Recibido: 15.02.2019 | **Aceptado:** 15.03.2019

Palabras clave: Examen, fisiología y reactivos.

Una manera sencilla de analizar

exámenes de opción múltiple

PATRICIA PÉREZ CORNEJO
gperez@uaslp.mx
FACULTAD DE MEDICINA, UASLP

Para los profesores universitarios no siempre es fácil estar al día en cuanto a las mejores prácticas docentes; se profundiza en el dominio del campo académico, pero se dedica menos tiempo a evaluar los métodos de enseñanza utilizados. A partir de 2010, bajo la tutoría de una colega de la Universidad de West Chester (West Chester, PA, USA), introduje cambios en mi práctica docente. A partir de entonces he asistido a congresos sobre educación donde he aprendido sobre nuevos métodos de enseñanza y en los últimos meses me he interesado por el tema de los exámenes.

En la Facultad de Medicina de la Universidad Autónoma de San Luis Potosí se evalúa a los estudiantes de la Licenciatura de Médico Cirujano mediante exámenes de opción múltiple, que presentan ventajas cuando se trabaja con grupos grandes, de 100 a 150 personas, ya que permite calificar con rapidez (algo que los alumnos siempre agradecen) y de manera automática mediante el uso de un lector óptico, que evita ambigüedades en la manera de calificar y permite hacer un análisis estadístico de los resultados. Además, el uso de guías para la elaboración de preguntas adecuadas permite construir un examen bien redactado que puede usarse para la evaluación de los estudiantes.

Las calificaciones obtenidas de las evaluaciones pueden usarse como un indicador del aprovechamiento académico. En consecuencia, un examen puede evaluar la eficacia de distintas estrategias de enseñanza; éste es el aspecto que me interesa. Por ejemplo, para mejorar mi práctica docente he probado diferentes estrategias de enseñanza, pero cómo saber si la que escogí es efectiva. Lo más sencillo es suponer que una estrategia efectiva dará como resultado un mayor aprovechamiento académico; en otras palabras, las calificaciones de los estudiantes mejoran al aplicar dicha estrategia, por lo tanto, uno podría comparar las calificaciones obtenidas por diferentes grupos, asumiendo que la única variable de cambio es el método de enseñanza empleado.

Un breve análisis comparativo

Para investigar si el método de enseñanza basado en casos de estudio era mejor que las clases frente a grupo, realicé un análisis retrospectivo que incluyó datos obtenidos durante 10 años dando clases a estudiantes de segundo año, en la Facultad de Medicina, inscritos en el curso de Fisiología. En el análisis sólo se incluyeron las calificaciones obtenidas en el examen de la sección de Fisiología Endocrina; la sección consta de 12 clases y una práctica de laboratorio. Los datos colectados se dividieron en dos mitades: grupos de aprendizaje pasivo (recibieron clases tradicionales frente a grupo) y grupos de aprendizaje activo (recibieron clase y además participaron durante ellas contestando preguntas y discutiendo en grupo diferentes casos de estudio). Los primeros sumaron un total de 615 estudiantes que tomaron clases durante los años 2005 (124), 2006 (108), 2007 (107), 2008 (137) y 2009 (139). Los segundos activo sumaron un total de 744 estudiantes que tomaron clases durante los años 2012 (178), 2013 (201), 2014 (105), 2015 (126) y 2016 (134). En los paréntesis se incluye el número de estudiantes inscritos en cada año escolar. En la figura 1 se comparan las calificaciones promedio obtenidas por los cinco grupos de aprendizaje pasivo (barras negras) contra las calificaciones de otros cinco grupos de aprendizaje activo (barras grises).

Como puede observarse, no hay una mejora sustancial en cuanto a calificaciones en los grupos de aprendizaje activo, pero, ¿podemos concluir que el nuevo método de enseñanza no beneficia a los estudiantes? Tal vez, pero esto contradice lo reportado en la literatura (Freeman *et al*., 2014). Al analizar los exámenes aplicados en cada uno de esos años, me di cuenta de que los exámenes tenían distintos grados de dificultad. ¿Es esto relevante? Benjamin Wright y Mark Stone comentan en su libro *Diseño óptimo de pruebas*: "Casi todo el mundo se da cuenta de que el cincuenta por ciento de respuestas correctas en una prueba fácil no significa lo mismo que cincuenta por ciento de respuestas correctas en una prueba difícil" (1979, p. 7).

