TEMA DE ACTUALIDAD

Criterios de calidad de los instrumentos evaluativos escritos

Quality criteria of written evaluative instruments

Dr. Raúl Martínez Pérez

Especialista de I y II grado en Angiología. Máster en Urgencias Médicas. Investigador y Profesor Auxiliar Facultad de Ciencias Médicas Mayabeque. Güines, Cuba. Correo electrónico: raulmart@infomed.sld.cu

Autor para la correspondencia. Dr. Raúl Martínez Pérez. Correo electrónico: raulmart@infomed.sld.cu

RESUMEN

Constantemente se emiten juicios y decisiones sobre los estudiantes basados en varios tipos de evaluación. Para hacer juicios imparciales, se debe entender cuidadosamente las fortalezas y limitaciones de las herramientas y procesos de evaluación sobre el cual se basan estas decisiones. En realidad, se requiere de evidencia para respaldar la validez de nuestras decisiones. Es necesario conocer los indicadores técnicos que definen la calidad del instrumento evaluativo que se emplea por su dimensión y por el poderoso impacto social que tiene, pues los defectos en la elaboración de los mismos tienen efectos negativos en los estudiantes. Con el objetivo de presentar algunos criterios que en la actualidad predominan en relación con la calidad de los instrumentos evaluativos escritos, se realizó una búsqueda bibliográfica a través de Infomed, Medline, SciELO y Redalyc.y Google Scholar. Se concluye que el actual modelo de Validez, por su carácter holístico, procesal, dinámico y abierto, resulta esencial como criterio de calidad en la evaluación. Cuando se emplean instrumentos escritos, el contenido de la prueba y su estructura interna constituyen las más importantes fuentes de evidencia para argumentar su validez.

Palabras clave: calidad de exámenes escritos; confiabilidad; validez
Descriptores: evaluación educacional**;** control de calidad; reproducibilidad de los resultados

ABSTRACT

Professors are constantly expressing judgments and decisions about students based on several types of assessments. To make impartial judgments, strengths and limitations of instruments and processes of evaluation must be carefully understood because they are based on those decisions. It is really required the evidences to Support the validity of the decisions. It is necessary to know the technical indicators that define the quality of the evaluative instruments which are used by its dimension and by the powerful social impact they have, because the defects in their elaboration have negative effects in students. A

library research by Infomed, Medline, SciELO y Redalyc.and Google Scholar was carried out with the objective to present several criteria that currently prevail in relation to the quality of the written evaluative instruments. It can be concluded that the current model of Validity, for been a process and for its holistic, dynamic and open character, results essential as a criterion of quality in evaluation.

Key words: written test quality; reliability; validity
Descriptor: educational measurement, quality control; reproducibility of results

Historial del trabajo.
Recibido: 21/12/2019
Aprobado:18/03/2020

INTRODUCCIÓN

La educación médica superior debe enfrentar los múltiples retos que implica el desarrollo del proceso de enseñanza-aprendizaje de un complejo conjunto de disciplinas, abarcadoras de una amplia gama de contenidos constituidos por diversos conocimientos, habilidades, actitudes y valores.

El proceso de enseñanza aprendizaje constituye un sistema en el que cada uno de sus componentes se interrelacionan, donde la evaluación es el mecanismo regulador del sistema y puede considerarse como el instrumento de control de la calidad del producto resultante del proceso docente educativo.[(1)]

En la educación superior cubana está reglamentado que la evaluación del aprendizaje es un proceso consustancial al desarrollo del proceso docente educativo que tiene como propósito comprobar el grado de cumplimiento de los objetivos formulados en los planes de estudio, mediante la valoración de los conocimientos y habilidades que los estudiantes van adquiriendo y desarrollando, así como por la conducta que manifiestan en el proceso docente educativo.[(2)]

La evaluación a su vez, constituye una vía para la retroalimentación y la regulación de dicho proceso ya que le permite al profesor indagar sobre el grado de aprendizaje y desarrollo de los estudiantes en su proceso de formación, así como la capacidad que poseen para aplicar los contenidos en la resolución de problemas de la profesión, brindándole información oportuna y confiable para descubrir aquellos elementos de su práctica que interfieren en los procesos de enseñanza y aprendizaje, de tal manera que pueda reflexionar en torno a estos para mejorarlos y reorientarlos permanentemente.[(2)]

La evaluación educativa, como proceso, involucra la elaboración, aplicación y análisis de los instrumentos de medición. La función principal de un instrumento de medición educativa, cuando se crea como medida para inferir las capacidades de las personas, es ofrecer información para la correcta toma de decisiones.

