Análisis de instrumentos evaluativos empleados en ciencias durante la pandemia: selección múltiple, indicadores y rendimiento

Facultad de Medicina

Gabriela Paz Urrejola-Contreras[a,§,*], Miguel Angel Pérez-Lizama[a,¶]

Resumen

Introducción: Las evaluaciones de opción múltiple constituyen el instrumento ampliamente usado en ciencias para evaluar al estudiantado; sin embargo, la reciente pandemia exigió adaptar este tipo de instrumento al entorno virtual. Este contexto requirió valorar la calidad de los instrumentos mediante índices de discriminación, consistencia interna y relacionarlo con el rendimiento académico.

Objetivo: Evaluar los instrumentos de evaluación empleados en modalidad online durante la pandemia por COVID-19, y el rendimiento de los estudiantes en ciencias de la salud.

Método: Se realizó la revisión de los 5 instrumentos de la asignatura Estructura y Función formados por 290 bancos de preguntas aleatorias para evaluar cada contenido en estudiantes de primer año durante el 2020 en la escuela de ciencias de la salud en la Universidad Viña del Mar. Se analizaron los datos obtenidos a partir de la plataforma virtual y se interpretaron los índices de discriminación, facilidad, eficiencia discriminativa, consistencia interna y rendimiento académico mediante un informe que fue compartido con los docentes para identificar los parámetros de calidad y validez.

Resultados: Del total de bancos de preguntas evaluados, un 70.2% de las preguntas presentaron adecuada discriminación y solo un 5.6% debieran ser eliminadas. El certamen dos obtuvo el menor rendimiento promedio 3.9 ± 0.99; sin embargo, presentó la consistencia interna más alta: 81%. Al comparar todos los instrumentos se observó una mejora gradual en la formulación, reflejada en el examen final, en el que además el rendimiento académico concuerda con el promedio del semestre 4.2 ± 0.92.

Conclusiones: El rendimiento académico debe ponderarse en relación con la calidad del instrumento formulado

[a] Escuela de Ciencias de la Salud, Universidad Viña del Mar, Viña del Mar, Chile.
ORCID ID:
[§] https://orcid.org/0000-0002-8370-4550
[¶] https://orcid.org/0000-0002-7257-1713
Recibido: 15-noviembre-2022. Aceptado: 27-diciembre-2022.

* Autora para correspondencia: Gabriela Paz Urrejola Contreras. Aguasanta 7055, Viña del Mar, Código postal: 2520000. Teléfono: (56) 32 2462400.
Correo electrónico: gabriela.urrejola@uvm.cl

en el que, a menor índice de facilidad, existe una mayor consistencia interna, representado por la mayor eficiencia discriminativa de las preguntas. El proceso de diseño y formulación de los instrumentos debe cuidar y examinar estas pautas para resguardar criterios de calidad.

Palabras clave: *Educación médica; rendimiento académico; indicadores de calidad; evaluación del aprendizaje; preguntas selección múltiple.*

Analysis of evaluative instruments used in science during the pandemic: multiple selection, indicators and performance

Abstract

Introduction: Multiple-choice assessments are the instrument widely used in science to assess students, however, the recent pandemic required adapting this type of instrument to the virtual environment. This context required evaluating the quality of the instruments through discrimination indices, internal consistency and relating it to academic performance.

Objective: Evaluate the evaluation instruments used in online mode during the COVID-19 pandemic and the performance of students in health sciences.

Method: The revision of the 5 instruments of the Structure and Function subject formed by 290 banks of random questions was carried out to evaluate each content in first-year students during 2020 in the school of health sciences at the Viña del Mar University. The data obtained from the virtual platform and the indices of discrimination, facility, discriminative efficiency, internal consistency and academic performance were interpreted through a report that was shared with the teachers to identify the parameters of quality and validity.

Results: Of the total number of question banks evaluated, 70.2% of the questions presented adequate discrimination and only 5.6% should be eliminated. Contest two obtained the lowest average performance 3.9 ± 0.99, however, it presented the highest internal consistency 81%. When comparing all the instruments, a gradual improvement in the formulation was observed, reflected in the final exam, in which the academic performance also agrees with the average of the semester 4.2 ± 0.92.

Conclusions: Academic performance must be weighed in relation to the quality of the formulated instrument, in which, at a lower ease index, there is greater internal consistency, represented by the greater discriminative efficiency of the questions. The design and formulation process must take care of and examine these guidelines to safeguard quality criteria.

Keywords: *Medical education; academic performance; quality control; learning assessment; multiple choice questions.*

INTRODUCCIÓN

La enseñanza y el aprendizaje en ciencias de la salud mediante entornos virtuales forma parte de los cambios que se han promovido de acuerdo con los nuevos paradigmas educativos[1,2]. En este proceso, la incorporación de nuevas estrategias y metodologías que han utilizado tanto educadores como estudiantes han permitido construir modelos híbridos que asocian el uso de herramientas sincrónicas como asincrónicas en dimensiones formativas y de evaluación[3,4].

Sin embargo, la reciente pandemia mundial por COVID-19 forzó y exigió migrar a escenarios virtuales utilizando nuevas herramientas de aprendizaje[5] y adoptar diferentes instrumentos evaluativos[6]. En este sentido, una primera etapa del desafío fue capacitar a los docentes en la adaptación de los instrumentos de evaluación mediante los recursos que disponen las plataformas virtuales que ampliamente se han utilizado en educación como Moodle, Collaborate Blackboard y/o Proctortrack[7,8].

En una segunda etapa se ha ponderado revisar el rendimiento de los estudiantes y analizar las variables que intervienen en ello[9,10]. Autores mencionan que tanto docentes como estudiantes reconocen fac-

tores facilitadores como el uso de métodos digitales ágiles y la flexibilidad temporal en la revisión de recursos asincrónicos, que intervienen en la consolidación del *e-learning* en ciencias de la salud[11,12].

En cuanto a las dificultades y barreras en el rendimiento de los estudiantes, los equipos docentes han detectado las falencias y han diseñado estrategias de mitigación y apoyo como el uso de cursos en línea masivos y abiertos (MOOC)[13,14], disponibilidad de instrumentos formativos de ejercitación previa, y herramientas asincrónicas que el estudiante realiza en forma complementaria para apoyar su proceso de aprendizaje1[5].

