Revista Mexicana de Anestesiología

ARTÍCULO DE REVISIÓN
Vol. 30. No. 3 Julio-Septiembre 2007
pp 158-164

Instrumentos de evaluación

Dr. Alberto Javier García-Garro,* Cit. Gregorio Ramos-Ortega,**
Dr. Manuel Antonio Díaz de León-Ponce,*** T.R. Alfredo Olvera-Chávez****

* Coordinador Médico de Programas de Estudios Técnicos. Coordinación de Educación en Salud. Maestría en Ciencias de la Educación.
** Profesor Titular del Curso Formativo de Citotecnólogos. Hospital General de Zona No. 57, La Quebrada.
*** Académico Titular de Academia Nacional de Medicina y Académico Emérito de la Academia Mexicana de Cirugía.
**** Profesor Titular del Curso Profesional Técnico en Terapia Respiratoria. Hospital de Especialidades, Centro Médico Nacional Siglo XXI.

Instituto Mexicano del Seguro Social Coordinación de Educación en Salud.
Coordinación de Programas de Estudios Técnicos.

Solicitud de sobretiros:
Dr. Alberto Javier García-Garro
Teléfono: 56276900 ext 21194.
E-mail: alberto.garcia@imss.gob.mx.

Recibido para publicación: 07-03-07
Aceptado para publicación: 08-05-07

RESUMEN

Para el desarrollo de buenas pruebas son importantes dos factores: la confiabilidad y la validez. Si hablamos en forma general, las pruebas más extensas son más confiables que otras. Esto es, la prueba de confiabilidad continúa utilizando puntuaciones en tanto existan pruebas no confiables. Si ocurre así, la prueba tiene confiabilidad. Todas las pruebas son indicadores imperfectos de cualidades o habilidades que tratan de medir. En todas las situaciones de prueba existen errores; de modo que, una prueba confiable también puede definirse como una prueba con un error estándar de medición reducido. Validez concurrente y validez predictiva: los reactivos no son fáciles de estructurar. Los reactivos interdependientes disminuyen la exactitud de una prueba y por tanto su validez. La mayoría de los aplicadores inteligentes de pruebas saben que las respuestas categóricas por lo general son erróneas. Si sólo una respuesta puede ser correcta y dos son iguales, entonces estas dos deben ser incorrectas; son reactivos relativamente fáciles de construir, pues son simples enunciados declarativos. Validez concurrente: Si la nueva prueba se valida por comparación con un criterio actual existente, tenemos validez concurrente. Validez predictiva: Algunas pruebas se diseñaron para predecir los resultados.
Validez total = validez de contenido + validez de criterio + validez de constructo
Así, la validez de un instrumento de medición se evalúa sobre la base de tres tipos de evidencia. La validez y la confiabilidad son medidas de grado, se entiende que una prueba puede tener mayor o menor grado de confiabilidad o validez.

Palabras clave: Instrumentos, evaluación.

SUMMARY

Two factors are important in the development of good tests: reliability and validity. Generally speaking, more extensive tests are more reliable than others, i.e., reliable tests will go on using scores as long as non reliable tests exist. This way, the test is reliable. All tests are less-than-perfect indicators of the qualities or abilities being measured. Errors may occur in all test settings. Thus, a reliable test may also be defined as a test with a small measurement standard error. Furthermore, test items are not easy to elaborate. Interdependent questions decrease the accuracy of a test, and hence its validity. Most smart testers are aware that categorical answers are usually wrong. If only one answer is right and two others are equal, then these two should be wrong; such items are relatively easy to elaborate, since they are mere declarative statements. Concurrent and predictive validity: If a new test is validated by comparison with an existing parameter, it has con-

current validity. Tests designed to predict their own results have predictive validity. Total validity = content validity + criterion validity + construct validity. *Thus, the validity of a instrument for measurement is assessed based on three types of evidence. Validity and reliability both measure a degree, not an absolute value, i.e., a test may have a higher or lower degree of reliability or validity.*

Key words: *Evaluation instruments, reliability, validity, test items.*

CARACTERÍSTICAS DEL INSTRUMENTO

Los instrumentos, como herramientas utilizadas para recolectar información nos ayudan a la medición, la cual constituye una actividad presente en la práctica docente, ésta a su vez nos aproxima al monitoreo y evaluación del proceso educativo. ¿Qué instrumentos y cuáles son las características para usarse en la recolección de datos con fines más allá de la medición, dando cuenta del monitoreo y la evaluación?, ¿Son las encuestas buenos instrumentos para tal fin?, ¿Serán mejor los debates de grupos focales?