Constantemente se emiten juicios y decisiones sobre los estudiantes, basados en varios tipos de evaluación. Para hacer juicios imparciales, se tiene que entender cuidadosamente las fortalezas y limitaciones de las herramientas y procesos de evaluación sobre el cual se basan estas decisiones. En realidad, se requiere de evidencia para respaldar la validez de nuestras decisiones.[3]

Cuando se utilizan instrumentos de alto impacto, como es el caso de los exámenes finales de las asignaturas, es necesario conocer los indicadores técnicos que definen la calidad del instrumento evaluativo que se emplea. Por su dimensión y por el poderoso impacto social que tienen, la elaboración de exámenes debe ajustarse a rigurosos estándares de calidad, pues los defectos en la elaboración de los mismos tienen efectos negativos en la calificación de los estudiantes.[4
]

Con el objetivo de presentar algunos criterios que en la actualidad predominan en relación con la calidad de los instrumentos evaluativos escritos, se realiza una amplia búsqueda de bibliografía relacionada con el tema a través de la red de Infomed, Medline, SciELO y Redalyc y con el motor de búsqueda de Google Scholar, seleccionando aquella publicada en revistas de impacto y por instituciones o autores reconocidos.

DESARROLLO

Las evaluaciones realizadas a través de instrumentos escritos tienen una larga historia en los distintos niveles educacionales, siendo utilizados como recurso diagnóstico o formativo y, además, como requisito para admisión, acreditación o premio.

Los mismos poseen importantes implicaciones educativas, manifiestas en los estudiantes, los profesores, el currículo (formal y oculto) y en el proceso docente educativo, las que son generadas por sus efectos potenciales positivos, como la motivación para estudiar, la mejora de la calidad educativa y el desarrollo profesional. Pero también por efectos potenciales negativos, que pueden distorsionar el proceso educativo y las prioridades de estudiantes y profesores.[5- 9]

Estas implicaciones suscitan el interés de profesores e investigadores en perfeccionar la metodología de elaboración de exámenes y los conceptos de validez en evaluación.

El examen se define como ¨recurso o procedimiento en el que una muestra sistemática de una conducta del sustentante, en un dominio específico es obtenida y calificada utilizando un proceso estandarizado¨,[10] y también como ¨instrumento de evaluación que se emplea para identificar el nivel de dominio de los sustentantes sobre un constructo específico¨.[11]

Los exámenes escritos se clasifican en dos grandes grupos, según los tipos de preguntas o ítems que los constituyen: Examen tipo ensayo y examen objetivo. Aunque pude considerarse un tercer grupo, mixto, formado por la combinación de ambos tipos.[7, 8]

Los exámenes tipo ensayo están constituidos por preguntas o ítems de respuestas abiertas, no estructuradas, en las que el examinado debe redactar las respuestas. Este tipo de

examen tiene la ventaja de que son más fáciles de confeccionar y permiten evaluar habilidades intelectuales de alto nivel y la capacidad para expresar las ideas de forma escrita, aunque presentan las desventajas de que limita sensiblemente la extensión del contenido a evaluar, el análisis estadístico de los resultados es complejo y es difícil lograr la objetividad en su calificación, la cual consume mucho tiempo, por lo que restringe su utilización en la evaluación de grupos numerosos. Para tratar de reducir estas desventajas se utilizan variantes como el ensayo de respuesta corta y el ensayo modificado.[8,12]

Los exámenes objetivos están integrados por ítems o reactivos de respuesta cerrada, estructurada, en las que el examinado debe escoger la respuesta entre el conjunto de opciones que se le proporcionan. Tienen las ventajas de que pueden evaluar un área más extensa de los contenidos, el análisis estadístico de los resultados es relativamente menos complejo, la calificación es objetiva y se realiza con rapidez, incluso, se pueden emplear medios automatizados para ello, por lo que son los más utilizados, sobre todo cuando es necesario examinar grandes grupos. Presentan la desventaja de que su elaboración exige mucho tiempo, pericia y experiencia, además, resulta complejo elaborar ítems que exploren habilidades intelectuales superiores.[12]

Aunque se describe una amplia gama de formatos de ítems para test objetivos, en la actualidad se prefieren los de selección múltiple de complemento simple, por ser los que aportan mayor evidencia de validez en relación con la estructura interna del examen.[5,7,13]