Por otro lado, la observación del rendimiento obtenido en evaluaciones mediante plataforma virtuales se ha puesto en foco revisar dos aspectos; uno de ellos relacionado con la incorporación de estrategias que limiten la copia y el plagio online[16,17]; y por otro lado, necesidad de examinar la validez, y la rigurosidad en los instrumentos de evaluación[18,19]. En este sentido, Carrillo y colaboradores (2020) han sido enfáticos en alertar y describir sobre las amenazas en la validez de los instrumentos de evaluación formados con preguntas de selección, en los que se presencian vicios como la formulación en negación, la elaboración incorrecta de los enunciados en el uso del lenguaje, el sesgo, las discordancias de dominio, y el *testwiseness*, con el fin de mejorar la calidad de las preguntas que apunten a ser realmente efectivas[20].

Otras conclusiones en esta temática sugieren revisar en profundidad la validez y la calidad de los instrumentos de evaluación basados en la selección múltiple mediante índice de dificultad, de discriminación, y coeficiente de consistencia interna, que en otros estudios han resultado tener diferencias significativas en la comparación de diferentes instrumentos[21].

Considerando estos indicadores, el índice de dificultad se relaciona con el índice de facilidad (IF) y busca pesquisar las diferencias entre los niveles de conocimiento de los estudiantes para los valores que oscilan entre un 30% y 70%. A modo de ejemplo, un instrumento con IF entre 21% y 34% puede considerarse difícil, mientras que un IF entre 35% y 65% es correcto para el estudiante promedio, y un IF entre 66% y 80% podría interpretarse como relativamente fácil.

Por otra parte, el índice de discriminación (ID) permite establecer la correlación entre las preguntas que dada su habilidad discriminativa es eficiente para separar a los estudiantes con mayor y menor capacidad.

Por tanto, la consistencia interna de un instrumento de evaluación denota el correcto uso de preguntas para diferenciar los grados de conocimiento en torno a una temática y la diferencia de habilidades que presentan los estudiantes. Valores cercanos a un 75% son considerados satisfactorios, mientras que valores inferiores a 64% sugieren la necesidad de someter a revisión y corregir la confección del instrumento, ya que las puntuaciones obtenidas podrían deberse al azar, y/o no necesariamente estar aportando la información requerida a partir del aprendizaje y desempeño del estudiante.

Por último, las experiencias en torno a la evaluación virtual en ciencias de la salud han sido enfáticas en sugerir alcances que debiesen profundizarse en materias como: mejorar la gestión y producción de instrumentos de evaluación[22], diseñar e incluir procesos de evaluación que cautelen la seguridad, y confiabilidad del instrumento a través de bancos de preguntas prescindiendo únicamente de la honestidad del estudiante, y sistematizar la instrumentación en los diferentes escenarios de enseñanza aprendizaje en ciencias de la salud; en instrumentos de selección múltiple, destinados a evaluar dominios de primer saber, ha sido requerido incluir análisis estadísticos para determinar la consistencia interna y validez[23].

OBJETIVO

Evaluar los instrumentos de evaluación empleados en la asignatura de Estructura y Función entre marzo y julio 2020, dictados en modalidad online durante la pandemia por COVID-19, y el rendimiento de los estudiantes.

MÉTODO

Este estudio corresponde a un trabajo cuantitativo, descriptivo, no experimental, transversal y retrospectivo que consideró el análisis de los instrumentos de evaluación utilizados durante el primer semestre en la asignatura de Estructura y Función, asignatura que revisa en forma integrada la anatomía y fisiología asociada a los sistemas del cuerpo humano, para

las carreras de obstetricia, nutrición, fonoaudiología, kinesiología, terapia ocupacional y enfermería en la Universidad Viña del Mar, Chile. Una vez terminado el semestre, el análisis incluyó la valoración mediante estadística descriptiva de la información obtenida a partir de los instrumentos de evaluación rendidos. Se consideró el índice de facilidad (IF), índice de discriminación (ID), eficiencia discriminativa (ED), coeficiente de consistencia interna, número de intentos y rendimiento. Para tales efectos se consideraron los valores de ID a partir del trabajo de Backhoff et al., que resume el poder de discriminación de las preguntas examinadas según su valor, además incluye una categoría cualitativa y recomendaciones sugeridas. Un ID >0.39 corresponde a una pregunta excelente por tanto debiera conservarse. A su vez, un valor entre 0.30 y 0.39 se considera como un reactivo bueno, pero con posibilidades de mejorar. Un reactivo que obtenga un valor entre 0.20 y 0.29 corresponde una pregunta de calidad regular que requiere ser revisada. Valores entre 0.00 y 0.20 son considerados como pregunta pobre que debe ser descartada o revisada en profundidad. Por último, valores negativos corresponden a preguntas de pésima calidad que se sugiere sean eliminadas definitivamente.

Se realizó el análisis de las evaluaciones aplicadas en la asignatura de Estructura y Función, correspondiente a los certámenes I, II, III, IV y el examen formulado mediante un banco de preguntas de selección única. En cuanto a la confección y formulación de los instrumentos de evaluación, tanto los certámenes y el examen disponen de la misma estructura y principalmente las diferencias entre los certámenes y el examen radica en el número de preguntas que componen el banco y el instrumento de evaluación.

En cuanto al diseño de los instrumentos de evaluación, los certámenes I, II, III y IV se configuraron mediante un banco de 290 preguntas. El banco de preguntas, a su vez, se subdivide en 35 bloques que albergan subtemas de contenidos específicos de cada unidad. Cada bloque agrupa entre 8 y 10 preguntas, de las cuales se selecciona entre 1 y 2 preguntas aleatorias para la formulación de cada certamen rendido individualmente por los estudiantes. Para los certámenes I, III, IV y examen, la complejidad de las preguntas se diseñó de acuerdo con la taxonomía de Bloom; un 60% de las preguntas de cada bloque aleatorio correspondió a niveles bajos de complejidad centrados en preguntas de reconocimiento o recuerdo de información, un 30% de las preguntas incluyó procesos mentales ligados a la asociación/comparación de conceptos claves, un 10% de las preguntas involucraron una complejidad mayor ligada a procesos de interpretación y aplicación. El certamen II siguió esta distribución de preguntas: 50% baja complejidad, 30% dificultad media, 20% mayor dificultad. Cada certamen estuvo compuesto por 40 preguntas y un puntaje total máximo de 40 puntos **(figura 1)**.