ENCUESTAS

Las encuestas son una de las herramientas de evaluación usadas con más frecuencia. Generalmente son una manera efectiva en términos de costos para recoger información comparable de un grupo grande de personas[1].

PREGUNTAS DE INTERPRETACIÓN ABIERTA Y PREGUNTAS CERRADAS

Una de las preguntas que surgen más a menudo en las investigaciones de encuestas es ¿qué tipo de pregunta es mejor: las de interpretación abierta o las cerradas? Veamos: las preguntas de interpretación abierta son aquéllas para las cuales no existe una respuesta predeterminada, mientras que las preguntas cerradas son aquéllas para las cuales sólo son posibles unas pocas respuestas que se incluyen en el cuestionario y el entrevistado selecciona; de allí se deberá tener en cuenta que en su construcción, las preguntas sean mutuamente excluyentes, procurando incluso darle oportunidad a los entrevistados la opción de decir «no sé» como respuesta a una pregunta planteada[2-4].

OTROS PUNTOS PARA DESARROLLAR ENCUESTAS EFECTIVAS

Mantenerla bien centrada, corta, sencilla, clara, efectuar una prueba preliminar y abordar la confidencialidad.

DEBATES DE GRUPOS FOCALES

Los debates de grupos focales son una técnica de investigación cualitativa privilegiada en la investigación de ciencias sociohumanas. En grupos pequeños, puede ser dominado por una o dos personas, pero si el grupo es de más de 10 ó 12 personas, su desarrollo puede ser difícil, por lo que es recomendable efectuarse al menos dos debates de grupos focales entre cada grupo participante[5].

UBICACIÓN

La ubicación para los debates de grupos focales debe procurar ambientes en la que los participantes se sientan cómodos y procedan de manera natural en razón del interés de la investigación. Elegir a los participantes puede ser difícil. La idea es tener un grupo homogéneo en términos de características relevantes a las preguntas de la investigación, pero a su vez heterogéneo para que fluya la diversidad con base en sus opiniones.

TONO DEL GRUPO

Mientras se efectúa un debate de grupo focal, es necesario recordar que lo habitual es que las personas revelarán más información en grupos donde se sientan apoyadas y no juzgadas.

EL MODERADOR

El moderador en su papel de dirigente del debate, deberá ser muy sensible para explorar, captar y explotar al máximo los temas a debatir.

EXAMEN

Por otra parte y particularizando la evaluación, en este caso del aprendizaje, entraremos en una dimensión esencial del proceso de enseñanza, aspecto de los más polémicos y de mayor interés en la práctica educativa por su pa-

pel fundamental dirigido a mejorar la enseñanza sin perder de vista que esta dimensión nos centra en el poder y el control que el docente ejerce sobre el alumno al valorar los conocimientos de habilidades y destrezas a los que han llegado como resultado del proceso docente, así como el proceso mismo de la construcción. El instrumento por excelencia empleado es sin duda el examen, el cual es resultado de diversas concepciones sobre el aprendizaje y no el motor que lo transforma[6].

Elaborar exámenes entraña una sólida formación docente-investigador expresados en su propuesta técnica, que dé cuenta de cómo elaborar exámenes, su manejo estadístico de datos, su construcción de reactivos objetivos, entre otros. Aun así hemos de considerar que dentro de las técnicas formales de la evaluación, estos son los instrumentos de uso más generalizado, por lo que es preciso considerar al menos que su intención es lograr una evaluación objetiva, libre de interpretaciones idiosincrásicas al establecer juicios sobre los aprendizajes de los alumnos.