Reconocidas organizaciones e investigadores en el desarrollo de instrumentos de evaluación del aprendizaje, han propuesto una serie de principios y mejores prácticas para diseñar exámenes con alta calidad.[4,7,10-12]

Múltiples son las características generales, atributos, criterios o estándares que han sido señalados para asegurar la calidad de los instrumentos evaluativos escritos, entre los que se pueden mencionar: utilidad, factibilidad, equidad, exactitud, pertinencia, equilibrio, eficacia, objetividad, pero sobre todos destacan la confiabilidad y la validez.[5,14,15]

La confiabilidad y la validez constituyen atributos fundamentales que se exigen en relación con los instrumentos evaluativos, pero es importante enfatizar que no son características de estos. La confiabilidad y la validez corresponden a propiedades de las interpretaciones, inferencias o usos específicos de las medidas que los test proporcionan[14] y, aunque están estrechamente relacionadas, con frecuencia se asigna a la confiabilidad un papel más importante del que realmente tiene, incluso separándola de la validez.[16]

Un instrumento que cumpla con las exigencias de la validez, tiene un alto grado de probabilidad de ser confiable, sin embargo, no necesariamente ocurre así a la inversa.[8]La confiabilidad es una cuestión relativa a la calidad de los datos, es una característica de unos resultados, de unas puntuaciones obtenidas en una muestra determinada, mientras que la validez se refiere a la calidad de la inferencia que se realiza a partir de esos datos.[14,17]

Un instrumento puede ser válido, porque mide lo que decimos que mide y queremos medir, pero lo puede medir con un margen de error grande, es decir, baja confiabilidad. Con

instrumentos parecidos o en mediciones sucesivas hubiéramos obtenido resultados distintos. También puede haber una confiabilidad alta, indicando que los sujetos están clasificados u ordenados con poco margen de error y, a la vez, el instrumento carecer de validez, porque no mide lo que se pretende o lo que se dice que se está midiendo.[17]

En el cálculo de la confiabilidad hay tres enfoques o métodos que, aunque parten de modelos teóricos idénticos o parecidos, siguen procedimientos distintos:
a) el test-retest
b) las pruebas paralelas
c) los coeficientes de consistencia interna**.**

Los coeficientes de consistencia interna son los más utilizado, por lo que cuando se habla de confiabilidad, sin especificar, hay que entender que se trata de confiabilidad en el sentido de consistencia interna.

Lo que expresan directamente estos coeficientes es hasta qué punto las respuestas son lo suficientemente coherentes (relacionadas entre sí) como para poder concluir que todos las preguntas o ítems miden un mismo rasgo y, por tanto, son sumables en una puntuación total única que representa y mide dicho rasgo.[17, 18]

Se trata de coeficientes de correlación que teóricamente significan la correlación del test consigo mismo, que pueden tomar valores entre 0 y 1, donde 0 significa confiabilidad nula y 1 representa confiabilidad total.[19]

Cuando se trata de preguntas o ítems con una escala de calificación politómica, es común emplear el coeficiente alfa de Cronbach. En cuestionarios de ítems dicotómicos y cuando existen alternativas dicotómicas con respuestas correctas e incorrectas se debe utilizar el coeficiente de confiabilidad de Kuder-Richarson.[19]

Existen tres factores que inciden en la magnitud del coeficiente de confiabilidad:

1. La homogeneidad de los ítems. En la medida en que los ítems midan el mismo rasgola confiabilidad será mayor, con preguntas muy distintas y poco relacionadas entre sí la confiabilidad será más baja.
2. Las diferencias entre los examinados (homogeneidad de la muestra). Si los sujetos tienen resultados muy parecidos la fiabilidad bajará. No se puede clasificar, ordenar bien, a los muy semejantes.
3. El número de ítems. A mayor número de ítems los examinados quedan mejor diferenciados.