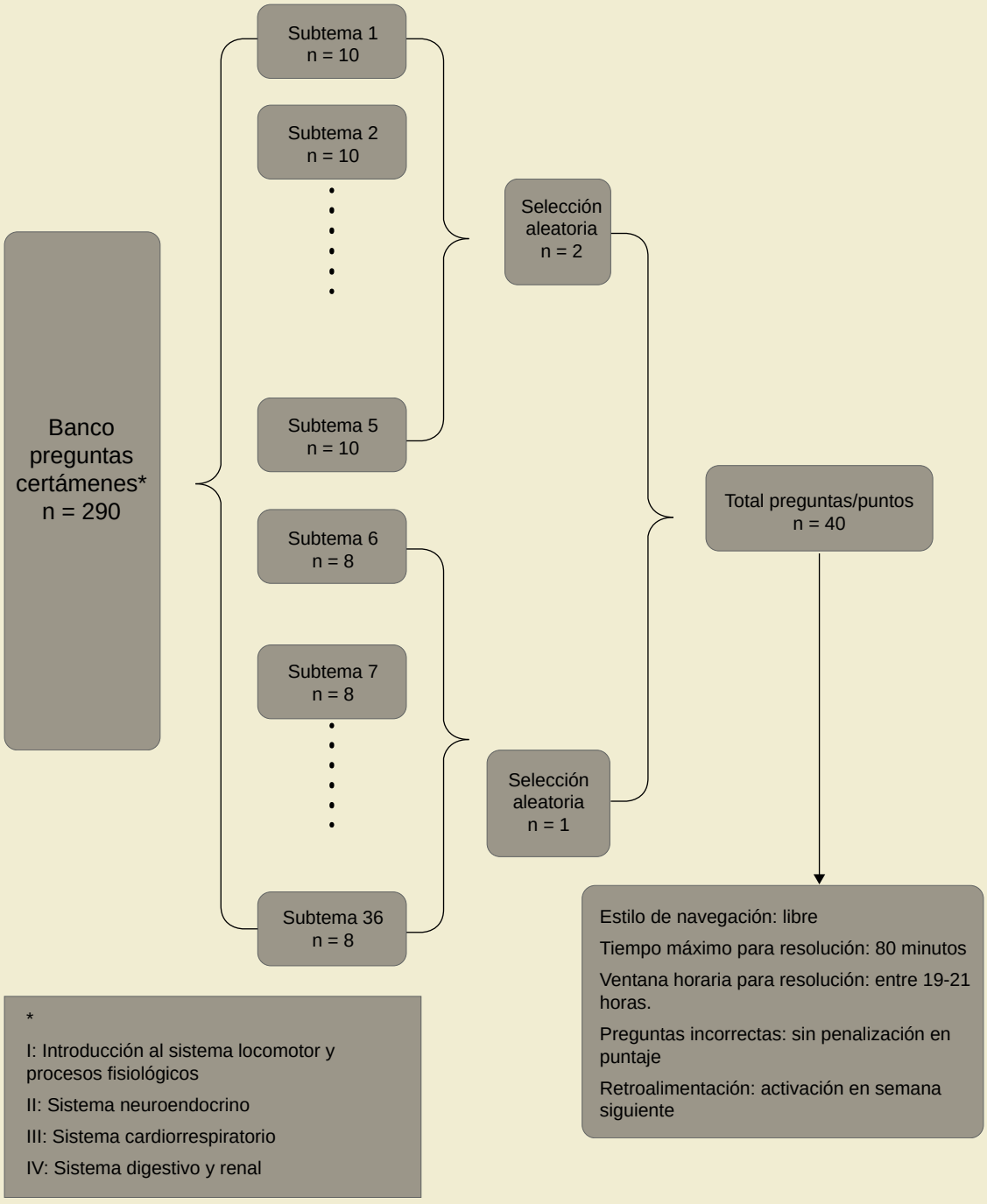
El examen incluyó la evaluación acumulativa de todos los contenidos revisados durante el semestre y consideró un banco de 370 preguntas. Este banco se compone de 42 bloques que agrupan entre 8 y 10 preguntas, de las cuales se seleccionó entre 1 y 2 preguntas aleatorias para la formulación del examen rendido al final del semestre individualmente por cada estudiante que presentara un promedio de calificaciones <5.5 o alguna calificación reprobatoria <4.0. Para todos los efectos se consideró una escala de calificaciones de 1.0 a 7.0. Cada certamen estuvo compuesto por 50 preguntas y un puntaje total máximo de 50 puntos.

Se utilizaron las siguientes consideraciones para la aplicación de los instrumentos en formato online mediante la plataforma virtual: solo un intento por estudiante, estilo de navegación libre (el estudiante puede devolver las preguntas si lo desea), tiempo máximo para resolución del certamen y examen de 80 y 90 minutos respectivamente, ventana horaria para rendir la evaluación entre las 19 y 21 horas, sin penalización del puntaje por respuestas incorrectas y la activación de la retroalimentación de las preguntas se realizó a la semana siguiente en horario designado por el docente.

Análisis estadístico

Todos los datos fueron analizados mediante el programa GraphPad Prism 8.01. Los datos se muestran como porcentajes descriptivos y estadísticos de tendencia central. El rendimiento académico obtenido por cada certamen y el coeficiente de consistencia fueron analizados utilizando ANOVA de una vía. Los posibles cambios en el índice de facilidad, índice de discriminación y eficiencia discriminativa fue-

Figura 1. Esquema de formulación de instrumentos de evaluación



El diagrama representa el proceso de confección utilizado para la creación de los certámenes I, II, III, IV y examen. Certámenes y examen presentan el mismo proceso para su formulación y las diferencias entre ellos, solo recae en el número de preguntas del banco y totales del instrumento de evaluación.

