Impacto de diferentes modalidades de evaluación en el promedio de notas de estudiantes de Medicina: un estudio de intervención no aleatorizado

Ignacio Álvarez Amuchástegui, María Candelaria Ramos, Camila Muslera, Florencia Gala Cullari, Lucía Victoria Campos Cervera, María Teresa Politi, Juan Guido Chiabrando, Julio César Rotondaro y Daniela Wojtowicz

RESUMEN

Introducción: en el proceso de enseñanza-aprendizaje existen múltiples conflictos al momento de seleccionar el tipo de evaluación que debería aplicarse a estudiantes de Medicina. Nuestro objetivo es comparar diferencias en la media de notas de tres modalidades de examen (oral, escrito para desarrollar y preguntas de opción múltiple) para así determinar cómo estas podrían afectar el desempeño de los estudiantes de Medicina en el campo de la Farmacología.

Material y métodos: estudio cuasi experimental con una intervención no aleatorizada en una muestra por conveniencia de estudiantes de Medicina. A fin de evaluar diferencias en la media de notas se hizo un análisis ANOVA para muestras pareadas y luego los correspondientes tests de T para muestras pareadas.

Resultados: enrolamos inicialmente a 36 estudiantes; 7 fueron excluidos (4 por ausencia y 3 por abandono), y se obtuvieron 29 participantes. La media de notas del examen oral y la de preguntas de opción múltiple fueron ambas significativamente superiores a la del examen escrito para desarrollar (oral vs. escrito: diferencia 1,8 puntos; IC 95% 0,8 a 2,7; $p < 0,01$; opción múltiple vs. escrito: diferencia 2,1 puntos; IC 95% 1,4 a 2,9; $p < 0,01$). No hubo diferencias estadísticamente significativas entre las notas medias del examen oral y del examen de preguntas de opción múltiple ($p = 0,37$).

Conclusión: los estudiantes de Medicina obtienen peores notas en el examen escrito para desarrollar en Farmacología, en relación con los exámenes oral y de preguntas de opción múltiple. Esto posiblemente se asocie al hecho de que aquella modalidad es menos frecuentemente empleada en la carrera de Medicina.

Palabras clave: estudiantes de Medicina, desempeño académico, habilidad para dar exámenes, farmacología.

IMPACT OF DIFFERENT TYPES OF EVALUATION ON THE AVERAGE GRADE OF MEDICINE STUDENTS: A NON-RANDOMIZED INTERVENTIONAL STUDY

ABSTRACT

Introduction: in the teaching-learning process, there are many problems in the selection of the most suitable type of exam for evaluating medical students. Our target was to compare differences in the average grade of medical students upon taking three different types of exam (oral, written, and multiple-choice questions) to determine how these different types of exam may affect the performance of medical students in the area of Pharmacology.

Material and methods: we conducted a quasi experimental study by applying a non-randomized intervention to a convenience sample of medical students. To evaluate differences in the average grades among three groups, an ANOVA analysis was applied followed by paired T-tests.

Results: we initially enrolled 36 students; 7 were excluded (4 were absent and 3 abandoned the intervention), arriving at a total sum of 29 participants. The average grades of the oral exam and multiple-choice questions were both significantly higher than the written exam (oral vs. written: difference 1.8 points; 95%CI 0.8 to 2.7, $p < 0.01$; multiple-choice vs. written: difference 2.1 points, 95%CI 1.4 to 2.9, $p < 0.01$). There were no significant differences between the average grades on the oral exam and the multiple-choice exam ($p = 0.37$).

Conclusion: medical students have worse grades on written exams in Pharmacology, as compared to oral and multiple-choice exams. This could possibly be associated with the fact that this type of exam is less frequently applied in Medical School.

Key words: medical students, academic performance, test taking skills, pharmacology.

Rev. Hosp. Ital. B.Aires 2019; 39(3): 86-93.