Fundamentalmente, la confiabilidad depende de las diferencias entre los sujetos, por lo que se puede cuestionar la confiabilidad de un examen como indicador necesario de su calidad. Si todos saben todo o casi todo (o casi nada), la confiabilidad tiende a bajar y esto no quiere decir que el examen sea deficiente o que se trate de un mal resultado. Un coeficiente de confiabilidad alto es claramente deseable cuando las diferencias entre los sujetos son legítimas y esperadas, que es lo que suele suceder en los exámenes finales, sobre todo si poseen un alto número de ítems y, más aún, en una elevada cantidad de evaluados, donde

es razonable esperar diferencias en rendimiento. Una confiabilidad alta nos dice que el examen deja a cada uno en su sitio y que, en exámenes parecidos, con otras preguntas del mismo estilo, los alumnos quedarían ordenados de manera semejante.[17]

No obstante, se debe señalar que construir pruebas que sean confiables según el modelo de medida es garantizar su insensibilidad con respecto a lo que se ha enseñado. El tipo de fiabilidad educativamente relevante es la fiabilidad inter jueces (distintos jueces harían el mismo juicio acerca de la misma actuación en la misma ocasión) e intrajueces (la misma persona haría el mismo juicio acerca de la misma actuación en dos ocasiones diferentes).[20]

La validez.

En términos generales, el propósito de la validez es indagar si un examen mide lo que debería medir.[21] No obstante, la manera de definir la validez ha variado sustancialmente desde su florecimiento, a mediados del siglo XX, hasta nuestros días.

Durante la primera mitad del pasado siglo, los educadores inicialmente reconocieron dos tipos de validez: *validez de contenido*, que se relaciona con la creación de ítems de evaluación, y la *validez de criterio*, que se refiere a cómo los puntajes se correlacionan con una medida estándar de referencia del mismo fenómeno, dado por exámenes similares que evaluaban las mismas habilidades. Por tanto, un examen se consideraba como válido exclusivamente si medía lo mismo que otros instrumentos ya existentes.[3,14,16,21]

Sin embargo, la validez de contenido casi siempre respaldaba la prueba y se reconoció que identificar y validar un estándar de referencia es muy difícil, especialmente para atributos intangibles. Como una alternativa, los teóricos propusieron la validez de constructo, en la que los atributos intangibles (constructos) están vinculados con atributos observables basados en una concepción o teoría del constructo, por lo que quedaron establecidos tres tipos de validez:

- De constructo: ¿Con qué alcance el instrumento mide realmente un rasgo determinado y con cuánta eficiencia lo hace?
- De contenido: ¿los contenidos que evalúan los ítems o preguntas de un instrumento son representativos del universo de contenido de la característica o rasgo que se quiere medir?
- De criterio: ¿Cómo se correlacionan los resultados obtenidos con el instrumento con los de alguna medida del criterio u otro procedimiento aplicado antes (*validez retrospectiva*), durante (*validez concurrente*) o después (*validez predictiva*) de aplicar el instrumento?

Estos constituyeron los ejes fundamentales de la validez de los exámenes hasta finales del siglo XX.[3,14,16 21]

El esquema mencionado ha sido utilizado por varias décadas en la comunidad de educadores, pero tiene el inconveniente de que genera una separación artificial entre los diversos tipos de validez, dando la impresión que son cosas diferentes e independientes entre sí. Además, propicia el concepto erróneo de que los exámenes son válidos o inválidos

por sí mismos, implicando que la validez es una propiedad intrínseca del instrumento que pudiera ser transferible a otros contextos.[16]

A finales de los años 80 del pasado siglo se suscitó la problemática de que los diferentes métodos de validez eran tratados como diversos recursos que podían ser empleados en distintas situaciones, de acuerdo a como los evaluadores consideraban que fuera más conveniente emplearlos. El punto fundamental era que no se contaba con un marco general de validación para los exámenes, puesto que la validez de contenido y concurrente dependían del tipo de examen que se quería validar.[21] Se propuso entonces abandonar el modelo de los diferentes tipos de validez, para migrar a un marco conceptual unificado en el que toda la validez es validez de constructo, que se alimenta de diferentes fuentes.[3,13,16]

El actual concepto de validez se refiere al grado en el que la evidencia empírica y la teoría apoyan las interpretaciones de los resultados de un examen para los usos para los cuales fue propuesto.[3,10]

La validación es un proceso de acumulación de pruebas para apoyar la interpretación y el uso de las calificaciones. Por tanto, el objeto de la validación no es el test, sino la interpretación de sus puntuaciones en relación con un objetivo o uso concreto. El proceso de validación se concibe como un argumento que parte de una definición explícita de las interpretaciones que se proponen, de su fundamentación teórica, de las predicciones derivadas y de los datos que justificarían científicamente su pertinencia.[14]

Dado que las predicciones suelen ser múltiples, una única prueba no puede sustentar un juicio favorable sobre la validez de las interpretaciones propuestas. Son necesarias pruebas múltiples y convergentes obtenidas en diferentes estudios. Por ello, se considera que la validación es un proceso dinámico y abierto. Obviamente, los usos y las interpretaciones relacionadas pueden ser muy variados. Por ello, las fuentes de validación son múltiples y su importancia varía en función de los objetivos.[14]