Tabla 1. Caracterización instrumentos de evaluación

Evaluación	Cert I	Cert II	Cert III	Cert IV	Examen
Nº intentos rendidos	220	220	205	189	162
Subtemas (bloques)	35	35	35	35	42
Nº preguntas banco	290	290	290	290	370
Calificación Promedio	4.6 ± 0.85	3.9 ± 0.99	4.1 ± 0.83	4.7 ± 0.89	4.2 ± 0.92
Índice de dificultad	64%	54%	57%	68%	59%
Coeficiente consistencia interna	68%	81%	71%	73%	73%
Índice Facilidad (IF) de las preguntas (%)					
Altamente fácil	14	4	4	11	5
Medianamente fácil	6	7	7	19	7
Dificultad media	23	13	59	54	58
Medianamente difícil	38	46	16	9	13
Altamente difícil	19	30	14	7	17
Índice Discriminación (ID) de las preguntas (%)					
Revisar y reformular con ajuste menor	54	35	33	39	34
Sugeridas de eliminar	2	2	4	6	14
Conservar	44	63	63	55	52
Eficiencia discriminativa (ED) de las preguntas (%)					
Débil discriminación	35	17	25	23	21
Inválidas	2	2	4	6	14
Adecuada discriminación	63	81	71	71	65
Rendimiento académico global	**Obstetricia Fonoaudiología**		**Kinesiología Terapia Ocupacional Nutrición**		**Enfermería**
Aprobación ≥4.0	86%		78%		83%
Reprobación <4.0	14%		22%		17%

ron analizados utilizando ANOVA de mediciones repetidas.

Consideraciones éticas

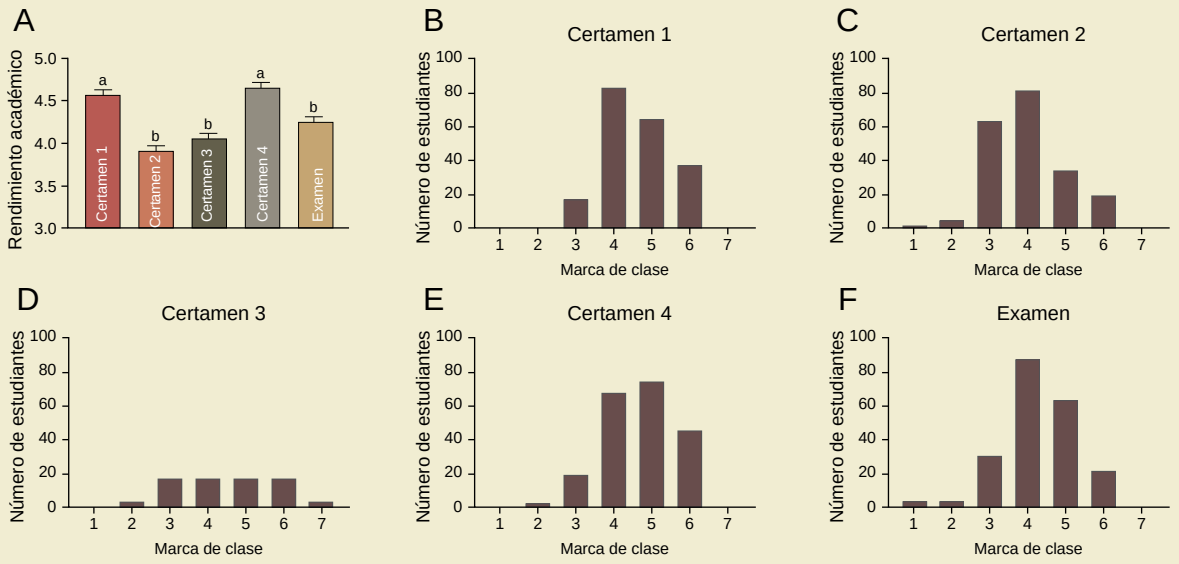
Este trabajo siguió el protocolo ético de la institución a través de los lineamientos establecidos por el comité de ética para el resguardo de los datos y no contempló el uso de consentimiento informado. El uso de estos solo tendrá fines investigativos.

RESULTADOS

En la **tabla 1** se resumen los antecedentes por cada certamen y/o examen. Se observa una disminución en el número de intentos rendidos entre el certamen 1 (220 intentos) al certamen 4 (189), para finalizar solo con 162 intentos en el examen. Reducción debida al retiro de los alumnos de la asignatura. Los promedios de los certámenes indican que las mejores calificaciones promedio fueron del certamen I y IV, siendo estas un 4.6 ± 0.85 y 4.7 ± 0.89 respectivamente. Ambas calificaciones promedio se consideraron aprobatorias. Por el contrario, el rendimiento más bajo fue observado en el certamen II, el que correspondió a un 3.9 ± 0.99, considerándose esta calificación como reprobatoria.

En general, el índice de dificultad es estable en todos los instrumentos, a excepción del certamen IV que es considerado ligeramente más fácil.

Figura 2. Rendimiento académico y distribución de frecuencia



a) Se encontraron diferencias significativas entre el rendimiento académico entre los instrumentos de evaluación ($p < 0.0001$). El certamen 1 y 4 obtuvieron un rendimiento académico significativo más alto en comparación con el certamen 2, 3 y examen. **b-f)** En todos los instrumentos de evaluación, el 40.7% en promedio de los alumnos concentra sus notas entre el 3.5 al 4.5 (marca de clase 4), mientras que el 17% de los alumnos concentra notas entre el 2.5 al 3.5 (marca de clase 3). El 27.8% de los alumnos concentra sus notas entre el 4,5 al 5,5 (marca de clase 5). Se destaca que el 12.8% de los alumnos obtiene notas entre 5.5 al 6.5 (marca de clase 6). Los datos representan el promedio ± SEM. Letras diferentes, indican diferencias significativas.