Las pruebas objetivas se caracterizan por estar construidas a base de reactivos cuya respuesta no deja lugar a dudas respecto a su corrección o incorrección, trabajando el estudiante sobre una situación estructurada a la que no aporta más que respuestas concretas; este tipo de prueba es posible emplearla con fines selectivos, diagnósticos, formativos, sumativos o de certificación, lo cual ya impone ciertas modalidades según el propósito para el que va a ser empleada.

El nivel de estructuración de los reactivos influye de manera importante en el tipo de procesos cognoscitivos y de aprendizaje significativo que logran los alumnos.

Un reactivo útil y valioso, combina su correcta construcción, la relación con los logros que se busca medir y su integración equilibrada al resto de reactivos incluidos en una prueba. Un buen reactivo se caracteriza por su validez de contenido; es decir su contenido corresponde al objetivo de aprendizaje para el cual fue elaborado[7].

En general este tipo de pruebas objetivas tienen alguna característica que no necesariamente se traducen en fortalezas del mismo, a pesar de ello señalamos a continuación:

- Son más fáciles de contestar que los de tipo ensayo. Los de opción múltiple permiten la medición en grandes grupos, ya que son relativamente fáciles de calificar al cotejar con la respuesta tipo (plantilla de respuestas).
- Gran parte de los reactivos pueden responderse por medio de aprendizajes memorísticos o aprendizajes poco significativos.
- Adecuados para medir los resultados en los niveles de aprendizaje de conocimiento, comprensión y aplicación; inadecuados para organizar y expresar ideas.
- No son válidos para explorar destrezas y actitudes.
- Proporcionan poca retroalimentación cualitativa sobre la situación de enseñanza.
- Generan ansiedad en los alumnos («ansiedad de prueba»).
- La capacidad de lectura y adivinar son factores que distorsionan las calificaciones.
- Es preciso determinar la complejidad de los objetivos a evaluar (a mayor cantidad y complejidad de objetivos más reactivos).
- Es menester identificar el tiempo disponible para su respuesta, por lo general se dispone de un tiempo predeterminado para la aplicación de la prueba.

A continuación se presentan los conceptos, usos y aplicación, así como las recomendaciones específicas para la elaboración de diferentes tipos de reactivos a incluir en una prueba objetiva.

REACTIVOS DE OPCIÓN MÚLTIPLE

Los reactivos de opción múltiple están constituidos en su forma clásica, por un enunciado incompleto o una pregunta (encabezado, tallo, tronco o base) en el que se plantea el problema a resolver y varias posibles respuestas (opciones o alternativas) una de las cuales es la correcta y las otras incorrectas (distractores).

Los aprendizajes que se pueden medir con los reactivos de opción múltiple se relacionan con contenidos declarativos (datos, hechos, conceptos y principios) en las categorías de conocimiento, comprensión, aplicación y análisis de la taxonomía de Bloom.

Los reactivos no son fáciles de estructurar. La habilidad y experiencia en la redacción son importantes. Cuando se emplea un número reducido de opciones; el reactivo disminuye su valor de medición.

- Cada reactivo debe ser independiente de los otros. Los reactivos interdependientes disminuyen la exactitud de una prueba y por tanto su validez.
- Si todas las opciones son homogéneas respecto al tópico (tema) abordado en el tallo, serán más razonables.
- La longitud de las opciones no debe dar la clave de la respuesta.
- Al construir la prueba, se debe evitar que un reactivo contenga la respuesta de otros.
- El tallo debe ser claro, simple y presentar sólo un problema.
- La mayoría de los aplicadores inteligentes de pruebas saben que las respuestas categóricas por lo general son erróneas.
- Si sólo una respuesta puede ser correcta y dos son iguales, entonces estas dos deben ser incorrectas.

- Se deben evitar asociaciones verbales entre la base y la respuesta correcta.

