Sección IV
Temas complementarios

Esta sección incluye algunos temas complementarios que están relacionados con la evaluación.

Capítulo 8
Interpretación de los resultados del análisis de los ítems

❖

Muchas facultades proporcionan a los profesores un informe con los resultados del análisis de los ítems después de cada examen de opción múltiple. Este informe es una excelente fuente de información sobre el ítem y es útil en la evaluación de su calidad, así como de la exactitud de la clave de respuestas.

A continuación se presentan ejemplos de resultados de cuatro ítems; cada uno ejemplifica una situación común. Se dividió a los alumnos que rindieron el examen en un grupo Superior y un grupo Inferior, según su rendimiento general en todo el examen. Si tiene un número reducido de alumnos, incluya el 50% superior de los alumnos en el grupo de nivel Superior y el otro 50% en el grupo de nivel Inferior. Si tiene un número elevado de alumnos, podría incluir el 25% superior de los alumnos en el grupo de nivel Superior y el 25% inferior en el grupo de nivel Inferior.

Normalmente, el informe de los resultados del análisis de los ítems indica el porcentaje de alumnos en cada grupo que seleccionó cada opción. A menudo, también incluye alguna medición de la dificultad del ítem (p. ej., el "valor p", o sea la proporción de alumnos que respondieron correctamente a la pregunta) y alguna medición de la discriminación (p. ej., un biserial o biserial puntual). Recomendamos concentrar la atención en el patrón de las respuestas en lugar del nivel de dificultad o índice de discriminación.

Para cada ejemplo de ítem que se presenta a continuación, se muestra el porcentaje de alumnos que seleccionó cada opción. La fila designada "Total" muestra el porcentaje del grupo entero que seleccionó cada opción. Por ejemplo, en el ítem N.º 1, 1% del grupo Superior seleccionó la opción A; 1% seleccionó la opción B; 91% seleccionó la C; 4% seleccionó la D; 1% seleccionó la E y el 2% seleccionó la F. En el mismo ítem, el 20% del grupo Inferior seleccionó la opción A; 6% seleccionó la B, etc. El asterisco en la opción B indica que ésta era la supuesta respuesta correcta.

Item N.º 1

Grupo	A	B*	C	D	E	F
Superior	1	**1**	91	4	1	2
Inferior	20	**6**	51	14	6	3
Total	**9**	**2**	**76**	**8**	**3**	**2**

Valor p: 2 **índice de discriminación: -0,21**

Interpretación: Este es el modelo típico de un ítem que tiene una clave equivocada: si la respuesta es la opción B, el ítem es muy difícil y el índice de discriminación es negativo. Si la clave es B, solamente el 2% de los alumnos respondió correctamente. La respuesta correcta es casi con seguridad la opción C, pero un experto en el contenido del ítem deberá revisarlo para estar seguro. Si la respuesta correcta es la opción C, el valor p se transforma en 76 y el índice de discriminación es de 0,46; ambos datos son excelentes desde la perspectiva estadística y no hay justificativos para realizar cambios en el texto del ítem.

Item N.º 2

Grupo	A	B	C*	D	E	F
Superior	0	1	**90**	3	3	3
Inferior	0	1	**60**	25	8	6
Total	**0**	**1**	**74**	**12**	**7**	**6**

Valor p: 74 **índice de discriminación: 0,33**

Interpretación: El 90% del grupo Superior y el 60% del grupo Inferior seleccionó la respuesta correcta. Estas son estadísticas generales excelentes. Se pueden volver a redactar las opciones A y B antes de volver a usar el ítem porque muy pocos alumnos seleccionaron esas opciones.

Item N.º 3

Grupo	A	B	C*	D	E	F
Superior	44	1	**50**	2	1	2
Inferior	20	15	**21**	22	20	2
Total	**32**	**7**	**34**	**14**	**11**	**2**

Valor p: 34 índice de discriminación: 0,30

Interpretación: El 50% del grupo Superior y el 21% del grupo Inferior seleccionó la respuesta correcta. Este es un ítem muy difícil que probablemente NO ESTA BIEN REDACTADO. Un gran número de alumnos del grupo Superior seleccionó la opción A; el ítem puede tener una redacción deficiente. Verifique la "imparcialidad" de la opción A. Asegúrese de que la opción A no sea igualmente correcta.