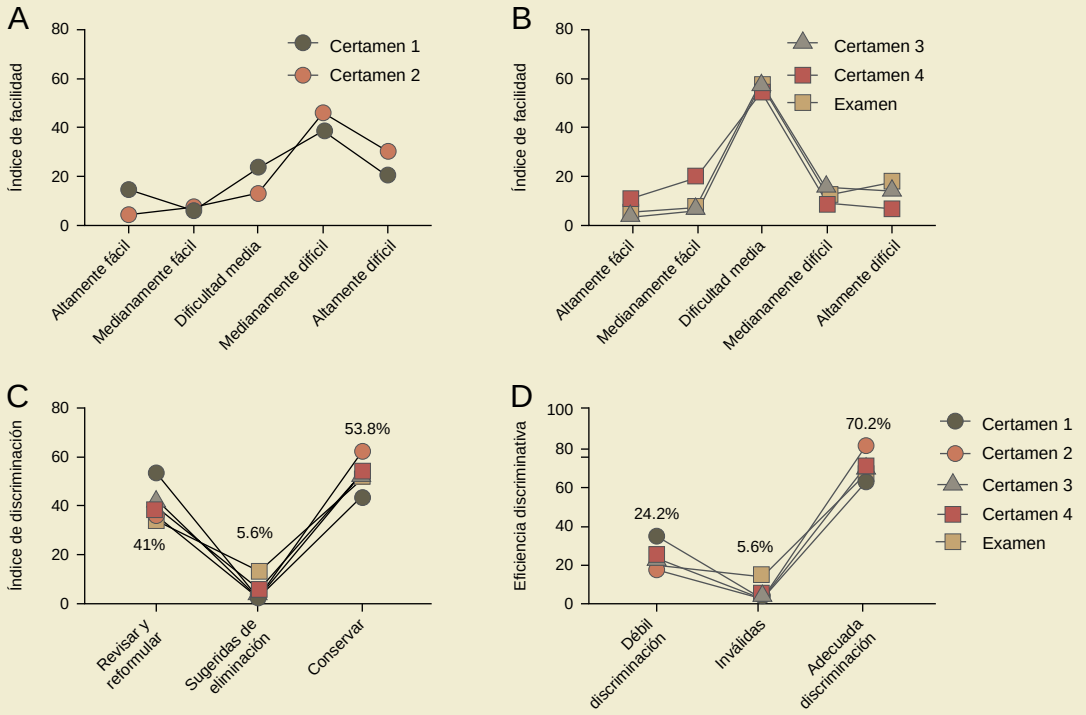
Al revisar la consistencia interna, cabe destacar que el mejor desempeño lo obtuvo el certamen II con un 81%, seguido del certamen IV y el examen final, ambos con un 73%. Según los valores de consistencia interna obtenidos, todos los instrumentos fueron considerados aceptables y satisfactorios; sin embargo, en el certamen I debieran revisarse, ya que obtuvo el porcentaje más bajo (68%).

Precisando el análisis a partir de las preguntas que componen los bancos de preguntas cada instrumento de evaluación, la efectividad de una pregunta para clasificar a los estudiantes más y menos capaces fue adecuada para la mayoría de las preguntas examinadas. Esto correspondió para un 52% a 63% de estas en los certámenes II, III, IV y examen final. En esta dimensión, para el certamen I, obtuvo un 44%, por tanto, se sugiere la revisión y/o el ajuste menor de un 54% de las preguntas.

Al contrastar con la eficiencia discriminativa, es decir qué tan buena es la discriminación en relación con la dificultad de las preguntas, se consideró adecuada en la mayoría de estas (63%-81%) **(tabla 1)**.

El rendimiento académico en la asignatura de Estructura y Función durante un entorno virtual se observa en la **figura 2a**. El análisis de ANOVA de una vía, demuestra que existen diferencias significativas entre los certámenes ($F_{(4,1009)} = 25.64$, $p < 0.0001$). El certamen 1 y 4 (certamen 1: 4.57 ± 0.05; certamen 4: 4.6 ± 0.06) obtuvieron un rendimiento académico significativo más alto en comparación con el certamen 2, 3 y examen (certamen 2: 3.9 ± 0,07; certamen 3: 4.0 ± 0.06; examen: 4.2 ± 0.06). A partir de las notas de cada instrumento de evaluación se han generado distribuciones de frecuencias de los rendimientos académicos **(figura 2b-f)**. En todos los instrumentos de evaluación, el 40.7% de los alumnos concentran sus notas entre el 3.5 al 4.5, mientras que el 17% de los alumnos concentran sus notas entre el 2.5 al 3.5. El 27.8% de los alumnos concentran sus notas entre el 4.5 al 5.5. Se destaca que el 12.8% de los alumnos obtienen notas entre 5.5 al 6.5 **(figura 2b-f)**.

Figura 3. Análisis del instrumento según índices de facilidad, discriminación o eficiencia discriminativa



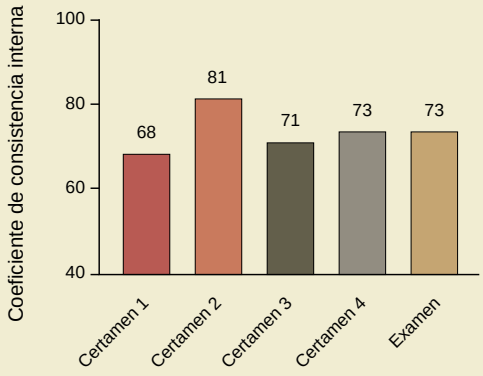
a-b) Índice de facilidad. Se agruparon los instrumentos de evaluación de acuerdo con sus similitudes, porque se pudieron establecer 2 grupos. Un grupo integra a los certámenes 1 y 2 **(a)**, mientras que el otro grupo integra a los certámenes 3, 4 y examen **(b)**. **c) Índice de discriminación.** Indica qué tan efectivas son las preguntas para clasificar/separar/discernir a los estudiantes que obtuvieron un puntaje alto de los que obtuvieron un puntaje bajo. Las preguntas en todos los certámenes poseen en promedio un 53.8% la capacidad de discriminar. El 41% de las preguntas requiere un ajuste y/o edición para mejorar su capacidad de discriminación, mientras que tan solo el 5% debieran eliminarse. Los datos representan el promedio.

El índice de facilidad o proporción de aciertos entrega información con respecto a las diferencias entre el nivel de conocimiento y preparación de los individuos. En la **figura 3a y 3b** se agruparon los certámenes de acuerdo con las características del índice de facilidad. En la **figura 3a**, se muestra el índice de facilidad del certamen 1 y 2, que presentan principalmente preguntas de dificultad media, medianamente difícil y altamente difícil, mientras que en la **figura 3b** se muestran el índice de dificultad de los certámenes 3, 4 y examen, los cuales presentan principalmente preguntas de dificultad media **(figura 3b)**.