REACTIVOS DE RESPUESTA ALTERNA (ALTERNATIVAS CONSTANTES)

Este tipo de reactivos se caracterizan por limitar la respuesta a una de dos opciones o alternativas (verdadero — falso; sí — no; nunca — siempre; correcto — incorrecto; o respuestas similares) para calificar una aseveración o enunciado.

Son reactivos relativamente fáciles de construir, pues son simples enunciados declarativos.

El número de respuestas calificables por cada mil palabras de examen o por cada minuto que éste dure es considerablemente superior al número de respuestas calificables en un examen de selección múltiple y probablemente muy superior a muchos otros tipos de examen.

Los reactivos buscan comprobar el conocimiento, por lo que se debe:

- Procurar construir reactivos que requieran sólo una respuesta correcta.
- Evitar la utilización de palabras que descubran la respuesta.
- Omitir más de tres palabras en un solo enunciado, omita sólo palabras o datos claves.
- Si la respuesta es numérica, indicar las unidades en que debe ser expresada.
- Proporcionar instrucciones claras y específicas sobre la forma de responder[7].

REACTIVOS DE RESPUESTA BREVE O SIMPLE

Son enunciados interrogativos que deben ser respondidos a través de una palabra, frase o enunciado corto.

Este tipo de reactivos se recomienda para medir conocimientos de asociaciones. No son adecuados para medir resultados complejos como comprensión, aplicación, análisis y organización.

RECOMENDAMOS PARA SU ELABORACIÓN

- Asegurarse que la pregunta plantea un problema concreto.
- Los grupos de respuestas y sus relaciones deben ser del mismo tipo y naturaleza.
- Elaborar columnas claras y ordenadas.

VALIDACIÓN DE LOS INSTRUMENTOS

Uno de los problemas más comunes con el uso de las pruebas es la interpretación errónea de las calificaciones. Ninguna prueba proporciona una imagen perfecta de las habilidades de una persona; una prueba sólo es una muestra pequeña de la conducta. En el desarrollo de buenas pruebas son importantes dos factores: la confiabilidad y la validez, atributos del instrumento a considerar antes de interpretar las calificaciones obtenidas tras la aplicación de los instrumentos[8,9].

A fin de perfeccionar la evaluación de test psicológicos, Cronbach (1951) introdujo en la evaluación dos elementos a considerar en la confiabilidad de los instrumentos: 1) equivalencia confiabilidad interna o consistencia interna) y 2) estabilidad (confiabilidad externa).

Ambos elementos confieren cierto grado de precisión de la prueba.

MÉTODOS Y CÁLCULO DE LA CONFIABILIDAD

Existen diversos métodos y procedimientos para medir la confiabilidad de los instrumentos. En general los coeficientes de confiabilidad aceptables para pruebas de rendimiento escolar se encuentran entre .60 y .80.

CONFIABILIDAD INTERNA

1) Método de la división por pares o método de mitades partidas (split-halves).

En este método se requiere sólo una aplicación de la medición. Específicamente la prueba se aplica a un grupo de sujetos y más tarde el conjunto total de reactivos es dividido en dos mitades. Si la prueba es confiable, entonces las puntuaciones de las personas en cada mitad deberían ser similares y el grado de similitud se evalúa utilizando la correlación. La forma más efectiva para mejorar la confiabilidad es agregar más conceptos a una prueba. Si hablamos en forma general, las pruebas más extensas son más confiables que otras.

Consiste en tomar a cada uno de los reactivos como unidad y compararlo con el resto de los reactivos que integran la prueba, esto nos va a proporcionar información sobre la consistencia interna de la prueba[10,11].

ANÁLISIS DE REACTIVO

Los reactivos producirán mayor confiabilidad en un cuestionario si discriminan bien entre los individuos. Hay dos métodos comunes para verificar el poder de discriminación de los reactivos. Para cada reactivo en la prueba o cuestionario, se calcula la correlación entre la puntuación de cada persona en el reactivo y su puntuación en la prueba como un todo.