Item N.º 4

Grupo	A	B	C*	D	E	F
Superior	18	10	**51**	17	2	2
Inferior	24	24	**21**	25	4	2
Total	**22**	**17**	**34**	**22**	**3**	**2**

Valor p: 34 índice de discriminación: 0,30

Interpretación: El desglose de los grupos Superior e Inferior en la opción C es igual al del ítem N.º 3; pero este ítem puede estar BIEN REDACTADO. A diferencia del ítem N.º 3, los alumnos que no conocen la respuesta correcta se distribuyen ampliamente entre los diferentes distractores. Obviamente, sería preferible revisar las opciones A, B y D para controlar su corrección y claridad.

Capítulo 9
Cómo establecer un estándar de aprobado/reprobado

❖

Definiciones y principios básicos

Los estándares pueden clasificarse como *relativos* o *absolutos.* Un *estándar relativo* se basa en el rendimiento del grupo que rinde el examen. Los alumnos aprueban o reprueban según el nivel de su rendimiento con respecto a los otros alumnos que rinden el examen. Los siguientes son ejemplos de estándares *relativos*.

> Aquellos alumnos que obtengan un puntaje menor a 1,2 desviaciones estándares por debajo de la media, no aprobarán el examen.
> El 20 por ciento inferior del grupo no aprobará el examen.

Por el contrario, un *estándar absoluto* no compara el rendimiento de un alumno con el de los otros que rinden el examen. Los alumnos aprueban o reprueban solamente según el nivel de su rendimiento, sin tener en cuenta el desempeño de los otros alumnos. Todos los alumnos pueden aprobar o todos pueden reprobar. El siguiente es un ejemplo de un estándar *absoluto*:

> Aquellos que respondan en forma correcta a menos del 60 por ciento de las preguntas, no aprobarán.

A menos que existan razones convincentes para reprobar a un número determinado de alumnos, es preferible un estándar absoluto (basado en el rendimiento del alumno) a uno relativo (basado en un índice de reprobación en particular).

Principios básicos para establecer los estándares

- Sin tener en cuenta el procedimiento utilizado, el establecimiento de los estándares requiere de un criterio determinado. En todos los casos, el establecimiento de estándares será arbitrario pero no necesariamente caprichoso.

- A menos que exista una razón específica para reprobar a un número determinado de alumnos (por ejemplo, solamente existe un número determinado de espacios disponibles), un estándar basado en el dominio que tiene el alumno del contenido del examen es preferible a un estándar basado en un índice de reprobación en particular.

- Es prudente que participen varios jueces informados en el proceso de establecimiento de estándares. Se presentarán diferencias de opinión, y el uso de varios jueces reducirá los efectos conocidos como "halcón/paloma" (más exigente/más indulgente).

- Se deberán proporcionar datos a los jueces sobre el rendimiento de los alumnos en algún momento del proceso de establecimiento de los estándares. El establecimiento de los estándares sin usar dichos datos podrá generar estándares basados en información insuficiente y resultados poco razonables.

Una referencia útil sobre cómo establecer estándares es:

Livingston SA, Zieky MJ. *Passing Scores: A Manual for Setting Standards of Performance on Educational and Occupational Tests.* Princeton, NJ: Educational Testing Service; 1982.

Dos métodos de determinación de estándares en base a juicios sobre los ítems

El método de Ebel modificado

- Un grupo analiza las características del "alumno en la frontera" (entre aprobado y reprobado): es decir, aquel alumno cuya aptitud es apenas suficiente como para permitirle aprobar el examen.

- Los jueces clasifican a los ítems como "esencial", "importante" o "indicado".

- Los jueces indican el número de ítems en cada categoría que obtendría un alumno en la frontera.

- El estándar de aprobado/reprobado se calcula como el porcentaje de puntos posibles que obtendría un alumno en la frontera.

El método de Angoff modificado

- Un grupo analiza las características del "alumno en la frontera" (entre aprobado y reprobado).
- Para cada ítem del examen, los jueces calculan el porcentaje de alumnos en la frontera que responderían correctamente al ítem.
- El estándar de aprobado/reprobado para el examen es el promedio de los porcentajes de los ítems.