El índice de discriminación de una pregunta corresponde a la diferencia entre las proporciones de aciertos entre los que obtuvieron un puntaje total de la prueba alto, de los que obtuvieron un puntaje bajo. Por lo tanto, expresa hasta qué punto la pregunta discrimina y establece diferencias entre los estudiantes de más alto y bajo rendimiento. En la **figura 3c** se observa que el 53.8% promedio de todos los instrumentos de evaluación permite una buena discriminación, mientras que un 41% de las preguntas se sugiere revisar y/o reformular para aumentar la discriminación, y solo el 5.6% de las preguntas en promedio se sugiere eliminar por nula discriminación.

La eficiencia discriminativa estima qué tan bueno es el índice de discriminación en relación con la dificultad de la pregunta. El 70% de las preguntas de todos los certámenes poseen una adecuada eficiencia discriminativa, el 24.2% de las preguntas de todos los certámenes poseen una débil discriminación, y solo el 5.6% de las preguntas presentan nula eficiencia discriminativa y se sugiere su eliminación **(figura 3d)**.

Figura 4. Coeficiente de consistencia interna



Los valores oscilan entre el 68% y 81%. Con un promedio del 73%. Los certámenes 2, 3, 4 y examen pueden ser clasificados como un instrumento satisfactorio, cuyas preguntas permiten discriminar entre los estudiantes de más alto y bajo rendimiento. El certamen 1 posee un coeficiente de consistencia interna superior al valor de corte. Sin embargo, se sugiere realizar medidas correctivas para incrementar su capacidad de discriminación.

En la **figura 4**, se describe el coeficiente de consistencia interna para todos los certámenes y examen, con valores que fluctúan entre el 68% y el 81%. Valores inferiores al 64% del coeficiente de consistencia interna indican que el instrumento completo es insatisfactorio y se deberían realizar medidas correctivas. El promedio del coeficiente de consistencia interna para todos los instrumentos de evaluación fue del 73%, por lo que corresponden a instrumentos satisfactorios, cuyas preguntas permiten discriminar entre los estudiantes de más alto rendimiento, de los de más bajo rendimiento, por lo que las puntuaciones totales obtenidas no se deben al azar **(figura 4)**.

DISCUSIÓN

Considerando el contexto de educación en línea en pandemia, el principal criterio ha sido examinar el rendimiento, ya que globalmente se vio incrementado en diferentes disciplinas en ciencias de la salud[24] por las ayudas que en línea los estudiantes logran disponer como apuntes, búsqueda simultánea y/o comunicación por redes sociales entre los estudiantes de mejor y menor desempeño, y plagio entre otros; sin embargo, estudios advirtieron que la educación en línea es menos efectiva que las clases en línea. Otro alcance reportado ante la mejora del rendimiento fue debatir que los procesos de educación en línea presentaban ineficiencia para mostrar integridad y rigurosidad académica[25].

En su contraparte, otros estudios revelaron que la mejora del rendimiento se debió al estudio continuo mediante recursos sincrónicos[26]; sin embargo, al comparar la preferencia del estudiantado entre la interacción sincrónica y asincrónica, esta última tuvo más adherencia por disponibilidad y comodidad en la flexibilidad que otorga respecto a los horarios destinados a revisar los contenidos[27].

En relación con lo anterior, este trabajo observó performance aprobatoria homogénea en las carreras, pero no registró un incremento en comparación a la modalidad presencial impartida previo a la pandemia. El análisis de los instrumentos aplicados permitió debatir de las fortalezas de este como la disponibilidad de extensos bancos de preguntas y la selección aleatoria de preguntas, lo que permitió que cada prueba/examen fuera prácticamente única y diferente para cada estudiante de la muestra estudiada.

Los hallazgos de este estudio indican que existe una relación entre el índice de facilidad y el índice de discriminación con el rendimiento obtenido en cada uno de los certámenes. Por ejemplo, al revisar el índice de facilidad, el certamen dos es el único instrumento que incluye una mayor proporción de preguntas medianamente difíciles y aunque esto pareció impactar en el menor rendimiento obtenido, el certamen dos es el que presentó la mayor consistencia interna de todos los instrumentos evaluados[28]. Esta condición observada permite la comprensión de que la aprobación o resultado general de un instrumento en función del rendimiento, no necesariamente denota validez, confiabilidad y calidad[29].

Por otra parte, la revisión en detalle del certamen dos, permitió observar que las preguntas que implicaban interpretación y aplicación en torno a un caso y/o problema expresado en la pregunta tuvo mayor complejidad para los estudiantes, ya que es requerida una comprensión más acabada de los contenidos, como también de la información entregada en el enunciado (encabezado) de la pregunta[30].

Al revisar todos los instrumentos, si bien el rendimiento global es homogéneo para todas las carreras, se observaron diferencias significativas en

el rendimiento entre los certámenes 1 y 4, que obtuvieron una mayor calificación promedio, respecto a los certámenes 2, 3 y examen. Al respecto, el examen reúne acumulativamente todos los contenidos tratados en el semestre, pero a partir de bancos de preguntas más extensos (ochenta preguntas más), que al analizarlos en detalle denotó una mejora gradual en la construcción del instrumento concordante con el informe posterior a ser aplicado: un menor número de preguntas presentaron una débil discriminación y, por tanto, requieren un ajuste y/o corrección menor. Además, esto se reflejó en la distribución de calificaciones obtenidas por los estudiantes que rindieron el instrumento, ya que se observó una menor dispersión hacia calificaciones bajas o altas; por el contrario, se concentraron las calificaciones cercanas al 4.0. Lo anterior sugiere que los docentes que formularon los instrumentos comprendieron los alcances relevantes entregados posteriormente a cada instrumento, y el proceso de mejora en la formulación de estos fue guiada, y reveló un perfeccionamiento progresivo, similar a como indican algunos autores en sus estudios[31].

