidad técnica de estas pruebas, específicamente en términos de su validez y confiabilidad, mediante la medición de atributos psicométricos, donde resaltan resultados como el promedio de facilidad de las pruebas que ubica a las evaluaciones en el rango de fácil, el índice de discriminación que se encuentra en el rango de adecuado y la eficiencia discriminativa que ubica a las evaluaciones en el rango de preguntas que no son particularmente buenas. Es por lo que se recomienda a los diseñadores de pruebas estandarizadas en la educación superior virtual prestar mayor atención a los aspectos psicométricos y el desarrollo de análisis rigurosos para garantizar la validez y confiabilidad de las pruebas. Además, se sugiere que se utilicen herramientas que ofrecen los sistemas de gestión de aprendizajes, como el caso de Moodle, para medir y mejorar la calidad técnica de las pruebas, que contribuye a la comprensión de las limitaciones de estas pruebas y evidencia alternativas para mejorarlas, lo que repercute en optimizar su calidad técnica.

REFERENCIAS BIBLIOGRÁFICAS

American Educational Research Association, American Psychological Association y National Council on Measurement in Education. (2014). ***Standards for educational and psychological testing. Editorial American Educational Research Association***. https://cir.nii.ac.jp/crid/1130000794987222016

Báez, L. (2020). Confiabilidad de las pruebas estandarizadas que se aplican en Colombia para medir y evaluar la calidad de la educación. ***Revista Espacios***, ***41***(35), 1-16. https://www.revistaespacios.com/a20v41n35/a20v41n35p01.pdf

Backhoff, E. (2018). Evaluación estandarizada del logro educativo. Contribuciones y retos. ***Revista Digital Universitaria***, ***19***(6), 1-15. https://www.revista.unam.mx/ojs/index.php/rdu/article/view/1374

Dominguez, L. C. y Vega, N. V. (2020). Efectos del mapa conceptual sobre la síntesis de información en un ambiente de aprendizaje interactivo: Un estudio preexperimental. ***Educación Médica***, ***21***(3), 193-197. https://www.sciencedirect.com/science/article/pii/S1575181318302523

Durán, R., Estay-Niculcar, C. y Álvarez, H. (2015). Adopción de buenas prácticas en la educación virtual en la educación superior. ***Aula Abierta, 43***(2), 77-86. https://www.sciencedirect.com/science/article/pii/S0210277315000037

Gutiérrez, J. y Acuña. L. (2021). Evaluación estandarizada de los aprendizajes: una revisión sistemática de la literatura. ***Revista de Investigación Educativa, 34.*** 321-351. https://dialnet.unirioja.es/servlet/articulo?codigo=8349965

Gutierrez, J. y Gamboa, L. A. A. (2022). ***Evaluación estandarizada del aprendizaje en la educación superior: un estudio de caso en México***. Scielo Preprints. https://preprints.scielo.org/index.php/scielo/preprint/view/5126/9951

Guevara, R. (2017). La calidad, las competencias y las pruebas estandarizadas: una mirada desde los organismos internacionales. ***Educación y ciudad, 33***, 159-170. https://dialnet.unirioja.es/servlet/articulo?codigo=6232098

Medina, M. y Verdejo, A. (2020). Validez y confiabilidad en la evaluación del aprendizaje mediante las metodologías activas. ***Alteridad, 15***(2), 270-283. http://scielo.senescyt.gob.ec/scielo.php?pid=S1390-86422020000200270&script=sci_arttext

McMillan, J. H., & Schumacher, S. (2019). ***Investigación educativa: Una introducción conceptual***. Pearson Educación. https://revistas.uam.es/tarbiya/article/download/7222/7583/15021

Moodle. (s.f.). ***Estadísticas del reporte del examen***. Moodle. https://docs.moodle.org/all/es/Estad%C3%ADsticas_del_reporte_del_examen

Moreira, T., Alfaro, L., Brizuela, A., Chacón, C., Gómez, E., Jiménez, K., Jiménez, F., Mena, P., Montero, E., Picado, H., Rojas, G., Rojas, L., Smith, V., Solórzano, M. y Villarreal, M. (2022). ***Cuadernos metodológicos: Estándares de calidad para pruebas de alto impacto en el contexto académico y profesional costarricense***. Costa Rica: Instituto de Investigaciones Psicológicas, Universidad de Costa Rica. https://iip.ucr.ac.cr/sites/default/files/contenido/Cuaderno%20Metodol%C3%B3gico%209%20%20Est%C3%A1ndares%20de%20Calidad%20para%20las%20Pruebas.pdf

Olivares, S. L. O., Cabrera, M. V. L. y Valdez-García, J. E. (2018). Aprendizaje basado en retos: una experiencia de innovación para enfrentar problemas de salud pública. ***Educación Médica***, ***19***, 230-237. https://www.sciencedirect.com/science/article/pii/S157518131730178X

Rodriguez-Alarcon, J. F., Vinelli-Arzubiaga, D., Aveiro-Róbalo, T. R., Garlisi-Torales, L. D., Delgado, J. E. H., Marticorena-Flores, R. K., . . . y Mejia, C. R. (2022). Repercusiones académicas de la educación virtual en los estudiantes de Latinoamérica: validación de una escala. ***Educación Médica, 23***(3), 100741. https://www.sciencedirect.com/science/article/pii/S157518132200033X

Rodríguez-Izquierdo, R. M. (2020). Aprendizaje Servicio y compromiso académico en Educación Superior. ***Revista de Psicodidáctica***, ***25***(1), 45-51. https://www.sciencedirect.com/science/article/abs/pii/S1136103419300085

Tristán, A. y Pedraza, M. (2017). La Objetividad en las Pruebas Estandarizadas. ***Revista Iberoamericana de Evaluación Educativa, 10***(1), 11-31. https://dialnet.unirioja.es/servlet/articulo?codigo=5913179

Weiner, I. B., y Greene, R. L. (2017). ***Handbook of personality assessment***. John Wiley & Sons. https://books.google.com.cu/books/about/Handbook_of_Personality_Assessment.html?id=xQS3DQAAQBAJ&redir_esc=y