Recibido: 7/04/19 Aceptado: 22/11/19

II Cátedra de Farmacología y Toxicología. Facultad de Medicina. Universidad de Buenos Aires. Argentina.
Correspondencia: galacuya@gmail.com

INTRODUCCIÓN

En el proceso de enseñanza-aprendizaje, las evaluaciones de los estudiantes cumplen un papel crucial en la planificación de la cursada. Sin embargo, existen múltiples conflictos al momento de seleccionar el tipo de evaluación que debería aplicarse a estudiantes de Medicina. Tanto estudiantes como docentes suelen tener valoraciones y preconceptos establecidos acerca de cada una de las diferentes modalidades de examen[1]. En la literatura médica existen muchos estudios que intentan analizar las ventajas y desventajas de diferentes métodos de evaluación aplicados a diferentes campos de la Medicina, y frecuentemente arriban a resultados contradictorios y discrepantes entre sí y con las ideas previas establecidas en la tradición educativa[2-5]. En general, se acepta que la selección de una determinada modalidad de evaluación dependería no solo de la disciplina específica que se va a evaluar sino también de las características del sujeto por evaluar[1]. Ante la ausencia de un modo unívoco de determinar el nivel de comprensión de un sujeto con respecto a un tema, esos estudios se han centrado en comparar las distintas modalidades de evaluación entre sí, en la valoración de diferentes disciplinas. La dimensión más frecuentemente analizada con ese objetivo ha sido la puntuación de la nota final asignada o la valoración cualitativa respecto de aprobación o desaprobación que surge de ella. A pesar de la multiplicidad de estudios que abordan este tema, no hemos hallado en la literatura alguno que haya intentado comparar simultáneamente los resultados de las tres modalidades más frecuentemente empleadas: el examen oral, el examen escrito para desarrollar y el examen por preguntas de opción múltiple. A su vez, no se han comparado estas modalidades de examen entre sí como herramientas para evaluar conocimientos de Farmacología en estudiantes de Medicina.

El objetivo del siguiente estudio es analizar si el desempeño de los estudiantes de Medicina en un examen de Farmacología cambia significativamente según la modalidad de examen: oral, escrito para desarrollar o por preguntas de opción múltiple.

MATERIALES Y MÉTODOS

Diseño del estudio y características de la población

Estudio cuasi experimental con una intervención no aleatorizada en una muestra por conveniencia de estudiantes de Medicina de la Universidad de Buenos Aires (UBA). Fueron invitados a participar todos los estudiantes de Farmacología II de la cursada de los días martes de 17 a 21 horas de la II Cátedra de Farmacología, de la carrera de Medicina de la UBA del primer cuatrimestre de 2018. Esto se llevó a cabo mediante una convocatoria oral realizada por los investigadores inmediatamente antes del inicio de las clases teóricas y prácticas de dicha cursada. Se ofreció un *e-mail* de contacto de un investigador para comunicación de los estudiantes interesados. Estos recibían luego un *link* electrónico en sus casillas de correo que conducía a un formulario en una plataforma digital (*Google Forms*). En dicho formulario, los estudiantes debían dejar constancia de su consentimiento informado y completar sus datos sociodemográficos. Asimismo, se les preguntaba si eran ayudantes de alguna materia de la carrera de Medicina y sus predicciones acerca de cuál sería la modalidad de examen en la cual alcanzarían un mejor desempeño. Fueron excluidos del estudio los estudiantes que abandonaron su participación, se encontraban ausentes el día de la intervención o se encontraban copiando durante el examen. Participaron como investigadores del estudio docentes y miembros de la Escuela de Ayudantes de la cursada.

Características de la intervención

La intervención consistió en una evaluación oral, una evaluación escrita para desarrollar y una evaluación de preguntas de opción múltiple, realizadas de manera sucesiva en la misma población de participantes durante la misma jornada. Se llevó a cabo durante 2 horas el día 24 de abril de 2018 durante el horario de clases, como una alternativa a la actividad de repaso tradicional; esto tuvo lugar 2 semanas antes del examen parcial del módulo de Farmacología Cardiovascular. A los estudiantes que eligieron participar del estudio se les asignó una hoja con un número aleatorio de 2 dígitos que se mantuvo guardada durante toda la intervención (es decir, fuera de la vista de los evaluadores). Dicha intervención se realizó en 2 aulas separadas: una para los exámenes escritos y de opción múltiple y otra para el examen oral. Los estudiantes que eligieron no participar de la intervención tuvieron una actividad de repaso tradicional de 2 horas de discusión en grupos, llevada a cabo en otras aulas separadas de la intervención.