En cuanto a las limitaciones de este trabajo se encuentran el disponer de los resultados y la retroalimentación de los indicadores de las preguntas posteriormente a la aplicación del instrumento[32], criterio que no pudo ser subsanado; sin embargo, una fortaleza es la posibilidad de mejorar la evaluación sistemáticamente a partir del extenso banco de preguntas examinado y del conocimiento sobre las preguntas que discriminan eficientemente frente a aquellas que presentan debilidad en este ítem[33,34].

Las oportunidades de mejora de este trabajo se resumen principalmente en la obtención de información crítica a partir de los indicadores. El análisis sugiere que una de cada cuatro preguntas requiere revisión y reformulación debido a una débil discriminación. Realizar esta adecuación permite comprender el error en el diseño y perfeccionar su formulación en el banco de preguntas, lo que constituye una instancia de aprendizaje para los docentes y estudiantes[35]. En este sentido, el informe obtenido a partir de cada instrumento después de ser aplicado se socializó con los docentes, así como también las sugerencias y directrices generales para mejorar la construcción de los instrumentos siguientes.

CONCLUSIONES

Considerando que los procesos evaluativos constituyen un recurso educativo y de aprendizaje a la vez, los instrumentos revisados demostraron consistencia, validez y pertinencia. El análisis profundo de los instrumentos denotó globalmente ser adecuado, conforme a la complejidad de las preguntas utilizadas en los bancos, y la eficiencia discriminativa de estas sobre los contenidos tratados para cada unidad temática similar a otros estudios en esta disciplina. Sin perjuicio de lo anterior, la calidad de los instrumentos puede ser superior, si la compresión del análisis es internalizada por los docentes que imparten la asignatura, y se dispone de docentes con competencias y habilidades específicas en educación médica que lideren y guíen el proceso de mejora[36,37].

Se sugiere y recomienda a las unidades y/o cuerpos académicos que se encuentran en procesos de transición e implementación de educación online, velen por resguardar una evaluación auténtica; la rigurosidad del proceso metodológico que permite el diseño de un instrumento de evaluación y que este sea pertinente y concordante en su capacidad discriminatoria y a la vez permita mitigar el sesgo en el incremento del rendimiento que no concuerda con un real aprendizaje.

CONTRIBUCIÓN INDIVIDUAL

- GU: Concepción y diseño de la investigación, recolección de datos, escritura del artículo, revisión crítica y aprobación de la versión final del manuscrito.
- MP: Redacción de los resultados, análisis de datos y formulación de la metodología, revisión crítica y aprobación de la versión final del manuscrito.

AGRADECIMIENTOS

Este estudio fue apoyado por la Unidad de Ciencias Aplicadas.

PRESENTACIONES PREVIAS

Ninguna.

FINANCIAMIENTO

Ninguno.

CONFLICTO DE INTERESES

Ninguno.

REFERENCIAS

1. Voutilainen A, Saaranen T, Sormunen M. Conventional vs. e-learning in nursing education: A systematic review and meta-analysis. Nurse Educ Today. 2017;50:97-103. doi:10.1016/j.nedt.2016.12.020
2. Lawn S, Zhi X, Morello A. An integrative review of e-learning in the delivery of self-management support training for health professionals. BMC Med Educ. 2017;17(1):183. doi: 10.1186/s12909-017-1022-0
3. Villaroel Quinchalef G del P, Fuentes Salvo M de los Á, Oyarzún Muñoz VH. Implementación de curso online de Anatomía y la percepción de los estudiantes de Kinesiología. Inv Ed Med. 2020;(35):75-84. doi: 10.22201/facmed.20075057e.2020.35.20226
4. Lisperguer S, Calvo M, Urrejola G, Pérez M. Clinical reasoning training based on the analysis of clinical case using a virtual environment. Educ Med. 2020;594:1-5. doi: 10.1016/j.edumed.2020.08.002
5. Goh PS, Sandars J. A vision of the use of technology in medical education after the COVID-19 pandemic. MedEdPublish. 2020;9:49. doi: 10.15694/mep.2020.000049.1
6. Gaur U, Majumder MAA, Sa B, Sarkar S, Williams A, Singh K. Challenges and Opportunities of Preclinical Medical Education: COVID-19 Crisis and Beyond. SN Compr Clin Med. noviembre de 2020;2(11):1992-7. doi: 10.1016/j.glt.2021.11.001
7. Arandjelovic A, Arandjelovic K, Dwyer K, Shaw C. COVID-19: Considerations for Medical Education during a Pandemic. MedEdPublish. 2020;9:87. doi:10.15694/mep.2020.000087.1
8. Medical Education Department, School of Medical Sciences, Health Campus, Universiti Sains Malaysia, 16150 Kubang Kerian, Kelantan, Malaysia, Abdul Rahim AF. Guidelines for Online Assessment in Emergency Remote Teaching during the COVID-19 Pandemic. Educ Med J. 30 de junio de 2020;12(2):59-68. doi: 10.52494/UCML9733
9. Álvarez-Vázquez M, Álvarez-Méndez AM, Bravo-Llatas C, Angulo-Carrere MT. Análisis multivariante del uso de espacios virtualizados por estudiantes pregraduados en ciencias de la salud. 2021;24(6):317-21. doi: 10.33588/fem.246.1159
10. Bautista-Rodríguez G, Gatica-Lara F. Factores relacionados con el rendimiento académico en una carrera técnica en salud impartida en línea. Inv Ed Med. 2020;(33):89-97. doi: 10.22201/facmed.20075057e.2020.33.19177
11. Regmi K, Jones L. A systematic review of the factors – enablers and barriers– affecting e-learning in health sciences education. BMC Med Educ. 2020;20(1):91. doi: 10.1186/s12909-020-02007-6
12. Alsoufi A, Alsuyihili A, Msherghi A, Elhadi A, Atiyah H, Ashini A, et al. Impact of the COVID-19 pandemic on medical education: Medical students' knowledge, attitudes, and practices regarding electronic learning. Plos One. 2020;15(11):e0242905. doi: 10.1371/journal.pone.0242905
13. Padilha JM, Machado PP, Ribeiro AL, Ribeiro R, Vieira F, Costa P. Easiness, usefulness and intention to use a MOOC in nursing. Nurse Educ Today. 2021;97:104705. doi: 10.1016/j.nedt.2020.104705
14. Yilmaz Y, Sarikaya O, Senol Y, Baykan Z, Karaca O, Demiral Yilmaz N, et al. RE-AIMing COVID-19 online learning for medical students: a massive open online course evaluation. BMC Med Educ. 2021;21(1):303. doi: 10.1186/s12909-021-02751-3
15. Logan RM, Johnson CE, Worsham JW. Development of an e-learning module to facilitate student learning and outcomes. Teach Learn Nurs. 2021;16(2):139-42. doi: 10.1016/j.teln.2020.10.007
16. Heidarzadeh A, Zehtab Hashemi H, Parvasideh P, Hasan Larijani Z, Baghdadi P, Fakhraee M, et al. Opportunities and Challenges of Online Take-Home Exams in Medical Education. J Med Educ; 2021;20(1). doi: 10.5812/jme.112512
17. Justo-Cousiño LA. ¿Podemos evaluar con garantías durante la pandemia de COVID-19? Evaluar sin devaluar las profesiones sanitarias. FEM. 2020;23(4):229. doi: 10.33588/fem.234.1075
18. Urrejola-Contreras GP, Tiscornia-González C. Retroalimentación estudiantil sobre herramientas sincrónicas y asincrónicas empleadas en ciencias de la salud en la pandemia por COVID-19. FEM. 2022;25(1):39. doi: 10.33588/fem.251.1168
19. Barteit S, Guzek D, Jahn A, Bärnighausen T, Jorge MM, Neuhann F. Evaluation of e-learning for medical education in low- and middle-income countries: A systematic review. Comput Educ. 2020;145:103726. doi: 10.1016/j.compedu.2019.103726
20. Carrillo-Avalos BA, Sánchez Mendiola M, Leenen I. Amenazas a la validez en evaluación: implicaciones en educación médica. Inv Ed Med. 2020;(34):100-7. doi: 10.22201/facmed.20075057e.2020.34.221
21. Giaconi E, Bazán ME, Castillo M, Hurtado A, Rojas H, Giaconi V, et al. Análisis de pruebas de opción múltiple en carreras de la salud de la Universidad Mayor. Inv Ed Med. 2021;(40):61-9. doi: 10.22201/fm.20075057e.2021.40.21365
22. Núñez J. Educación médica durante la crisis por COVID-19. 2020. 21(3):157. doi: 10.1016/j.edumed.2020.05.001
23. Luna de la Luz V, González P. Transformaciones en educación médica: innovaciones en la evaluación de los aprendizajes y avances tecnológicos (parte 2). Inv Ed Med; 2020. 9(34):87-99. doi:10.22201/facmed.20075057e.2020.34.20220
24. Almahasees Z, Mohsen K, Omar Amin M. Faculty's and Students' Perceptions of Online Learning During COVID-19. Front Educ 2021. 6:638470. doi:10.3389/feduc.2021.63847
25. Mukhtar K, Javed K, Arooj M, Sethi A. Advantages, Limitations and Recommendations for online learning during COVID-19 pandemic era: Online learning during COVID-19 pandemic era. Pak J Med Sci. 2022;36:COVID19-S4. doi: 10.12669/pjms.36.COVID19-S4.2785
26. González T, De la Rubia M, Hincz K, Comas M, Subirats L, Fort S, et al. Influence of COVID-19 confinement on students' performance in higher education. Pak J Med Sci 2020;15(10):e0239490. doi: 10.12669/pjms.36.COVID19-S4.2785
27. Fabriz S, Mendzheritskaya J, Stehle S. Impact of Synchronous and Asynchronous Settings of Online Teaching and