Figura 1.

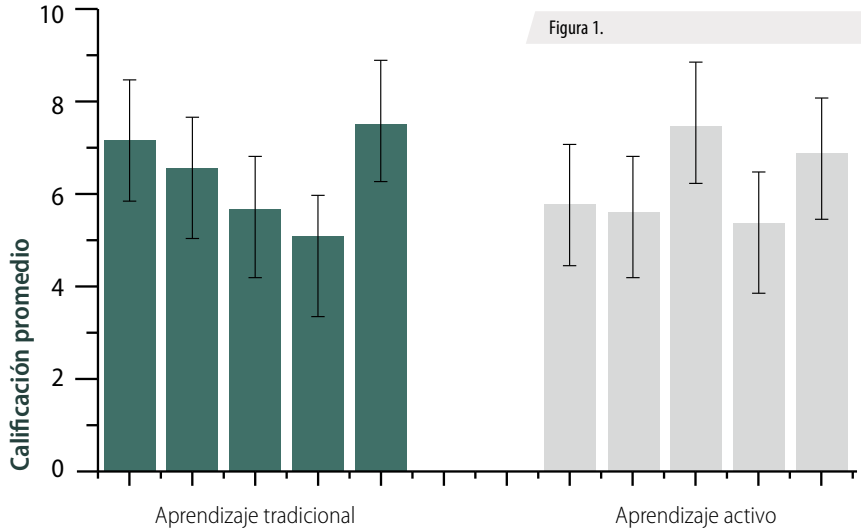


Tabla 1.

Reactivo	1	2	3	4	5	6	7	8
Grado de dificultad	0.46	0.33	0.62	0.38	0.10	0.75	0.71	0.33
Índice de discriminación	0.40	0.40	0.46	0.66	0.26	0.2	0.33	0.26

Figura 2.

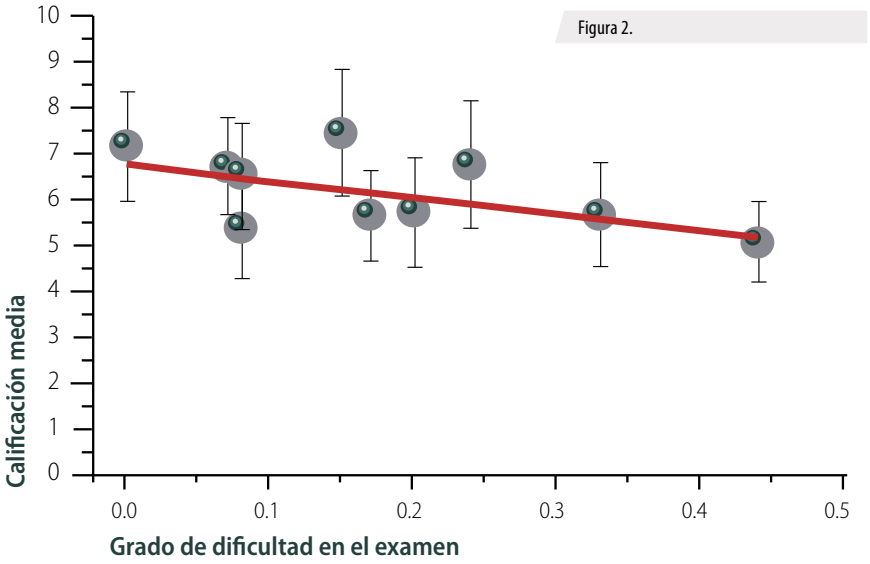
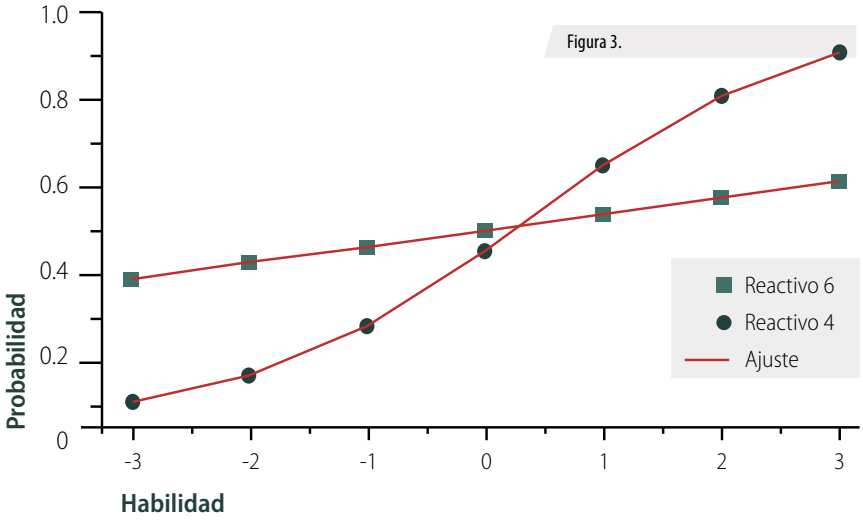