Después se totalizan las puntuaciones de estos dos grupos de personas para cada reactivo en la prueba. Esto es, la prueba de confiabilidad continúa utilizando puntuaciones en tanto existan pruebas no confiables.

CONFIABILIDAD EXTERNA

Confiabilidad de test-retest. Consiste en aplicar la prueba en un momento y al cabo de dos semanas, volverlo a aplicar. Si los resultados son más o menos similares, la prueba tiene confiabilidad; sin embargo, hay varias razones por las cuales las personas obtienen la misma puntuación, la segunda vez en la misma prueba. Si transcurre poco tiempo entre las pruebas, los alumnos pueden distinguir las respuestas correctas por recuerdo, lo cual generará un defecto de confiabilidad[12].

Las versiones son similares en contenido, instrucciones, duración y otras características. Los patrones de respuesta deben variar poco entre las aplicaciones.

Calificación real: Todas las pruebas son indicadores imperfectos de las cualidades o habilidades que tratan de medir. En todas las situaciones de prueba existen errores, en ocasiones, los errores son en su contra, pero si pudiera presentar la prueba una y otra vez sin cansarse o sin aprenderse de memoria las respuestas, su buena y mala suerte terminaría y el promedio de las calificaciones de la prueba se acercarían a una calificación real; sin embargo, en realidad, los estudiantes presentan una prueba sólo una vez. Esto significa que la calificación que recibe cada alumno está constituida por la calificación real hipotética más cierta cantidad de error. En las pruebas estandarizadas, las personas que desarrollan pruebas toman esto en consideración y realizan cálculos de cuánto variarían las calificaciones de los estudiantes si presentaran la prueba en repetidas ocasiones. Este cálculo se llama error estándar de medición. De modo que, una prueba confiable también puede definirse como una prueba con un error estándar de medición reducido.

VALIDEZ

Si una prueba es lo suficientemente confiable, la pregunta siguiente es ¿qué tan válida es?, o en forma más exacta, si los juicios y decisiones que se basan en la prueba son válidos.

Para tener validez, las decisiones e inferencias que se basan en la prueba deben tener respaldo por evidencia. Se habla de validez de una prueba al grado en que mide el atributo o característica para la cual fue elaborada; así una prueba es válida si realmente mide lo que se supone debe medir[13,14,18].

Una prueba sobre conocimientos de historia debe medir esto y no conocimientos de literatura histórica.

EVIDENCIAS RELACIONADAS CON LA VALIDEZ DE CONTENIDO

Una prueba tiene validez de contenido si está hecha con una muestra representativa de los objetivos y contenidos abordados.

EVIDENCIAS RELACIONADAS CON LA VALIDEZ DE CRITERIO

La validez de criterio establece la validez de un instrumento de medición comparándola con un criterio externo. Hay dos tipos de validez de criterio que difieren sólo en términos de la puesta a punto de la prueba de criterio. Validez concurrente y validez predictiva.

VALIDEZ CONCURRENTE

Si la nueva prueba se valida por comparación con un criterio actual existente, tenemos *validez concurrente*. La validez concurrente, hace referencia; por lo tanto a la relación que existe entre las calificaciones obtenidas con la prueba y un criterio universalmente aceptado como válido para medir lo que la prueba pretende evaluar.

VALIDEZ PREDICTIVA

Algunas pruebas se diseñaron para predecir los resultados. Las pruebas SAT, por ejemplo, tienen el propósito de pronosticar el desempeño en la universidad. En otras palabras, las calificaciones de la prueba son indicadores bastante exactos del criterio (qué tan bueno será el desempeño que tendrá el estudiante en la universidad).

La validez de constructo suele determinarse mediante un procedimiento denominado «análisis de factores». La evidencia de validez de constructo se recopila durante varios años.

La evidencia de validez de constructo, también puede demostrarse cuando los resultados de la prueba se correlacionan con los de otras medidas válidas y bien establecidas del mismo constructo[18,19].