Variantes comunes en el método de Angoff

- Los jueces pueden o no tener las respuestas correctas a las preguntas.
- Los jueces pueden o no disponer de la información referente al porcentaje de los alumnos que respondieron correctamente a cada pregunta.
- Luego de un período de capacitación, los jueces pueden continuar trabajando en grupo o individualmente.

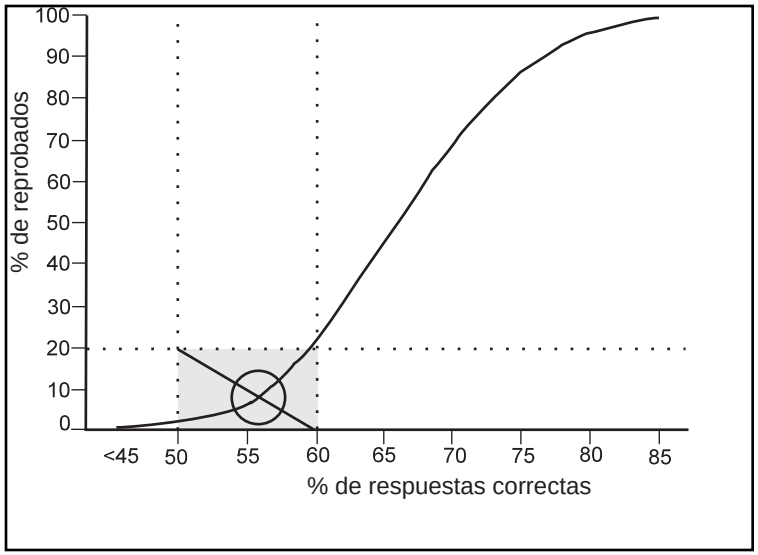
Estándares de compromiso relativo/absoluto: el método de Hofstee

Recientemente, se han desarrollado varios "modelos de compromiso" que utilizan las ventajas de los procedimientos de establecimiento de estándares tanto relativos como absolutos. Uno de estos métodos es el de Hofstee, que se describe a continuación.

1. Los jueces revisan una copia del examen.

2. Luego, los jueces asignan los valores que se indican a continuación, que definen los estándares aceptables:
   Porcentaje mínimo aceptable de alumnos reprobados (índice mínimo de reprobación)
   Porcentaje máximo aceptable de alumnos reprobados (índice máximo de reprobación)
   El puntaje más bajo que permitiría que un alumno apruebe el examen (punto mínimo de aprobación)
   El puntaje más elevado requerido para que un alumno apruebe (punto máximo de aprobación)

3. Después del examen, se grafica una curva que muestre el índice de reprobación en función del puntaje de aprobación. (En la figura que se muestra, la curva se extiende desde la parte inferior izquierda hasta la parte superior derecha.)



4. Los cuatro valores obtenidos en el punto N.º 2 se trazan para formar un rectángulo. A menudo, se usan los valores medianos del grupo de jueces. En el ejemplo, se estableció que el índice de reprobación apropiado estaba entre 0 y el 20% (ver las líneas horizontales); se determinó que el punto adecuado de aprobado/reprobado estaba entre un 50% y un 60% de respuestas correctas (ver las líneas verticales).

5. Se traza una línea en la diagonal desde la parte superior izquierda hasta la parte inferior derecha. El punto de intersección con la curva es el estándar (es decir, un poco más del 55% de respuestas correctas en la figura).

Una referencia útil en los métodos de compromiso es:

de Gruijter D. Compromise models for establishing examination standards. *Journal of Educational Measurement.* 1985;22:263-269.

Capítulo 10
Reflexiones varias sobre temas relacionados con la evaluación

❖

A continuación se presentan unos comentarios sobre una mezcla variada de temas relacionados con exámenes. En general, los puntos que se tratan son conjeturas y se basan en anécdotas más que en la evidencia. Es decir, reflejan nuestros prejuicios en lugar de los resultados de una investigación.

Exámenes de múltiples estaciones (también conocidos como Exámenes prácticos, Carreras de obstáculos, OSCE (examen clínico objetivo estructurado))

Si bien es complejo instalar y administrar este tipo de exámenes, desde el punto de vista logístico, son muy útiles en el área de las ciencias básicas, particularmente para evaluar las destrezas prácticas que no se pueden medir en exámenes de papel y lápiz (por ejemplo, la capacidad de usar un microscopio, de realizar un procedimiento de laboratorio). Además, la reproducción de algunos tipos de materiales (por ejemplo, resultados de estudios de diagnóstico por imágenes, materiales ilustrativos en colores) es muy costoso; en dichas situaciones, el método de múltiples estaciones se puede usar para reducir los costos de la administración del examen.