La evaluación de modalidad oral consistió en 3 preguntas dirigidas que permitían al estudiante expresar oralmente su conocimiento sobre el tema. Cada evaluación duró entre 10 y 15 minutos, y fue realizada por un docente con al menos 5 años de antigüedad en la cursada de Farmacología y con experiencia previa en realizar evaluaciones orales. Tal como se hace en la práctica habitual de toma de exámenes orales en la cursada, se compartió una lista de 25 preguntas sugeridas para evaluar, permitiendo al evaluador elegir entre estas preguntas u otras. La evaluación de preguntas de opción múltiple consistió en un examen de 30 preguntas (con 4 opciones por pregunta y 1 sola correcta), para completar la cual los estudiantes tuvieron hasta 40 minutos. Dichas preguntas fueron obtenidas de exámenes realizados en cursadas previas. La evaluación de modalidad escrita estaba compuesta por 3 preguntas

para desarrollar brevemente, para lo cual los estudiantes contaron con 25 minutos. Las preguntas de opción múltiple y las de la modalidad escrita fueron redactadas y corregidas por docentes con al menos 5 años de antigüedad en la cursada de Farmacología y con experiencia en la confección y corrección de exámenes.

Los estudiantes fueron divididos en dos grupos de manera aleatoria: un grupo comenzó con el examen oral y el otro con los exámenes de preguntas de opción múltiple y escrito para desarrollar. Luego, cada grupo reemplazó a la otra parte en la intervención.

Variables

Los puntajes de cada modalidad fueron definidos del 1 al 10. Por consenso entre los investigadores –y por analogía con el reglamento de examen de la carrera de Medicina de la Universidad de Buenos Aires– se consideró como valor de corte para la aprobación de cada examen un puntaje igual a 4 o mayor. En la modalidad oral, el puntaje se definía a criterio del examinador, por considerarse que esta práctica era lo más similar a la que ocurre en la práctica real. En el examen escrito para desarrollar, por consenso entre los evaluadores, se establecieron los puntajes en función del nivel de dificultad de cada pregunta (pregunta 1: 4 puntos; pregunta 2: 4 puntos; pregunta 3: 2 puntos). Se consideró que la información correcta sumaba puntos y que información incorrecta restaba puntos. Cada examen escrito para desarrollar fue corregido de manera independiente por dos examinadores (también docentes de la II Cátedra de Farmacología con al menos 5 años de antigüedad), considerándose la nota promedio entre ambos. Se evaluó el grado de acuerdo entre ambos evaluadores mediante un gráfico de Bland-Altman. En la modalidad de preguntas de opción múltiple se consideró como nota final la décima parte del porcentaje de preguntas correctas realizadas (es decir, 100% fue un 10; 60% fue un 6; 40% fue un 4).

Asuntos éticos

El estudio fue desarrollado según los principios de la Declaración de Helsinki. Cada participante dejó constancia de su consentimiento informado en una plataforma digital. La confidencialidad de los datos se mantuvo reemplazando el nombre y cualquier otra información que pudiera identificar de forma unívoca al estudiante, por un código de 2 dígitos. La única persona con acceso a dicha codificación fue un docente que no participó de ninguna de las instancias de evaluación durante la intervención. En el examen parcial real de la materia (realizado de manera oral 2 semanas después de la intervención), ningún docente que hubiera evaluado oralmente a un estudiante como parte de la intervención podía participar como evaluador.

La decisión de realizar la intervención de este estudio de investigación como parte del repaso del examen –y no como parte del examen real– se debió a razones éticas. No se consideró apropiado intervenir en una instancia de tal vulnerabilidad para el estudiante como la instancia real de evaluación, aún más habiéndose anticipado al comienzo de la cursada que la modalidad de examen parcial sería siempre y únicamente oral. Por ello se decidió que fuese voluntario y previo al examen.

Asuntos estadísticos

Se evaluó la hipótesis nula de que no hay diferencias en los puntajes promedio en cada una de las 3 modalidades de examen evaluadas.

En primer lugar, se realizó un análisis inicial unidireccional para evaluar la distribución de cada variable. Las variables continuas se resumieron utilizando la media y el desvío estándar (DE) para variables con distribución normal. Para las variables con distribución no normal se utilizó la mediana y el rango intercuartílo (RIC). Las variables categóricas se informaron con porcentajes en cada categoría. Las comparaciones entre grupos se hicieron con un test de T para variables continuas con distribución normal y con un test de U de Mann-Whitney para variables continuas de distribución no normal. Las comparaciones entre grupos para variables categóricas se hicieron con el test de chi-cuadrado o con el test exacto de Fisher. El grado de acuerdo entre los evaluadores del examen escrito para desarrollar se estimó mediante un gráfico de Bland-Altman (Fig. 1). La valoración del desenlace primario se hizo de la siguiente manera: al tratarse de 3 grupos de notas (es decir, oral, escrito para desarrollar y preguntas de opción múltiple), la diferencia de medias de notas se analizó en primer lugar mediante un test de ANOVA para muestras pareadas. Luego, de encontrarse diferencias estadísticamente significativas entre los grupos, se planificó un test de T pareado entre cada par de modalidades de evaluación. La valoración del desenlace exploratorio del porcentaje de aprobados por modalidad de examen se realizó mediante un test de chi-cuadrado. En todos los casos se hizo un ajuste por Bonferroni para comparaciones múltiples con un nivel de significancia de $\alpha \leq 0{,}017$. Todos los tests de hipótesis se valoraron a 2 colas. A excepción del análisis del desenlace primario, en todos los demás tests de hipótesis se consideró estadísticamente significativo un nivel de significancia de 0,05. Todos los análisis se realizaron utilizando el *software* STATA versión 13®.