Learning in Higher Education on Students' Learning Experience During COVID-19. Front Phsycol. 2021;12:733554. doi:10.3389/fpsyg.2021.733554
28. Coughlin PA, Featherstone CR. How to Write a High Quality Multiple Choice Question (MCQ): A Guide for Clinicians. Eur J Vasc Endovasc Surg. 2017;54(5):654-8. doi: 10.1016/j.ejvs.2017.07.012
29. Abhijeet I, Purushottam G, Mohan D. Study on item and test analysis of multiple choice questions amongst undergraduate medical students. Int J Community Med Public Health. 2017;4(5):1562-5. doi:10.18203/2394-6040.ijcmph20171764
30. Kolomitro K, MacKenzie LW, Lockridge M, Clohosey D. Problem-solving strategies used in anatomical multiple-choice questions. Health Sci Rep. 2020;3(4). doi:10.1002/hsr2.209
31. Douthit N, Norcini J, Mazuz K, Alkan M, Feuerstein M, Clarfield M, et al. Assessment of global health education: the role of multiple-choice questions. Int J Appl Basic Mes Res. 2021;9(640204). doi: 10.3389/fpubh.2021.640204.
32. Butler AC. Multiple-choice testing in education: Are the best practices for assessment also good for learning? J Appl Res Mem Cogn. 2018;7(3):323-31. doi: 10.1016/j.jarmac.2018.07.002
33. Burud I, Nagandla K, Agarwal P. Impact of distractors in item analysis of multiple choice questions. Int J Res Med Sci. 2019;7(4):1136. doi: 10.18203/2320-6012.ijrms20191313
34. Scott K, King A, Estes M, Conlon L, Phillips A. Evaluation of an Intervention to Improve Quality of Single-best Answer Multiple-choice Questions. West J Emerg Med. 2018;20(1):11-4. doi: 10.5811/westjem.2018.11.39805
35. Moore S, Nguyen HA, Stamper J. Examining the Effects of Student Participation and Performance on the Quality of Learnersourcing Multiple-Choice Questions. En: Proceedings of the Eighth ACM Conference on Learning @ Scale. Virtual Event Germany: ACM; 2021, p. 209-20. doi:10.1145/3430895.3460140
36. Gupta P, Meena P, Khan A, Malhotra R, Singh T. Effect of faculty training on quality of multiple-choice questions. Int J Appl Basic Med Res. 2020;10(3):210. doi: 10.4103/ijabmr.IJABMR_30_20
37. Przymuszała P, Piotrowska K, Lipski D, Marciniak R, Cerbin-Koczorowska M. Guidelines on Writing Multiple Choice Questions: A Well-Received and Effective Faculty Development Intervention. SAGE Open. 2020;10(3):215824402094743. doi: 10.1177/2158244020947432