Este modelo holístico de validez, también denominada validez argumentativa, ha sido ampliamente aceptado por la comunidad internacional de investigadores en evaluación, pasando a ser el concepto más importante en evaluación educativa, ya que permea por todos lados el proceso educativo y determina la congruencia interpretativa del uso de los resultados de exámenes, de acuerdo a los fines para los que fueron diseñados.[3,15,16,21]

Se describen cinco diferentes fuentes de evidencia para argumentar la validez del proceso de evaluación, relacionadas con:[4,14,16]

a) El contenido de la prueba: debe ser una muestra representativa del constructo o dominio de contenido definido en el programa educativo. Se centra en la relevancia y representatividad del contenido objeto de la evaluación, por lo que ha de reflejar fielmente los núcleos básicos de contenidos y sus esencialidades, con el nivel de asimilación establecido, existiendo una correspondencia entre el fondo de tiempo asignado en la planificación docente y la cantidad de ítems que los exploran.[8,15,22]

En relación con el nivel de asimilación, tradicionalmente las preguntas de examen se clasificaban en preguntas de retención, interpretación o resolución de problemas (memoria, comprensión y razonamiento), según los procesos cognitivos que se requerían para contestar la pregunta. Típicamente, las ¨preguntas de retención¨ son aquellas que evalúan el conocimiento que tiene el estudiante sobre definiciones o hechos. Las ¨preguntas de interpretación¨ requieren que los estudiantes analicen determinada información y lleguen a alguna conclusión. Las ¨preguntas de resolución de problemas¨ presentan una situación en la que los estudiantes deben tomar alguna decisión.[7]

La dificultad relacionada con estas clasificaciones es que los procesos cognitivos que se necesitan para contestar la pregunta dependen tanto de la formación del estudiante como del contenido de la pregunta. Además, la selección de los tipos de preguntas depende de su propósito de aplicación: para una evaluación sumativa, el uso de preguntas que requieren habilidades de pensamiento de orden superior y la aplicación de conocimientos sería más recomendable que las preguntas simples de retención. El uso de preguntas de retención puede ser de mayor utilidad para fines de evaluación formativa.[7]

Los procesos cognitivos que se requieren para contestar una pregunta son específicos a cada estudiante, por lo cual este enfoque taxonómico resulta difícil de usar. Un enfoque alternativo, divide a las preguntas en dos categorías: la aplicación de conocimientos frente a la retención de contenidos.[7]

Otro elemento fundamental concerniente al contenido de la prueba lo constituye la calidad técnica de la elaboración de los ítems en lo relacionado con el formato, el estilo y la redacción de cada uno.[4,7,11]

b) El proceso de respuesta: debe incluir el análisis de los procesos, las estrategias de resolución de problemas y las representaciones mentales que emplean los participantes para resolver los ítems. Se obtendrá evidencia de validez cuando los procesos utilizados se ajustan a los que se postulan en las teorías relativas al constructo medido. La metodología de estudio es muy diversa: entrevistas a los examinados para que describan cómo resuelven las tareas, análisis de los movimientos oculares, los tiempos de respuesta y otros.

c) La estructura interna: El análisis de la estructura interna persigue verificar empíricamente si los ítems se ajustan a la dimensionalidad prevista en la elaboración dela prueba. Se refiere a las características estadísticas o psicométricas del instrumento, de las preguntas o ítems que lo conforman y de las respuestas.[13]

La psicometría incluye un conjunto sistemático de análisis de la dificultad, de la discriminación, de las opciones de respuesta (para ítems objetivos) y de la confiabilidad. Su realización se fundamenta en dos enfoques principales: la teoría clásica de los test (TCT) y la teoría de respuesta al ítem (TRI).[5-8, 22-24]
El primero, es un conjunto articulado de procedimientos desarrollados fundamentalmente en la primera mitad del siglo pasado, que se conoce también como modelo de la puntuación verdadera o teoría del error de medición. Implica la diferenciación de los conceptos

¨puntuación verdadera¨ y ¨puntuación observada¨ como resultado de la aplicación de una prueba. Se asume que la ¨puntuación verdadera¨ de una persona no cambia entre ocasiones, por lo que la variabilidad de las ¨puntuaciones observadas¨ se debe a la influencia de un error de medida aleatorio, no sistemático (producido por causas desconocidas e incontrolables en esa situación). La cantidad de error en cada caso sería la diferencia entre una ¨puntuación observada¨ y la ¨puntuación verdadera¨

