

ID:564

METODOLOGÍA PARA DETERMINAR LA CALIDAD DE LOS INSTRUMENTOS DE EVALUACIÓN.

Díaz Rojas, Pedro A; Leyva Sánchez. Elizabeth. Cuba

RESUMEN

Introducción. La evaluación del aprendizaje es el mecanismo regulador del proceso enseñanza aprendizaje, favorece la retroalimentación del proceso y permite direccionar las acciones en pos de la calidad del producto resultante. La evaluación del aprendizaje acredita y certifica al estudiante como nuevo profesional y al residente como nuevo especialista. **Objetivo.** Obtener evidencias de la calidad de los instrumentos que aplicamos en estos tipos de exámenes. **Método.** Se aplicaron indicadores de calidad como el índice de dificultad, el índice de discriminación, el coeficiente de discriminación y el alfa de Cronbach para determinar consistencia interna. **Resultados.** Propuesta de una metodología para la evaluación científica de la calidad de los instrumentos de evaluación. **Conclusiones.** La aplicación de la metodología propuesta tiene como objetivo ayudar a la realización de un análisis científico y normalizado, que permita determinar los contenidos con resultados deficientes y establecer su tratamiento posterior. El resultado final del análisis de calidad de un instrumento debe incluir la valoración que realicen los profesores de los colectivos de las asignaturas en su planeación y los resultados de la aplicación de la metodología propuesta.

Palabras claves: Evaluación del aprendizaje, calidad de instrumentos de evaluación.

INTRODUCCIÓN

La evaluación del aprendizaje es el mecanismo regulador del proceso enseñanza aprendizaje, nos posibilita su retroalimentación, dirección y perfeccionamiento. Determina los modos de actuación que los estudiantes incorporan en su proceso formativo y acredita y certifica el cumplimiento de los objetivos terminales del futuro profesional y especialista.¹

Salas Perea define la evaluación como la actividad que consiste en el análisis e interpretación de los resultados de diferentes instrumentos de medición sobre la actuación de un educando o candidato ante una actividad y/o tarea, en comparación con un patrón o modelo estandarizado, y que se efectúa con la finalidad de tomar una buena decisión.²

La resolución 210-2007 expresa que la evaluación es un proceso consustancial al desarrollo del proceso docente educativo. Tiene como propósito comprobar el grado de cumplimiento de los objetivos formulados en los planes y programas de estudio de la educación superior, mediante la valoración de los conocimientos y habilidades que los estudiantes adquieren y desarrollan; así como, por la conducta que manifiestan en el proceso docente educativo. Constituye a su vez, una vía para la retroalimentación y la regulación de dicho proceso.³

Su resultado final es emitir un juicio que se expresa en una calificación que lo define. Lo esencial en todos los casos radica en que la calificación sea objetiva, que exprese una correcta correspondencia con la calidad de la asimilación y por lo tanto, la necesidad de establecer criterios uniformes, que disminuyan la influencia del factor subjetivo, y que frente a una actuación determinada, con iguales resultados, el grupo de profesores otorguen iguales calificaciones.^{1, 3, 4.}

En la actualidad, con el incremento del contenido de las ciencias y la tendencia a su integración,⁶ podemos definir como pasos para la selección de los contenidos de una evaluación, la identificación de los núcleos básicos de contenidos y sus esencialidades, su relación con los problemas que los educandos deben ser capaces de enfrentar y/o resolver, las tareas que debe desarrollar para darles solución y que definen su futura competencia.

De esta forma logramos que los instrumentos evaluativos planeados⁵ reúnan los requisitos de que sus contenidos estén directamente relacionados con los objetivos educativos y que al aplicarlos demuestren su validez y confiabilidad.^{1, 3, 7-9.}

Para lograr estas propiedades debemos tener presente el análisis del colectivo de profesores en el proceso de planeación y elaboración de las preguntas e ítems, la forma de definir la clave de calificación y de determinar el éxito y el fracaso. La experiencia por parte del colectivo de profesores en la temática evaluación es esencial para el logro de un instrumento válido y confiable.

El colectivo podrá hacer las modificaciones y rectificaciones necesarias cuando se obtengan los resultados del instrumento, es por ello la importancia de brindar una metodología que permita determinar la calidad de los instrumentos evaluativos, que sirva de guía para el perfeccionamiento de este importante componente del proceso enseñanza aprendizaje.