Figura 3.



La figura 2 muestra que en los grupos comparados la calificación promedio (media ± desviación estándar) disminuyó conforme aumentó la dificultad del examen, lo que nos indica que la variación observada en las calificaciones se debe a la dificultad del examen y no al método de enseñanza utilizado (clases frente a grupo versus el método de casos donde participaron los estudiantes). Esto sugiere que al hacer comparaciones debemos asegurarnos de que el instrumento de medición (el examen) funcione de manera adecuada. Lo recomendable sería aplicar exámenes equivalentes en los grupos a comparar. Pero, ¿cómo saber si mis exámenes son adecuados?

El método de Rasch

En el campo de la psicología se desarrolló el método de Rasch, usado por psicólogos educativos en la aplicación de pruebas estandarizadas, las cuales son evaluaciones a gran escala que permiten comparar distintas poblaciones de estudiantes.

El método de Rash usa una escala lineal para medir tanto las habilidades de los estudiantes como la dificultad de las preguntas. Con esos datos se construye un modelo matemático que permite determinar la calidad de cada una de las preguntas (Tristán, 1998). Por ejemplo, Martín y colaboradores (2011) lo aplicaron para calibrar un examen diagnóstico de química compuesto por 12 reactivos (o preguntas). Un examen calibrado de química es uno que puede aplicarse a distintas poblaciones para hacer una medición con propiedades invariantes (Martín *et al*., 2011).

Análisis del examen

No siempre es posible calibrar los exámenes que aplicamos; sin embargo es importante tener presente que su calidad está dada por la excelencia de cada uno de los reactivos. Un examen de calidad debe ser confiable (medir consistentemente el nivel de aprendizaje) y válido (medir realmente lo que intenta medirse) de manera que sirva para evaluar el aprendizaje de los jóvenes. La confiabilidad de un examen se ve afectada por dos factores importantes: el grado de dificultad (indica que tan difícil es la pregunta para los alumnos) y la capacidad de discriminación de cada reactivo (indica si la pregunta es capaz de diferenciar entre estudiantes con mayor o menor conocimiento) (Engelhardt, 2009).

Entonces, ¿puedo usar estos parámetros para hacer un análisis rápido de mi examen? A continuación se muestra como ejemplo el análisis de ocho reactivos incluidos en el examen ordinario de la materia de Fisiología, impartida en 2018, conformado en total por 50 preguntas

de opción múltiple, de las cuales las ocho de las analizadas correspondían a la sección de Fisiología endocrina. El examen fue presentado por 52 estudiantes.

Para cada reactivo se calculó el índice de dificultad al dividir el número de aciertos entre el número de estudiantes que presentaron el examen. El rango de valores para este indicador va de 0 (nadie contestó correctamente la pregunta) a 1 (todos contestaron bien la pregunta), de manera que mientras mayor sea el índice de dificultad más fácil será la pregunta (Engelhardt, 2009). Asimismo, se calculó el índice de discriminación para cada reactivo. Para ello se ordenaron las calificaciones de mayor a menor, luego la lista se dividió en tres partes. Estudiantes con mayor conocimiento (27 por ciento) localizados en el tercio superior (S) y estudiantes con menor conocimiento (27 por ciento) localizados en el tercio inferior (I). Luego se sumaron los aciertos en cada grupo y se aplicó la formula D= (S-I)/(N/2). Donde N es el número total de estudiantes que presentaron el examen (Engelhardt, 2009).