Validez total = *validez de contenido + validez de criterio + validez de constructo;*

así, la validez de un instrumento de medición se evalúa sobre la base de tres tipos de evidencia. Entre mayor evidencia de validez de contenido, validez de criterio y validez de constructo tenga un instrumento de medición éste se acerca más a representar lo que pretende medir; y aún surgen nuevas preguntas acerca de la validez. ¿Cuáles son las consecuencias de utilizar un planteamiento de evaluación particular para la enseñanza y el aprendizaje? Sam Messic (1975) formuló dos preguntas importantes que deben considerarse al tomar cualquier decisión sobre el uso de una prueba: ¿La prueba es una buena medida de la característica que se supone debe evaluar? ¿Se debe utilizar la prueba para los fines propuestos? La primera pregunta se asocia con la validez del constructo; la segunda se refiere a la ética y los valores (Moss, 1992).

Algunos factores pueden interferir en la validez de las pruebas que se aplican en las situaciones de un salón de clases, las pruebas de rendimiento estandarizadas deben seleccionarse de modo que los incisos en la prueba midan los conocimientos adquiridos en las clases; así mismo, los estudiantes deben contar con las habilidades necesarias para presentar la prueba. Si los estudiantes obtienen calificaciones bajas en una prueba de ciencias no por su falta de conocimientos sobre las ciencias, sino porque tienen dificultades para leer las preguntas, no comprenden las instrucciones o no tienen tiempo suficiente para terminar, entonces la prueba no es una medida válida del rendimiento en ciencias de esos estudiantes.

Las pruebas, cuestionarios, guías de observación, listas de cotejo y escalas evaluativas son sólo algunos ejemplos de instrumentos de medición, los cuales requieren de un procedimiento específico para su elaboración.

La validez y la confiabilidad son medidas de grado, por lo cual se entiende que una prueba puede tener mayor o menor grado de confiabilidad o validez.

Una prueba debe ser confiable a fin de ser válida. Sin embargo, la confiabilidad no garantizará validez; una prueba puede ser confiable pero no válida; mas una prueba que es válida necesariamente tiene que ser confiable. Los siguientes lineamientos podrían ayudarle a incrementar la confiabilidad y validez de las pruebas estandarizadas.

Asegúrese de que la prueba en realidad cubra el contenido de la unidad de estudio:

1. Compare las preguntas de la prueba con los objetivos del curso.
2. Una matriz de contenido de la conducta podría ser útil en este caso.
3. Utilice las pruebas de rendimiento y normas locales siempre que sea posible.
4. ¿Sus alumnos experimentan alguna dificultad con la prueba como no tener tiempo suficiente, el nivel de lectura y demás? De ser así, analice estos problemas con personal idóneo de la escuela.

Asegúrese de que sus alumnos sepan cómo utilizar todos los materiales de la prueba

Ejemplos:

1. Siga las instrucciones para administrar la prueba con exactitud.
2. Asegúrese de que los estudiantes estén tan cómodos como sea posible durante la prueba.
3. Recuerde que ninguna calificación en las pruebas es perfecta.

EVALUACIÓN POR NORMAS O POR CRITERIOS

Los métodos de evaluación deben ser los más adecuados para evaluar integralmente. De lo antes dicho, debe ser evidente el contraste entre evaluación basada en competencias y la evaluación tradicional por norma. Mientras que la evaluación basada en competencias evalúa el desempeño de un individuo a partir de criterios preestablecidos, la evaluación por norma se encarga de comparar el desempeño de un individuo con el desempeño del grupo; no obstante, los métodos que se utilizan para realizar una evaluación basada en normas son similares a los métodos tradicionales.

Los enfoques integrados buscan combinar conocimiento, comprensión solución de problemas, habilidades técnicas, actitudes y valores en la evaluación:

- Estar orientada al problema.
- Ser interdisciplinaria.
- Considerar la práctica.
- Cubrir grupos de competencias.
- Demandar habilidades analíticas.
- Combinar la teoría y la práctica.