MATERIAL Y MÉTODO

Se aplicaron los métodos teóricos y empíricos, con un profundo análisis documental de las variantes publicadas para evaluar la calidad de los instrumentos de evaluación, se sometió a análisis de informantes claves y se procedió a ajustar las técnicas y procedimientos a nuestras condiciones en la metodología propuesta.

RESULTADOS

Metodología propuesta:

I. Análisis del examen aplicado.

En este acápite se debe tener en cuenta: Tipo de examen, cantidad de preguntas del examen y temáticas que abordan, cantidad de ítem por preguntas y área temática, tipos de preguntas que contiene el instrumento, clave de calificación y decisión de éxito o fracaso, resultado obtenido en cada ítem, en términos de correcto e incorrecto y las calificaciones por preguntas y del examen en general.

II. Elaboración de la hoja de cálculo con los datos primarios de los resultados de cada estudiante en su examen.

En una hoja de cálculo de Microsoft Excel se tabulan los resultados del examen objeto de estudio. Las columnas recogerán los ítems de cada pregunta y a continuación la nota otorgada a la pregunta. A los ítems se le asigna la categoría de Correcto o Incorrecto. Al final, la última columna corresponde a la calificación obtenida en el examen. Las filas se asignan a los estudiantes, para lo cual es muy importante asignar en la primera columna una clave que permita identificar su grupo. Ello favorecerá el análisis por los colectivos docentes y unidades docentes que intervienen en la impartición de la asignatura.

III. Correspondencia del fondo de tiempo asignado en el P1 a los contenidos y la cantidad de ítems que lo exploran.

Se debe tabular los contenidos con el número de horas que se asignan en el modelo de planificación docente y a continuación la cantidad de ítems que exploran el contenido. En cada caso se obtiene el porcentaje correspondiente.

Este análisis puede hacerse también en función de áreas temáticas agrupando varios temas de interés y de igual manera se puede realizar por ciencias particulares en aquellas asignaturas que son integradas por más de una ciencia.

El resultado que se obtenga nos permitirá establecer la correspondencia entre el tratamiento de contenidos y su exploración en el examen. Debe encontrarse una proporción adecuada de ítems en función del tiempo que se le asignó al tratamiento del contenido.

En las asignaturas integradas por varias ciencias nos daría la proporción de cada una de ellas en el examen y su peso en tiempo.^{10, 11.}

Para concluir el análisis se puede aplicar el coeficiente de correlación de Spearman.¹²

IV. Definición de la escala de clasificación de la dificultad según expertos.

Para poder comparar los resultados y clasificar los ítems, preguntas y el examen, atendiendo al índice de dificultad, nos apoyaremos en los especialistas de la asignatura, que definirán los intervalos según la proporción de estudiantes aprobados para cada componente.

Se utiliza la propuesta de Bachkofft que define que el examen debe clasificarse en ítems o preguntas difíciles, medianamente difíciles, de dificultad media, medianamente fáciles y fáciles.

Un ejemplo pudiera ser:

- Menos de 40 % de aprobados se consideran difíciles.
- Entre 41 y 50 por cientos de aprobados se consideran medianamente difíciles.
- Entre 51 y 80 por cientos de aprobados se consideran de dificultad media.
- Entre 81 y 90 por cientos de aprobados se consideran medianamente fáciles.
- Entre 91 y 100 por cientos de aprobados se consideran fáciles.

Esta escala puede variar atendiendo al criterio de los especialistas de más experiencia de la asignatura, pero debe ser congruente con la calidad que se exige en los resultados de un examen. Para poder comparar posteriormente con otros instrumentos evaluativos, se debe mantener la misma clasificación.

V. Obtención de indicadores de validez del instrumento.

▪ Índice de dificultad.

○ De cada ítems.

Se determina el índice de dificultad de cada uno de los ítems, se divide el número de estudiantes que contestó correctamente el ítem entre el número total que contestaron el ítem. Esta proporción se le denota como Di y se utiliza la siguiente fórmula:^{8 - 11.}

$$Di = \frac{A}{N}$$

Donde: Di = Índice de dificultad del ítem. A = Número de respuestas correctas del ítem. N = Número total de estudiantes que contestaron el ítem.