RESULTADOS

Inicialmente se evaluó la participación de 36 estudiantes: 7 de ellos fueron excluidos (4 por ausencia y 3 por abandono), y se obtuvo un número final de 29 participantes enrolados. Las principales características de los estudiantes que participaron del estudio se encuentran resumidas en el cuadro 1. La mayoría eran hombres jóvenes que se habían desempeñado como ayudantes de otras materias. Casi la mitad de los estudiantes predijeron que obtendrían

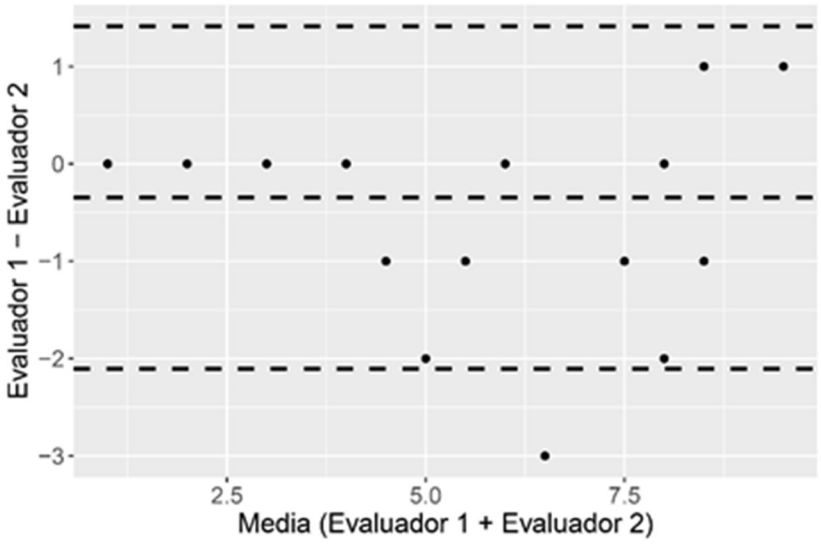

Figura 1. Gráfico de Bland-Altman de acuerdo entre dos evaluadores independientes de los exámenes escritos para desarrollar. Las líneas horizontales punteadas representan la media de la diferencia entre los dos evaluadores y el intervalo de confianza 95% (IC 95%) de estas diferencias.

Cuadro 1. Características sociodemográficas basales de los participantes. *Variable con distribución no normal. RIC: rango intercuartilo

Variable	Resultados (n = 29)
Hombres - n (%)	15 (51,7)
Edad - años * (mediana, RIC)	23 [23-23]
Ayudantes - n (%)	17 (58,6)
Años de ayudantía - años * (mediana, RIC)	1 [1-2]
Predicción - n (%)	
-Oral	14 (48,3)
-Opción múltiple	10 (34,5)
-Escrito	5 (17,2)

Cuadro 2. Promedio de notas obtenidas en cada una de las modalidades de examen.
Las notas se encuentran expresadas en puntos. Todas las variables presentaron una distribución normal y están expresadas como media y desvío estándar (DE). IC 95%: intervalo de confianza de 95%

	Oral	Escrito para desarrollar	Preguntas de opción múltiple
Oral	7,3 ± 2,3	Diferencia: 1,8 IC 95% 0,8 a 2,7 p < 0,01	Diferencia: 0,4 IC 95% -0,4 a 1,2 p = 0,37
Escrito para desarrollar		5,5 ± 2,6	Diferencia: 2,1 IC 95% 1,4 a 2,9 p < 0,01
Preguntas de opción múltiple			7,6 ± 1,3