La TRI, por su parte, cuyos fundamentos comienzan a ser publicados en los años 80 del siglo pasado, conocida inicialmente como Teoría del rasgo latente, aproxima el análisis de las respuestas en una prueba de forma radicalmente diferente, enfocándose en los componentes constituyentes de la misma, es decir los ítems, en vez del resultado global de la medición. Está dirigida a estimar estadísticamente, de manera independiente, los parámetros de las personas y de los ítems en un continuo latente a partir de las respuestas observables.[13, 23-25]

A pesar de que entre los expertos en psicometría existe consenso general sobre la superioridad teórica de la TRI, el enfoque principal en contextos aplicados para el análisis de los resultados de los test continúa siendo la TCT. Específicamente en el área de la educación médica, son escasos los estudios que analizan los datos de instrumentos de evaluación dentro del marco de la TRI.[13, 23-25]

d) Su relación con otras variables: Las relaciones de las calificaciones del test con otras variables externas a la prueba constituyen una importante fuente de validación por cuanto se trata de justificar la utilidad de la prueba para predecir un criterio (medida de la variable de interés).

La utilidad de la prueba se suele cuantificar mediante la correlación entre sus puntuaciones y las de alguna medida del criterio (*coeficiente de validez*), o mediante otros procedimientos, tales como diferencia en las puntuaciones entre grupos de distinto nivel en el criterio, grado de acuerdo en las clasificaciones en categorías diagnósticas realizadas mediante el test y expertos, y otros.

La elección de un criterio fiable y válido (suficiente, objetivo y representativo de la conducta de interés) es el punto crítico que determina la bondad del proceso de validación. En función del momento temporal en el que se evalúa el criterio, se distinguen distintos tipos de recogida de datos: retrospectiva (el criterio se ha obtenido antes de administrar el test), concurrente (las puntuaciones del test y del criterio se obtienen en la misma sesión) y predictiva (el criterio se mide en un momento posterior).

e) Las consecuencias para la persona que es objeto de la evaluación: plantea la previsión de las posibles consecuencias del uso de los test como parte del proceso de validación. Desde esta perspectiva, el análisis y justificación de las consecuencias ocupan un lugar preponderante cuando los test vayan a emplearse para tomar decisiones críticas para personas e instituciones: selección, promoción, graduación, categorización profesional, evaluación de programas, etc. La literatura psicométrica denomina estos usos como de alto riesgo o alto impacto. En estos casos, la pertinencia del uso no se limita a la comprobación

de que las puntuaciones representan adecuadamente los constructos y a la justificación teórica de la red nomológica que vincula los constructos con los criterios de interés, por cuanto las aplicaciones de alto riesgo tienen efectos colaterales de carácter personal y social.

CONSIDERACIONES FINALES

Por el poderoso impacto de la evaluación sobre estudiantes, profesores y el propio proceso docente educativo -del cual es elemento constitutivo sustancial- es ineludible ajustarse a rigurosos estándares de calidad, para lo cual el actual modelo de Validez, por su carácter holístico, procesal, dinámico y abierto, resulta esencial.

Cuando para fines evaluativos se emplean instrumentos escritos, el contenido de la prueba y su estructura interna constituyen las más importantes fuentes de evidencia para argumentar la validez de los mismos.

REFERENCIAS BIB LIOGRÁFICAS

1.Salas Perea RS, Salas Mainegra A. Modelo formativo del médico cubano. Bases teóricas y metodológicas [Internet]. La Habana: Editorial Ciencias Médicas; 2017 [citado 6 Mar 2020]. Disponible en: http://www.bvs.sld.cu/libros_texto/modelo_formativo_medico_cubano/modelo_formativo.pdf

2.Ministerio de Educación Superior. Resolución Ministerial No. 02. Reglamento de trabajo docente y metodológico en la educación superior [Internet]. La Habana: MES; 2018 [citado 6 Mar 2020]. Disponible en: https://instituciones.sld.cu/faenflidiadoce/files/2018/08/Resoluci%c3%b3n-2-del-2018.pdf

3.Cook DA, Brydges R, Ginsburg S, Hatala R. A contemporary approach to validity arguments: a practical guide to Kane's framework. Medical Education [Internet]. 2015 [citado 6 Mar 2020];49:560–75. Disponible en: https://onlinelibrary.wiley.com/doi/full/10.1111/medu.12678