○ **Índice de dificultad de las temáticas.**

En este caso se suma el total de estudiantes que respondieron correctamente cada uno de los incisos de la temática y se divide entre el número de incisos que explora el tema. Se calcula la media de respuestas correctas para el tema y se divide por el total de examinados, como se representa en la siguiente fórmula:

$$Dt = \frac{\sum(A_{t1} + A_{t2} + \dots + A_{tt})}{\frac{n_{tt}}{Nt}}$$

Donde: Dt = Índice de dificultad de la temática. A_t = Número de respuestas correctas de los incisos de la temática. n_{tt} = Número total de incisos que exploran la temática. Nt = Número total de presentados en el examen.

○ **Índice de dificultad de las preguntas.**

Se suman el total de estudiantes que respondieron correctamente cada uno de los ítems que corresponden a la pregunta y se divide entre el número de ítems de la pregunta. Se calcula la media de respuestas correctas para la pregunta y se divide por el total de examinados. Se utiliza la fórmula anterior aplicada a la pregunta como se muestra a continuación:

$$Dp = \frac{\sum(A_{p1} + A_{p2} + \dots + A_{pp})}{\frac{n_{tp}}{Nt}}$$

Donde: Dp = Índice de dificultad de la pregunta. A_p = Número de respuestas correctas de los incisos de la pregunta. n_{tp} = Número total de incisos de la pregunta. Nt = Número total de presentados en el examen.

○ **Índice de dificultad de las ciencias integradas.**

Se suman el total de estudiantes que respondieron correctamente cada uno de los ítems que corresponden a los contenidos particulares de una ciencia explorada en el examen y se divide entre el número de ítems de esa ciencia presente en el examen. Se calcula la media de respuestas correctas para la ciencia y se divide por el total de examinados. Se utiliza la fórmula anterior aplicada a la pregunta como se muestra a continuación:

$$Dc = \frac{\sum(A_{c1} + A_{c2} + \dots + A_{cp})}{\frac{n_{tc}}{Nt}}$$

Donde: Dc = Índice de dificultad del contenido de la ciencia. A_c = Número de respuestas correctas de los incisos que exploran la ciencia. n_{tc} = Número total de incisos que exploran la ciencia. Nt = Número total de presentados en el examen.

• **Comparación de los resultados según índice de dificultad obtenido (real) con la dificultad esperada.**

A partir de la obtención de los índices de dificultad, estamos en condiciones de clasificar cada ítem y pregunta o temática según la escala de clasificación definida en el colectivo. Estos resultados se tabulan. Para ello confeccionamos una tabla donde expresamos los resultados que debíamos esperar con los obtenidos en realidad según la clasificación y sus respectivos por cientos.

Escala de clasificación de los ítems según la dificultad esperada, considerada en la literatura sobre esta temática es como sigue:⁹

5 % para ítems fáciles.

20 % para ítems medianamente fáciles.

- 50 % para ítems de dificultad media.
- 20 % para ítems medianamente difíciles.
- 5 % para ítems difíciles.

Para ratificar el comportamiento de los ítems y preguntas podemos utilizar el Chi cuadrado de heterogeneidad, siempre y cuando no existan valores menores de 5, si esto ocurriese se puede aplicar el test de Kolmogorov-Smirnov como test no paramétrico que es menos restrictivo.¹²

▪ **Análisis de la discriminación de los ítems y preguntas.**

La discriminación, conocida también como poder de discriminación, expresa la medida en que el ítem o pregunta discrimina entre los buenos estudiantes y los peores estudiantes. Se entiende que los buenos estudiantes deben haber respondido en su mayoría correctamente el ítem y los estudiantes malos respondieron incorrectamente el ítem.

Este indicador nos permite además predecir el comportamiento del ítem o pregunta dentro del examen, para la elaboración de nuevos instrumentos con estructura bastante similar.

Para el estudio de este indicador podemos utilizar el Índice de discriminación y el Coeficiente de discriminación.^{9 - 11.}

• **Índice de discriminación.**

Para poder aplicar este indicador debemos organizar los datos, de la siguiente forma, en una columna se ponen los resultados de la pregunta y en la columna siguiente los resultados del ítem. A continuación se ordena de la nota mayor a la nota menor la columna de la pregunta, siempre se debe aceptar la opción “ampliar la selección” de la aplicación Excel. Con esta organización determinamos que número de estudiantes representa el 27 % de los examinados en ambos extremos.

Procedemos entonces a seleccionar el grupo de mejores notas (en sentido descendente a partir de la más alta) y el grupo de notas más bajas (en sentido ascendente a partir de la más baja).