Cuadro 3. Porcentaje de aprobados en cada una de las modalidades de examen

	Oral	Escrito para desarrollar	Preguntas de opción múltiple
Oral	96,6%	p = 0,69	p = 0,73
Escrito para desarrollar		75,9%	p = 0,07
Preguntas de opción múltiple			96,6%

mejores calificaciones en el examen oral. La corrección del examen escrito para desarrollar, realizado de manera independiente por dos evaluadores, derivó en un grado de acuerdo apropiado, si bien con cierta heterogeneidad según el valor de la nota, valorado por el gráfico de Bland-Altman (véase Fig. 1). La diferencia de medias entre las notas del examen escrito para desarrollar valoradas por ambos evaluadores fue 0,34 (IC 95% 0,69-0,00) puntos. Con respecto al desenlace primario del estudio, en el test de ANOVA la diferencia entre grupos fue 45,33 puntos, mientras que la diferencia dentro de los grupos fue 4,43 puntos ($F = 10,23$; $p < 0,01$). En los tests de T subsiguientes, tanto la nota promedio del examen oral como la del examen de preguntas de opción múltiple fueron significativamente superiores a la del examen escrito para desarrollar (Cuadro 2). No hubo diferencias estadísticamente significativas entre las notas promedio del examen oral y del examen de preguntas de opción múltiple, si bien la primera fue ligeramente mayor (véase Cuadro 2). En todos los casos se consideró un valor α ajustado por Bonferroni a comparaciones múltiples.

Con respecto al porcentaje de aprobados en cada modalidad de examen, no se encontraron diferencias estadísticamente significativas entre las distintas modalidades de evaluación (Cuadro 3).

Asimismo, se evaluó si los estudiantes eran capaces de predecir acertadamente cuál era la modalidad en la que su propio rendimiento sería más alto. El 31% de la muestra fue capaz de predecir acertadamente la modalidad en la que obtendrían una mayor calificación (IC 95% 14,2% a 47,9%).

DISCUSIÓN

Este estudio sugiere que no existen diferencias significativas entre la modalidad de examen oral y por preguntas de opción múltiple en la evaluación de conocimientos de Farmacología en estudiantes de Medicina. En cambio, en la modalidad de examen escrito para desarrollar, en comparación con cualquier otra modalidad, los resultados sugieren que los estudiantes obtienen peores calificaciones.

La modalidad de examen oral tendría una clara limitación relacionada con la subjetividad del evaluador[2,6]. Durante la evaluación oral, la subjetividad del encuentro entre sujeto evaluador-sujeto evaluado indefectiblemente desempeña un papel tanto en la valoración del primero como en el desempeño del último[2,6]. Ventouras y cols. exploraron cómo el estrés psicosocial asociado a una situación de evaluación oral podría favorecer a personalidades más extrovertidas y con mayor seguridad en sí mismas, desdibujando la relación entre la nota de examen y el grado de conocimiento del tema para evaluar[3]. Dado que esta subjetividad es una característica inherente a la modalidad de examen oral, decidimos respetarla en nuestra intervención, dejando mínimamente pautados los temas para evaluar y convocando a varios docentes a tomar examen oral simultáneamente. Otra característica del examen oral es el proceso de devolución o *feedback* inmediato que viene asociado a una dinámica de preguntas y respuestas entre el estudiante y el docente. Creemos que esta característica de la modalidad oral podría haber beneficiado más a los participantes que iniciaron la intervención con la evaluación oral y luego pasaron a las demás modalidades de examen, ya que habrían recibido una devolución oral por parte de un docente de temas relacionados con los que se estaban evaluando solo unos minutos antes de las evaluaciones escritas y por preguntas de opción múltiple. Si bien este interrogante podría contestarse mediante un análisis de subgrupos, según qué modalidad de examen fue evaluada primero en cada participante, dicho dato no fue recolectado por considerarse que no contaríamos con el poder suficiente para contestar esta pregunta. Queda entonces esta pregunta abierta para futuros estudios.

La modalidad de examen por preguntas de opción múltiple tendría varias ventajas: la objetividad de una evaluación que es idéntica para todos los estudiantes, la posibilidad de elegir entre enunciados ya formulados, la posibilidad de saltear una pregunta que genera dificultades para volver a ella después y la disponibilidad de tiempo para distribuir según la dificultad de las preguntas[6]. La falta de una situación de exposición personal directa ante un docente, tal como se presenta en un examen oral, también representa una característica de este método. Una de las desventajas del poder discriminativo de este método de evaluación es la posibilidad de acertar un cierto porcentaje de preguntas simplemente por azar[3]. Otra es su capacidad limitada para medir dimensiones cognitivas complejas y de alto nivel,