En el contexto de medicina, los métodos que incluyen más niveles de evaluación integrada que los exámenes formales son: Problemas con el manejo del paciente por ejemplo: simulaciones escritas sobre problemas del paciente ***(caso clínico)***.

REFERENCIAS

1. Almeida N. Desarrollo de instrumentos en la investigación epidemiológica. En: Epidemiología sin números. Washington, D.C.: Organización Panamericana de la Salud; 1992:43-57.
2. Bedolla G. Cómo estimar e interpretar y mejorar la confiabilidad de una evaluación. Monterrey, México: UDME; 1983:9-10.
3. Davey DD, McGoogan E, Somrak MT, et al. Competency assessment and proficiency testing. Act Cytol 2000;44:939-43.
4. Donají RC, Espinosa AP. Competencia clínica en hipertensión arterial sistémica de alumnos de pregrado de dos escuelas de medicina. Rev Invest Clin 2000;52:132-9.
5. García CF. La medición y evaluación educativa. En: Lifshitz A, Editor. Educación Médica. Enseñanza y aprendizaje de la clínica. México: Auroch; 1997:168-9.
6. Díaz BA. El problema de la teoría de la evaluación y cuantificación del aprendizaje. En: Díaz Barriga, A. (compilador) El examen. Textos para su historia y debate. México, CESU/UNAM/ Plaza y Valdés Editores 2000:304-314.
7. Rojas MI. La educación basada en normas de competencia (EBNC) como nuevo modelo de formación profesional en México. En: Valle Flores, MA (Coordinadora). Formación en

competencias y certificación profesional México CESU/UNAM (Colección Pensamiento Universitario. Tercera Época (No. 9) 2000:203-205.
8. Rueda BM. Notas para una agenda de discusión sobre evaluación de la docencia en universidades. En: ¿Hacia una nueva cultura de la evaluación de los académicos? México CESU/UNAM Colección Pensamiento Universitario. Tercera Época No. 88 1999:203-205.
9. Landis JR, Koch GG. The measurement of observer agreement for categorical data. Biometrics 1977;33:159-174.
10. Rube FI. Experience in managing a large-scale rescreening of Papanicolaou smear and the pros and cons of measuring proficiency with visual and written examinations. Act Cytol 1989;4:479-83.
11. Salazar L. Evaluación de desempeño de estudiantes de salud, usando el modelo de calidad de cuidado percibido por la comunidad. Colombia Med 1999;30:62-8.
12. Thompson, WD. Canadian Experience in Cytology. Proficiency testing. Act Cytol 1989;33:484-6.
13. Vargas HJG. Las reglas cambiantes de la competitividad global en el nuevo milenio. Las reglas cambiantes de la competitividad global en las universidades. Las competencias en el nuevo paradigma de la globalización, En Crecemos. Revista Hispanoamericana de Desarrollo Humano y Pensamiento. Año 7, 2004;1:17-20.
14. Gonczi A. Enfoques de la educación basada en competencias: la experiencia de Australia (segunda parte). La Academia. Anónimo: 56-60.
15. Gonczi A, Athanasou 1. Evaluación. En: Argüelles A, (Compilador) Competencia laboral y educación basada en normas de competencia. México: Limusa; 1996:284-5.
- Guía Técnica para elaborar programas educativos por competencias para profesionales del área de la salud. IMSS. Área de Estudios para personal técnico. Coordinación de Educación Médica del IMSS, 1999.
16. Sabido-Siglher M, Viniegra L. Competencia y desempeño clínicos en diabetes. Rev Inv Clin 1998;50:211-6.
17. Nagy KG, Collins ND. False positive and false negative proficiency test. Results in cytology. Act Cytol 1991;35:3-7.
18. Furlan A. La evaluación de los académicos. En: ¿Hacia una nueva cultura de la evaluación de los académicos?, México, CESU/UNAM Colección Pensamiento Universitario. Tercera Época No. 88 1999:56-66.
19. Vooijs GP, Davey DD, Somrak MT, et al. Computerized training and proficiency testing. Act Cytol 1998;42:141-7.