La fórmula para su cálculo es:

$$D_i = \frac{GA_{aciertos} - GB_{aciertos}}{N_{grupomayor}}$$

Donde:

Di: Índice de discriminación del ítem i

GA_{aciertos}: Número de aciertos en el ítem i del 27% de personas con las puntuaciones más altas en el test.

GB_{aciertos}: Número de aciertos en el ítem i del 27% de personas con las puntuaciones más bajas en el test.

N_{grupomayor}: Número de personas en el grupo más numeroso (GA o GB).

Entre más alto es el índice de discriminación, el reactivo diferenciará mejor a las personas con altas y bajas calificaciones. Si todas las personas del GA contestan correctamente un reactivo y todas las personas del GB contestan incorrectamente, entonces D = 1 (valor máximo de este indicador); si sucede lo contrario, D = -1 (valor máximo negativo); si ambos grupos contestan por igual, D = 0 (valor mínimo de discriminación).

Algunos autores han definido la siguiente regla para determinar la calidad de los reactivos, en términos del índice de discriminación. A continuación se muestran los valores D y su correspondiente interpretación. Asimismo, se señalan las recomendaciones para cada uno de estos valores.

D	Calidad	Recomendaciones
> 0,39	Excelente	Conservar.
0,30 – 0,39	Buena	Posibilidades de mejorar.
0,20 – 0,29	Regular	Necesidad de revisar.
0,00 – 0,19	Pobre	Descartar o revisar a profundidad.
< -0,01	Pésima	Descartar definitivamente.

Debemos tomar en consideración que el máximo valor del índice de discriminación de un ítem tiene estrecha relación con su índice de dificultad, ya que un reactivo con valor muy alto o muy bajo de índice de dificultad en realidad no discrimina, pues es solucionado por la mayoría de los estudiantes o por la minoría de ellos respectivamente.

- **Coefficiente de discriminación del punto biserial (r_{pbis}).**

Este indicador permite calcular la efectividad discriminativa de un reactivo, entiéndase por ello un ítem o una pregunta.

La correlación del punto biserial (r_{pbis}) se emplea para saber si los estudiantes aplicados son los que obtienen las respuestas correctas, qué tanto poder predictivo tiene el reactivo, y cómo puede contribuir a las predicciones.¹⁹

Se utiliza la siguiente fórmula:
$$r_{pbis} = \frac{\overline{X}_1 - \overline{X}_0}{S_x} * \sqrt{\frac{n_1 \cdot n_0}{n(n-1)}}$$

Donde:

\overline{X}_1 = Media de las puntuaciones totales de aquellos que respondieron correctamente el ítem.

\overline{X}_0 = Media de las puntuaciones totales de aquellos que respondieron incorrectamente el ítem.

S_x = Desviación estándar de las puntuaciones totales.

n_1 = Número de estudiantes que respondieron correctamente el ítem.

n_0 = Número de estudiantes que respondieron incorrectamente el ítem.

$n = n_1 + n_0$

Según estándares internacionales la distribución de los resultados del coeficiente discriminación punto de correlación biserial (r_{pbis}) es:

- < 0 discriminan negativamente.
- 0 - 0.14 discriminan pobremente.
- 0.15 - 0.25 discriminan regular.
- 0.26 - 0.35 buen poder discriminativo.
- > 0.35 excelente poder de discriminación.

El r_{pbis} puede ser calculado para cada ítem y para cada pregunta.

VI. Análisis de la confiabilidad de las preguntas y el examen.

Existe una variedad de indicadores que se utilizan para evaluar la confiabilidad de un instrumento evaluativo. En nuestro caso proponemos usar el coeficiente de alfa de Cronbach que permite determinar la consistencia interna de un instrumento y brinda información relacionada con el aporte que hace cada pregunta a la consistencia general del examen.

- **Coefficiente α de Cronbach.**

La fórmula que se emplea es la siguiente:
$$\alpha = \frac{nr}{1 + r(n-1)}$$

Donde n es el número de elementos del examen y r su correlación media.

Este coeficiente se puede obtener utilizando las aplicaciones estadísticas profesionales como el SPSS y el SyStat.

El alfa de Cronbach toma un valor positivo. Para muchos autores un valor aceptable para un examen es de 0,60 a 0,70.²²

Los resultados deben tabularse atendiendo al alfa del examen y de cada pregunta, que se calcula excluyendo los datos de esa pregunta.