4.Rivera Jiménez J, Flores Hernández F, Alpuche Hernández A, Martínez González A. Evaluación de reactivos de opción múltiple en medicina. Evidencia de validez de un instrumento. Inv Ed Med [Internet]. 2016 [citado 6 Mar 2020];6(21):8-15. Disponible en: https://www.researchgate.net/publication/303808901_Evaluacion_de_reactivos_de_opcion_multiple_en_medicina_Evidencia_de_validez_de_un_instrumento/link/583ce4ad08ae3cb63655973c/download

5.Rosales FA. Capacidad de discriminación de las preguntas de un examen escrito. Rev Agron Noroeste Argent [Internet]. 2014 [citado 6 Mar 2020];34(2):94-6. Disponible en: https://studylib.es/doc/6738850/capacidad-de-discriminaci%C3%B3n-de-las-preguntas-de-un-examen...

6.Blanco Pereira ME, Martínez L, González Gil A, Jordán Padrón M. Calidad del examen final teórico de Morfofisiología Humana I en la Facultad de Ciencias Médicas de Matanzas. Cursos 2012-2013 y 2013-2014. Rev Med Electrón [Internet]. 2015 Ago [citado 6 Mar 2020];37(4):323-32. Disponible en: http://scielo.sld.cu/scielo.php?script=sci_arttext&pid=S1684-18242015000400003&lng=es.

7.Paniagua MA, Swygert KA. Cómo elaborar preguntas para evaluaciones escritas en las áreas de ciencias básicas y clínicas [Internet]. Philadelphia: National Board of Medical Examiners (NBME); 2016 [citado 6 Mar 2020]. Disponible en: http://sedici.unlp.edu.ar/handle/10915/8529

8.Díaz Rojas PA, Leyva Sánchez E. Metodología para determinar la calidad de los instrumentos de evaluación. Educ Med Super [Internet]. 2013 [citado 6 Mar 2020];27(2):269-86. Disponible en: http://scielo.sld.cu/scielo.php?script=sci_arttext&pid=S0864-21412013000200014&lng=es

9.Sánchez-Mendiola M, Delgado-Maldonado L. Exámenes de alto impacto: implicaciones educativas. Investigación Educ Médica [Internet]. 2017 Mar [citado 6 Mar 2020];6(21):52-62. Disponible en: http://www.scielo.org.mx/scielo.php?script=sci_arttext&pid=S2007-50572017000100052&lng=es

10.American Educational Research Association. Standards for educational and psychological testing [Internet]. Washington, DC: AERA; 2014 [citado 6 Mar 2020]. Disponible en: https://eric.ed.gov/?id=ED565876

11.Instituto Nacional para la Evaluación de la Educación. Criterios técnicos para el desarrollo y uso de instrumentos de evaluación educativa 2014-2015 [Internet]. Mexico, DF: INEE[; 2014 [citado 6 Mar 2020]. Disponible en: https://www.inee.edu.mx/wp-content/uploads/2019/02/CRITERIOS_TECNICOS_PARA_EL_DESARROLLO_Y_USO_DE_INSTRUMENTOS_10_ABRIL_2014.pdf

12.Moya Ricardo D, Pérez Gómez Y, Ruiz Cordovéz R. La Teoría de Respuesta al Ítem para la evaluación del aprendizaje en Matemática. Edu Sol [Internet]. 2016 [citado 6 Mar 2020];16(55):92-104. Disponible en: https://www.redalyc.org/jatsRepo/4757/475753050020/475753050020.pdf

13.Jurado Núñez A, Flores Hernández F, Delgado Maldonado L, Sommer Cervantes H, Martínez González A, Sánchez Mendiola M. Distractores en preguntas de opción múltiple para estudiantes de medicina: ¿cuál es su comportamiento en un examen sumativo de altas consecuencias? Inv Ed Med [Internet]. 2013 [citado 6 Mar 2020];2(8):202-10. Disponible en: https://www.sciencedirect.com/science/article/pii/S2007505713727133

14.Prieto G, Delgado AR. Fiabilidad y validez. Papeles del Psicólogo [Internet]. 2010 [citado 6 Mar 2020];31(1):67-74. Disponible en: www.papelesdelpsicologo.es/pdf/1797.pdf