La tabla 1 resume los índices de dificultad y discriminación que se calcularon para cada reactivo. En la primera fila se muestra el grado de dificultad calculado. Para este parámetro un valor de 0.5 se considera óptimo por tener una dificultad media (Engelhardt, 2009). Un reactivo fácil tiene valores mayores a 0.7 y uno difícil menores a 0.3. Los reactivos 1, 2, 3, 4 y 8 tienen un grado de dificultad adecuado (0.3 a 0.6), mientras que los 6 y 7 son fáciles (>0.7) y el 5 es difícil (0.1). Como ya se explicó, el índice de discriminación se calculó con base en el patrón de respuesta de los estudiantes con mayor y menor conocimiento. El rango de valores para este parámetro va de -1 a +1. Reactivos con valores de 0.3-1.0 se considera discriminan de manera adecuada (Engelhardt, 2009). Como puede observarse en la tabla 1, los reactivos 5, 6 y 8 no discriminan de manera adecuada porque tienen un índice menor a 0.3.

De acuerdo con la teoría de respuesta al ítem (TRI), la cual utiliza el método de Rasch, se establece que la probabilidad de que un estudiante responda correctamente a un reactivo depende de su habilidad (conocimiento) y de las características del reactivo (dificultad y poder de discriminación). Esta relación entre la probabilidad y la habilidad se ajusta a una función logística, lo que da lugar a la curva característica del reactivo (CCR).

La figura 3 muestra la CCR para los reactivos 4 y 6. En esta representación gráfica (que incluye los parámetros de dificultad y poder de discriminación calculados) puede observarse que el reactivo 4 es muy diferente al 6. El índice de dificultad y discriminación se muestran como el punto de inflexión y la pendiente de la curva (Tristán, 1998). Aquí puede verse que el reactivo 6 tiene un grado de discriminación no óptimo (discriminación=0.2), ya que la pendiente de su curva es poco profunda. El caso del reactivo 4 muestra una curva con un punto de inflexión y una pendiente adecuadas (dificultad=0.38 y discriminación=0.66). Así, la CCR nos permite inspeccionar de manera visual la calidad de cada reactivo.

Conclusiones

El análisis de los exámenes puede llevarse a cabo utilizando la teoría de respuesta al ítem, la cual permite ubicar la habilidad de los estudiantes y la dificultad de las preguntas en una sola escala. Al calcular los índices de dificultad y de discriminación de cada pregunta incluida en el examen, se aplica esta teoría. De esta manera es posible hacer un análisis rutinario de exámenes de opción múltiple con el fin de garantizar la calidad de los reactivos. Mi motivación inicial fue comparar las calificaciones obtenidas con distintos métodos de enseñanza para saber si usar uno nuevo da como resultado un mayor aprovechamiento académico. Con este análisis puedo decir que si elaboro preguntas con grados de dificultad similar, podré aplicarlas para comparar calificaciones obtenidas por diferentes grupos.

En resumen, al asegurar que los exámenes aplicados son adecuados, puede evaluarse con mayor confianza el aprendizaje de nuestros estudiantes. UP

PATRICIA PÉREZ CORNEJO

Es doctora en ciencias por la Universidad de Rochester en Nueva York, Estados Unidos de América. Es profesora investigadora de tiempo completo nivel VI en la Facultad de Medicina de la UASLP. Trabaja en el proyecto "Interacciones entre canales iónicos y lípidos en la membrana plasmática".

Referencias bibliográficas:

Engelhardt, P. V. (2009). An introduction to classical test theory as applied to conceptual multiple-choice tests. *Getting Started in Physics Education Research*, *2*(1), pp. 1-40.

Freeman, S., Eddy, S. L., McDonough, M., Smith, M. K., Okoroafor, N., Jordt, H. y Wenderoth, M. P. (2014). Active learning increases student performance in science, engineering, and mathematics. *Proceedings of the National Academy of Sciences*, *111*(23), pp. 8410-8415.

Martín Guaregua, N., Díaz Torres, C., Córdoba Herrera, G., y Picquart, M. (2011). Calibración de una prueba de química por el modelo de Rasch. *Revista Electrónica de Investigación Educativa*, *13*(2), 132-148.

Tristán, A. (1998). *Análisis de Rasch para todos: Una guía simplificada para evaluadores educativos. México: Centro Nacional de Evaluación para la Educación Superior.*

Wright, B. D. y Stone, M. H. (1979). *Best Test Design: Rasch Measurement [Diseño óptimo de pruebas]*. Chicago, IL: Mesa Press.