Exámenes para completar en casa

Los exámenes para completar en la casa pueden constituir una experiencia de aprendizaje importante ya que estimulan a los alumnos a leer en profundidad y ampliamente los temas importantes. Lamentablemente, los alumnos tienden a producir libros como respuestas y no queda claro si las respuestas presentadas por los alumnos representan su propio trabajo. Se pueden obtener las mismas ventajas mediante la distribución de (un gran conjunto de) preguntas de examen con anticipación y la administración de (un subconjunto de) estas preguntas en forma de examen de tiempo fijo.

Exámenes a libro abierto

Estos exámenes pueden ser una buena idea debido al impacto que tienen sobre el tipo de preguntas que prepara el profesor. En los exámenes a libro abierto, no tiene sentido realizar preguntas sobre hechos aislados que pueden buscarse rápidamente en una sola página del libro de texto; por lo tanto, el material de evaluación desarrollado para este tipo de exámenes tiende a concentrarse más en la comprensión de principios y conceptos fundamentales de situaciones problemáticas.

Pruebas breves frecuentes o exámenes poco frecuentes

Las evaluaciones poco frecuentes convierten a cada examen en un acontecimiento importante; es posible que los alumnos dejen de asistir a clases para prepararse, y esta situación es indeseable. Además, con los exámenes poco frecuentes, los alumnos quizás no podrán determinar si estudian el material correcto o aprenden con la suficiente profundidad. A pesar de que podrán exigir más tiempo al profesor, las evaluaciones periódicas reducen la importancia de cada examen individual y ayudan a los alumnos a evaluar mejor su avance. En general, las evaluaciones frecuentes son preferibles, aunque es probable que los alumnos se quejen de todos modos sin considerar el método adoptado.

Guardar los exámenes de manera "segura" o permitir que los alumnos se queden con ellos

Debido a que los exámenes pueden tener un efecto de "dirección" importante en el aprendizaje de los alumnos, el permitir que ellos retengan el material de evaluación puede ayudarles a concentrarse en temas clave y a reforzar los objetivos del plan de estudios y del curso (si se supone que los materiales de examen los reflejan). Sin embargo, la preparación de preguntas adecuadas para un examen implica dedicar mucho tiempo y la calidad del material de la prueba se puede deteriorar con el paso de los años si el profesor tiene que desarrollar nuevos materiales de examen cada vez que enseña un curso. El método más apropiado puede ser el de disponer de una muestra de preguntas de buena calidad a fin de guiar el aprendizaje de los alumnos pero mantener un banco de preguntas "seguras" para uso repetido. Es necesario recordar que es probable que la seguridad sea deficiente ya que los alumnos muchas veces memorizan las preguntas y las intercambian.

Uso de exámenes acumulativos

Los exámenes acumulativos que responsabilizan a los alumnos de todo el material presentado hasta la fecha fomentan la concentración en las interrelaciones entre los temas, particularmente si las preguntas del examen requieren de la comprensión tanto de los temas recientemente presentados como de los anteriores. El uso de exámenes que abarcan solamente el material presentado desde el examen anterior estimula a los alumnos para que estudien los temas aislados; se pueden perder las relaciones entre los temas de unidades diferentes. Ya que los alumnos pueden tener un mal rendimiento en una serie de exámenes porque nunca dominan el material básico, este método también puede, por otra parte, motivar a los alumnos para que corrijan sus deficiencias.

Uso de exámenes integradores entre cursos diferentes

Al igual que con el uso de los exámenes acumulativos, los integradores entre cursos motivan a los alumnos para que analicen las interrelaciones entre las disciplinas y los temas; esto debería ser muy útil para la retención a largo plazo y para la aplicación de conocimientos en el área de las ciencias básicas a situaciones clínicas. Generalmente, se necesitan profesores de ciencias básicas y de departamentos clínicos para la preparación de dichos exámenes. Si bien implican dedicar mucho tiempo, este esfuerzo en conjunto puede dar como resultado mejores materiales de examen y además producir discusiones útiles entre los profesores sobre el material que debe incluirse en el plan de estudios.