tales como la creatividad y la habilidad para resolver problemas[7]. A su vez, se le critica a esta modalidad de evaluación no acreditativa su enfoque reduccionista, al no permitir evidenciar el proceso del análisis llevado a cabo por el estudiante para la resolución del problema sin posibilidad de un *feedback* sobre este[8]. Finalmente, en el proceso de elaboración de preguntas de opción múltiple, es necesario construir preguntas con un poder discriminativo adecuado y que cuenten con distractores igualmente plausibles. Como este proceso representa un desafío considerable, en términos de dificultad y tiempo requerido, en ocasiones las preguntas elaboradas no cumplen con tales requisitos, tendiendo a ser simples, superficiales y rápidas de construir. Un examen construido con preguntas de estas características ulteriormente evalúa solo datos memorísticos y sacados de contexto, en lugar de conceptos fundamentales[9].

Por último, en la modalidad de examen escrito es donde los estudiantes de este estudio han demostrado un peor rendimiento. Esto posiblemente se encuentre asociado a que es la modalidad de examen menos frecuentemente empleada en la carrera de Medicina de la Universidad de Buenos Aires y a que, por consiguiente, los estudiantes están menos habituados a ella y no tendrían un entrenamiento apropiado para enfrentarse a este tipo de examen. Tal como sugiere Swanwick, un correcto desempeño en una evaluación escrita para desarrollar requerirá ciertos elementos críticos por parte del sujeto evaluado: una adecuada memoria de corto y de largo plazo, la capacidad de construir oraciones gramaticalmente correctas y no ambiguas, y la habilidad de organizar sus conocimientos a los fines de responder a la pregunta[1]. El autor reflexiona que la modalidad de examen escrito para desarrollar será una modalidad apropiada si estas competencias son elementos críticos del programa académico y que, por consiguiente, fueron ejercitadas durante la cursada. Por el contrario, si estas competencias son prerrequisitos de la cursada, el autor plantea el interrogante de si otras modalidades de evaluación serían más apropiadas. Otra posible explicación para este resultado es la creciente dificultad observada en la expresión adecuada por escrito[6]. Los resultados de nuestro estudio discrepan con los resultados de estudios de otros autores que no han encontrado diferencias significativas entre las modalidades de examen escrito para desarrollar y por preguntas de opción múltiple[3]. Estas diferencias posiblemente estén vinculadas con diferencias locales en las estrategias de educación y de evaluación.

Uno de los marcos conceptuales más utilizados para evaluar habilidades clínicas, competencias y desempeños fue propuesto por Miller[10]. La pirámide de Miller describe una serie de niveles que abordan desde el conocimiento acerca de la información pertinente (saber) en la base de la pirámide, pasando por la aplicación del conocimiento (saber cómo), el desempeño en un entorno estructurado (mostrar cómo), alcanzando finalmente la traducción del conocimiento en habilidades prácticas (hacer), ubicado en el vértice de la pirámide. De este modo, la pirámide de Miller se mueve desde la cognición, en los niveles inferiores, hasta la aplicación práctica y la implementación, en los niveles superiores. Apelando a este análisis, la modalidad de evaluación de preguntas de opción múltiple –y posiblemente la evaluación escrita para desarrollar– indagarían únicamente acerca del conocimiento de información pertinente, ubicado en la base de la pirámide. Por el contrario, la modalidad de evaluación oral permitiría valorar la aplicación del conocimiento y, en ciertos escenarios, incluso el desempeño en un entorno estructurado, apelando a niveles superiores de dominio del conocimiento. De ser correcta esta apreciación, los resultados de nuestro estudio se encontrarían más relacionados con el grado de familiaridad que los estudiantes presentan frente a cada modalidad de evaluación que a los dominios mismos que dichas modalidades evalúan.

Existe un preconcepto ampliamente difundido entre los estudiantes en cuanto a que la manera de estudiar para una evaluación debería modificarse sobre la base de la modalidad de examen, teniendo en cuenta las características propias de cada modalidad. La evaluación de un mismo cuerpo de conocimientos a través de diferentes modalidades de examen aplicadas de manera secuencial en un mismo día de evaluación podría eliminar este problema, permitiendo abordar la evaluación del tema de interés desde un enfoque integrador, reuniendo las ventajas y desventajas de cada modalidad.