15.González Machado EC, Reyes Piñuelas EP, López Ortega M. Construcción de una prueba para evaluar aprendizajes en educación superior [Internet]. San Luis de Postosí: Congreso nacional de investigación educativa; 2017 [citado 6 Mar 2020]. Disponible en: http://www.comie.org.mx/congreso/memoriaelectronica/v14/doc/2300.pdf

16.Sánchez Mendiola M. Mi instrumento es más válido que el tuyo: ¿Por qué seguimos usando ideas obsoletas? Inv Ed Med [Internet]. 2016 [citado 6 Mar 2020];5(19):133-5. Disponible en: https://www.elsevier.es/es-revista-investigacion-educacion-medica-343-articulo-mi-instrumento-es-mas-valido-S2007505716300308

17.Morales Vallejo P. Estadística aplicada a las Ciencias Sociales. La fiabilidad de los test y escala [Internet]. Madrid: Universidad Pontificia Comillas; 2007 [citado 6 Mar 2020]. Disponible en: http://www.upco.es/personal/peter/estadisticabasica/Fiabilidad.pdf

18.Zamora Araya JA. Análisis de la confiabilidad de los resultados de la prueba de diagnóstico matemática en la Universidad Nacional de Costa Rica utilizando el modelo de Rasch. Actualidades en Psicología [Internet]. 2015 [citado 6 Mar 2020];29(119):153-65. Disponible en:

https://www.researchgate.net/publication/326697432_Analisis_de_la_confiabilidad_de_los_resultados_de_la_Prueba_de_Diagnostico_Matematica_en_la_Universidad_Nacional_de_Costa_Rica_utilizando_el_modelo_de_Rasch/link/5b5fdf1e458515c4b254446e/download
19.Corral Y. Validez y confiabilidad de los instrumentos de investigación para la recolección de datos. Revista Ciencias Educación [Internet]. 2009 [citado 6 Mar 2020];19(33):228-47 https://www.researchgate.net/publication/302415291_Validez_y_confiabilidad_de_los_instrumentos_de_investigacion_para_la_recoleccion_de_datos
20.Biggs, J. Calidad del aprendizaje universitario. Madrid: Narcea; 2005.
21.Mendoza Ramos A. La validez en los exámenes de alto impacto. Un enfoque desde la lógica argumentativa. Perfiles Educativos [Internet]. 2015 [citado 6 Mar 2020];37(149):169-86. Disponible en: http:// www.iisue.unam.mx/perfiles/descargas/pdf/2015-149-169-86
22.Sánchez Hernández E, Medina Pavón M, Rodríguez García M, Vega Van Der Meer L, De la Torre Vega G. Indicadores de calidad para un examen teórico de la especialidad de medicina general integral. Medisan [Internet]. 2015 [citado 6 Mar 2020];19(2):150-7. Disponible en: http://scielo.sld.cu/scielo.php?script=sci_arttext&pid=S1029-30192015000200002&lng=es
23.González Machado EC, Reyes Piñuelas EP, López Ortega M. Construcción de una prueba para evaluar aprendizajes en educación superior. Congreso nacional de investigación educativa [Internet]. San Luis Potosí: Congreso Nacional de Investigación Educativa; 2017 [citado 6 Mar 2020]. Disponible en: http://www.comie.org.mx/congreso/memoriaelectronica/v14/doc/2300.pdf
24.Carvajal Álzate DE, Méndez Sánchez H, Torres Angulo MB. Análisis de la confiabilidad y de algunos parámetros psicométricos de un test realizado en el colegio Vista Bella de la ciudad de Bogotá [Tesis]. Bogotá: Fundación Universitaria Los Libertadores. Departamento de Ciencias Básicas, Especialización en Estadística Aplicada; 2016 [citado 6 Mar 2020]. Disponible en: https://repository.libertadores.edu.co/handle/11371/620
25. Leenen I. Virtudes y limitaciones de la teoría de respuesta al ítem para la evaluación educativa en las ciencias médicas. Inv Ed Med [Internet]. 2014 [citado 6 Mar 2020];3(9):40-55. Disponible en: https://www.sciencedirect.com/science/article/pii/S2007505714727243

Conflicto de intereses.
El autor declara que no tiene conflictos de intereses para la publicación del artículo.

Citar como: Martínez Pérez R. Criterios de calidad de los instrumentos evaluativos escritos. Medimay [Internet]. 2020[citado: fecha de citado]; Abr-Jun;27(2):240-51. Disponible en: http://www.medimay.sld.cu/index.php/rcmh/article/view/1662

Contribución de autoría
El autor se responsabiliza con el texto que se publica.