Su interpretación se realiza tomando en cuenta como aumenta o disminuye el alfa sin la pregunta, si el valor es igual o inferior al alfa del examen la pregunta aporta consistencia interna al examen, pues al eliminarla el coeficiente desciende, en caso contrario, si el valor del alfa del examen se eleva al eliminar la pregunta, es un indicador de que la pregunta afecta la consistencia interna y disocia los resultados del examen.

CONCLUSIONES

La calidad de un instrumento depende en gran medida de la planeación que se realice con suficiente antelación y constante revisión y actualización.

La aplicación de la metodología propuesta tiene como objetivo ayudar a la realización de un análisis científico y normalizado, que permita determinar los contenidos con resultados deficientes y establecer su tratamiento posterior.

El resultado final del análisis de calidad de un instrumento debe incluir la valoración que realicen los profesores de los colectivos de las asignaturas en su planeación y los resultados de la aplicación de la metodología propuesta.

REFERENCIAS

1. Álvarez de Zayas C. “La dirección del proceso docente educativo”. Tomado de: Hacia una escuela de Excelencia. La Habana: Editorial Academia; 1996.
2. Salas Perea RS. La evaluación en la educación superior contemporánea. Biblioteca de Medicina. Volumen XXIV. UMSA, La Paz, Bolivia, 1998. En CD-ROM Maestría en Educación Médica, ENSAP; 2005.
3. Reglamento de Trabajo Docente y Metodológico. Ministerio de Educación Superior. Resolución 210; 2007. Gaceta Oficial de la República de Cuba. 2007.
4. González Pérez M. La evaluación del aprendizaje.- tendencias y reflexión crítica. Revista Cubana de Educación Superior, Vol. XX, N° 1, 2000:47-62
5. Salas Perea R. Planeamiento de un examen. Biblioteca de Medicina. Volumen XXIV. UMSA, La Paz, Bolivia, 1998. En CD-ROM Maestría en Educación Médica, ENSAP; 2005.
6. Vicedo Tomey A. La integración de conocimientos en la educación médica. Rev Cubana Educ. Med Sup. 2009; 23(4).
7. Segura Castillo, M. La perspectiva ética de la evaluación de los aprendizajes desde un enfoque constructivista. Rev Actualidades Investigativas en Educación [Internet]. 2007 ene-abr [Citado 27 septiembre 2011]; 7(1). [Aprox. 22 p]. Disponible desde: dialnet.unirioja.es/servlet/articulo?codigo=2273870
8. Pérez Tapia JH, Acuña Aguila N, Arratia Cuela E. Nivel de dificultad y poder de discriminación del tercer y quinto examen parcial de la cátedra de cito-histología 2007 de la carrera de medicina de la UMSA. Rev Cuadernos [Internet]; 2008 [citado 22 abril 2011]; 53(2): [Aprox. 14 p]. Disponible desde: www.revistasbolivianas.org.bo/pdf/chc/v53n2/v53n2a03.pdf
9. Backhoff E, Larrazolo N, Rosas M. Nivel de dificultad y poder de discriminación del Examen de Habilidades y Conocimientos Básicos (EXHCOBA). Revista Electrónica de Investigación Educativa [Internet]; 2000. [citado 2 de octubre de 2011]; 2 (1): [Aprox. 14 p]. Disponible en: <http://redie.uabc.mx/vol2no1/contenido-backhoff.html>.
10. Domínguez Mendoza TN, Godoy Calderón S. Recomendaciones metodológicas para el análisis de reactivos de opción múltiple en exámenes que integran diversas áreas del conocimiento. Instituto Politécnico Nacional. Ciudad México. 2008.

11. Caraballo Carballo RM, Rico Romero L, Lupiáñez Gómez JL. Análisis de los ítems de las evaluaciones autonómicas de diagnóstico en España: 2008-2009. Rev Iberoamericana de Educación Matemática [Internet]; 2011 Junio [citado 2 de octubre de 2011]; (26): [Aprox. 16 p]. Disponible en: http://www.fisem.org/web-/union/index.php?option=com_content&view=article&id=32&Itemid=3
12. Díaz Rojas P. Introducción a la Investigación en Ciencias de la Salud. La Habana: Escuela Nacional de Salud Pública.; 2010. ISBN: 978-759-7158-47-7.