En cuanto a las limitaciones de este estudio, en primer lugar, contamos con un tamaño muestral fijo restringido (es decir, 29 participantes) que posiblemente haya limitado su poder estadístico. Esto quizás haya restringido las posibilidades de detectar diferencias estadísticamente significativas entre la modalidad de preguntas de opción múltiple y la modalidad oral. En segundo lugar, a pesar de que todos los exámenes fueron elaborados y corregidos por docentes con al menos 5 años de antigüedad en la cursada de Farmacología y con experiencia en la confección y corrección de exámenes, siguiendo la práctica usual de la cursada, la validez de formas y de contenidos de cada modalidad de examen no fue formalmente valorada. Este hecho podría limitar su poder discriminativo. En tercer lugar, el hecho de que la evaluación se haya realizado como una intervención en un contexto de investigación –y no como una instancia real de evaluación– posiblemente haya tenido una influencia sobre los participantes tanto subjetiva (es decir, que los estudiantes no hayan tenido la misma motivación ni presión externa para tener su mejor rendimiento en el examen) como objetiva (al haberse realizado la intervención 2 semanas antes del verdadero examen, posiblemente no hayan contado con el tiempo necesario para estudiar tal como lo harían para un examen

real). Ambos fenómenos tendrían como consecuencia, en esta intervención, que los participantes no reflejen su verdadero rendimiento en un examen. Una cuarta limitación se encuentra relacionada con la estrategia de selección de la muestra. Al haber tomado una muestra por conveniencia de los estudiantes de la cursada, autoconvocados por decisión propia ya que la participación era voluntaria, una posible consecuencia es que hayamos analizado a estudiantes particularmente motivados y comprometidos con el estudio. Un dato de nuestro análisis que apoya esta presunción es que un 59% de los participantes eran ayudantes de otras materias. Quedaría por evaluar si todos los estudiantes de la cursada tienen el mismo perfil de rendimiento que aquellos que decidieron participar. Finalmente, en el momento de análisis de resultados, se constató que no se había definido "a priori" la estrategia de conversión entre el número de respuestas correctas en el examen de preguntas de opción múltiple y la nota de evaluación. Por este motivo se decidió analizar las notas de esta modalidad de examen de dos maneras distintas: una considerando el 40% de respuestas correctas como aprobado (equivalente a un 4) y la segunda considerando el 60% de respuestas correctas como aprobado (también equivalente a un 4) (véase Material suplementario). Mediante el análisis de estos diferentes valores de corte como criterio de aprobación para la modalidad de examen de preguntas de opción múltiple se confirmó la conclusión de la no diferencia con la modalidad oral, pero no se pudo confirmar la conclusión de una mayor nota promedio con respecto a la modalidad escrita para desarrollar. Serán necesarios otros estudios de mayor magnitud dirigidos específicamente a indagar acerca de esta discrepancia para confirmar tales resultados. En conclusión, las modalidades de examen oral y de preguntas de opción múltiple derivan en notas más elevadas que la modalidad escrita para desarrollar. Creemos que es importante tener en cuenta estas diferencias entre una modalidad de examen y otra al seleccionar un método de evaluación en Medicina. Asimismo, serían necesarios estudios adicionales para confirmar la hipótesis de que existen dificultades en la expresión escrita entre estudiantes de Medicina y, de confirmarse, deberían promoverse estrategias de educación dirigidas a este problema.

MATERIAL SUPLEMENTARIO

Comparación de distintos valores de corte para la aprobación en el examen de preguntas de opción múltiple

En un intento por explorar la influencia de diferentes valores de corte para la aprobación del examen de preguntas de opción múltiple, se evaluó si considerar un 60% de respuestas correctas (es decir, 18 preguntas y, a partir de este valor, aumentando en 1 punto la nota por cada 2 respuestas correctas) como criterio de aprobación (es decir, equivalente a una nota de 4) modificaría nuestros resultados. Este valor de corte mayor que el utilizado en nuestro estudio (es decir, 40% de respuestas correctas o 12 preguntas totales como equivalente a una nota de 4) se tuvo en cuenta por la posibilidad que existe, en esta modalidad de examen de preguntas de opción múltiple, de seleccionar la respuesta correcta simplemente por azar. Considerando como valor de corte de aprobación un 60% de respuestas correctas, el promedio de notas en la modalidad de preguntas de opción múltiple (6,2 ± 1,8 puntos) fue inferior al promedio de notas de la modalidad de examen oral (7,3 ± 2,3 puntos), si bien estas diferencias continuaron sin tener significancia estadística (diferencia: -1,1 puntos; IC 95% -1,9 a -0,2; p = 0,017 considerado no significativo para un valor α ajustado por Bonferroni de 3 comparaciones). Esto confirma los resultados obtenidos en nuestro estudio acerca de la no diferencia significativa entre la modalidad de examen oral y de preguntas de opción múltiple. Sin embargo, la comparación entre la nota promedio de la modalidad de preguntas de opción múltiple y la nota promedio del examen escrito para desarrollar (5,3 ± 2,5 puntos) deja de ser estadísticamente significativa con este nuevo valor de corte (diferencia: 0,9 puntos; IC 95% -0,1 a 1,7; p = 0,03 considerado no significativo para un valor α ajustado por Bonferroni a 3 comparaciones), discrepando con la conclusión de nuestro estudio en este punto. A modo de conclusión, podemos reafirmar que no habría diferencias estadísticamente significativas entre el promedio de las notas por modalidad de preguntas de opción múltiple y por modalidad oral, incluso modificando el valor de corte de aprobación del primero. Sin embargo, no podríamos confirmar la conclusión de que la nota promedio de la modalidad de preguntas de opción múltiple sería mayor que la de la modalidad escrito para desarrollar, ya que esta pareciera depender del valor de corte utilizado.

Comparación de resultados de examen entre grupo ayudantes y no ayudantes

En la modalidad de preguntas de opción múltiple se observó que la nota promedio de examen de los estudiantes que eran ayudantes de otras materias de la carrera de Medicina de la Universidad de Buenos Aires (7,5 ± 1,3 puntos), en comparación con la de los estudiantes que no lo eran (7,8 ± 1,3 puntos), no presentaba diferencias estadísticamente significativas (diferencia: 0,3 puntos; IC 95% -0,7 a 1,3; p = 0,61). Lo mismo se observó en la modalidad escrito para desarrollar (diferencia media: 0,1 puntos; IC 95% -2,1 a 1,9; p = 0,92) y en la modalidad oral (diferencia media: 1,6 puntos; IC 95% -3,3 a 0,1; p = 0,07). En todos los casos se consideró un valor α ajustado por Bonferroni a comparaciones múltiples.

A modo de conclusión, en ninguna de las tres modalidades se pudo observar una diferencia estadísticamente significativa entre los estudiantes que a su vez eran ayudantes de otras materias, con respecto a los que no lo eran. Posible-

mente esta conclusión esté limitada por el tamaño muestral del estudio, al no ser un análisis de subgrupos planeado *a priori*. Por lo pronto, podría sugerirse que ser ayudante no necesariamente implicaría un mejor desempeño académico en ninguna de las tres modalidades de examen.

Agradecimientos: Se agradece a todo el cuerpo docente y a los estudiantes de la II Cátedra de Farmacología del Departamento de Farmacología y Toxicología de la Facultad de Medicina de la Universidad de Buenos Aires, sin cuyo apoyo y colaboración este estudio hubiera sido imposible.

Conflictos de interés: los autores declaran no tener conflictos de interés.

REFERENCIAS

1. Swanwick T. Understanding Medical Education: Evidence, Theory and Practice. 2nd edition. Oxford: Wiley-Blackwell; 2013.
2. Huxham M, Campbell F, Westwood J. Oral versus written assessments: a test of student performance and attitudes, Assessment & Evaluation in Higher Education, 2012; 37:1,125-36
3.Ventouras E, Triantis, D, Tsiakas, et al. Comparison of examination methods based on multiple-choice questions and constructed-response questions using personal computers. Comput Educ. 2010; 54(2): 455-61.
4. Golda SD. A case study on multiple-choice testing in anatomical sciences. Anat Sci Educ. 2011;4(1):44-8.
5. Washburn S, Herman J, Stewart R. Evaluation of performance and perceptions of electronic vs. paper multiple-choice exams. Adv Physiol Educ. 2017; 41(4):548-55.
6. McTighe J, Wiggins G. Essential questions: opening doors to student understanding. Alexandria: ASCD; 2013.
7. Van der Vleuten C, Sluijsmans D, Joosten-ten Brinke D. Competence Assessment as Learner Support in Education. In: Mulder M. Competence-based Vocational and Professional Education. Switzerland: Springer International Publishing; 2017. p. 607-30.
8. Harrison CJ, Könings KD, Schuwirth L, et al. Barriers to the uptake and use of feedback in the context of summative assessment. Adv Health Sci Educ Theory Pract. 2015; 20(1):229-45.
9. Raymond MR, Stevens C, Bucak SD. The optimal number of options for multiple-choice questions on high-stakes tests: application of a revised index for detecting nonfunctional distractors. Adv Health Sci Educ Theory Pract. 2019; 24(1):141-50.
10. Miller GE. The assessment of clinical skills/competence/performance. Acad Med. 1990; 65(9 Suppl):S63-7.