Revista de Enseñanza Universitaria 1998, Nº 13; 105-111

ANALISIS EVALUATIVO DE CALIDAD DE LA PRUEBA OBJETIVA TIPO TEST (PREGUNTAS DE ELECCIÓN MÚLTIPLE)

Ignacio Gómez de Terreros Sánchez
Hospital Universitario "Virgen del Rocío"
Facultad de Medicina
Universidad de Sevilla

Resumen

El examen tipo test o preguntas de elección múltiple, es un método de evaluación impuesto en las Facultades de Medicina. Su dificultad de confección justifica la conveniencia de disponer de una herramienta que permita el análisis individualizado de sus numerosas preguntas en la consecución de una prueba objetiva de calidad que potencie su carácter formativo. La actual tecnología informática facilita la ejecución de dicho objetivo a través del calculo de los siguientes indicadores: Identificación de los grupos fuertes y débiles. Indice de dificultad. Indice de discriminación. Respuestas de los distintos ítems por pregunta. Datos recogidos en hoja evaluativa por pregunta, que prevé futuros análisis en estrategia de feedback. Beneficio en la calidad de las cuestiones planteadas y en el proceso de tutorías de revisión de exámenes han sido ostensible, permitiendo profundizar en el proceso formativo de cada alumno.

Abstrac

The multiple choice test is an evaluation method used in the Faculties of Medicine. It is difficult to elaborate but this makes it a convenient tool for individualizing the analysis of the numerous questions, thus acheiving quality and educational character. The current computer technology enables us to obtain our aim through the following indicators: Identification of two groups of students (the most outstanding and the mediocre). Index of difficulty. Discrimination Index. Answers to the different items per question. Data collected from the evaluation sheet, which wil enable an analysis of feedback.

It benefits the quality of the questions and tutorial revisions of the examinations, thus helping the training process of each student.

INTRODUCCIÓN

El examen tipo test o también denominado preguntas de elección múltiple (PEM) es un método de evaluación cuyo empleo y difusión se ha impuesto en las Facultades de Medicina, máxime al verse abocados los alumnos a la prueba MIR, sin olvidar que su

actual elevado número hace difícil mantener los métodos tradicionales. Además, el examen test permite la exploración de gran cantidad de áreas en los conocimientos del alumno dentro de una indudable objetividad.

En el presente trabajo no nos planteamos el análisis de los distintos tipos y subtipos de pruebas objetivas, ni los criterios a tener en cuenta para la redacción de ítems (preguntas con una sola respuesta verdadera), que eviten los errores, no infrecuentes, que se pueden cometer durante su confección. Diversas publicaciones existen al respecto y, dentro de nuestra propia Facultad de Medicina, los profesores Bedoya y Uralde (1976) y González Hachero (1988) han escrito sobre el tema.

Partiendo de la dificultad en la confección del PEM, se pone en evidencia la conveniencia de disponer de una herramienta que permita el análisis individualizado de la calidad de las múltiples preguntas que constituyen dicho examen tipo test. Tengamos en cuenta que una vez formuladas engrosarán o se reintegrarán al banco de preguntas del profesor o profesores de la asignatura, sirviendo para la construcción de futuros exámenes, por lo que si no están correctamente elaboradas será factible la reiteración de posibles errores.

La generalizada implantación de este tipo de exámenes en las Facultades de Medicina no ha hecho desaparecer los argumentos en su contra ante una utilización inadecuada o deficiente en su construcción, sino todo lo contrario. En realidad han tenido el efecto de generar consecuentemente múltiples discusiones en cuanto a sus ventajas e inconvenientes. En esa discusión las alternativas barajadas han ido, en un amplio abanico de opciones, desde los que se muestran favorables a su pura y simple su
ción hasta los que prefieren compleme
los con otro tipo de exámenes. Los qu
sitúan en esta última opción de la con
mentariedad pueden optar a su vez por
rrir a exámenes más complejos, como
den ser los de solución de problema
otros de más fácil elaboración, que re
ren mas tiempo de corrección e imp
pérdida en la objetividad, como puede
las pruebas escritas con respuestas abi
y cortas o el examen programado. Este
mo método es el desarrollado por la N
nal Board of Medical Examiners en Est
Unidos que, manteniendo las caracterís
del tipo test, permite medir por simula
clínica sobre el papel la competencia
conocimientos del alumno.

Si bien las pruebas objetivas no son
fectas, se reconoce que permiten alcanz
evaluación de la mayor parte de la tax
mía clásica en pedagogía. Sirven para
mar el conocimiento, la compresión, la
cación, el análisis, la síntesis y otros niv
demostrando por tanto su utilidad para
mar el grado de aprendizaje en el ám
cognoscitivo. Exigen, por supuesto,
muy cuidadosa construcción que cumpl
objetivos que se pretende con cada pre
ta, pues si no se realiza con coherencia,
de todas sus ventajas.

Partimos, pues, de la base de qu
pesar del esfuerzo del profesor que forr
las preguntas, es inevitable que se pro
can transgresiones pedagógicas en los í
utilizados. Ello implica la exigencia de
estrategia localizadora y correctora en b
del objetivo de conseguir una prueba ob
va de calidad, que a su vez redund
potenciar el carácter formativo de la mi

La actual tecnología informática
facilita la ejecución de dicho objetiv

permitir un análisis individualizado de cada pregunta sobre su efectividad y construcción. Es ésta la empresa en la que nos hemos implicado y cuyo desarrollo procedemos a exponer.

EXAMEN DE LA PREGUNTA DE ELECIÓN MULTIPLE (PEM) MOTIVO DE ANÁLISIS

El examen de tipo test motivo de análisis, corresponde al implantado en el Área de Pediatría del Departamento de Farmacología, Pediatría y Radiología de la Facultad de Medicina de la Universidad de Sevilla, en cuya utilización fuimos entrenados por el Profesor González Hachero, al iniciarse las actividades Docentes en el Hospital Infantil Universitario "Virgen del Rocío".

Constituye una prueba objetiva de tipo test con que se evalúa el contenido teórico de la asignatura de Pediatría. Su construcción se elabora con cinco posibles opciones al planteamiento de la pregunta, de las cuales una sola es la verdadera, utilizándose como factor corrector de la respuesta acertada al azar la fórmula:

$$R = A - \frac{E}{N - 1}$$

en la que R= resultado final; A= aciertos; E= errores y N= número de respuestas posibles (Velasco,1972).

Para eliminar materia se exige la contestación del 60% de las 80 preguntas que constituyen el examen, trasladándose a la puntuación clásica de 0 a 10 con calificación de suspenso, aprobado, notable, sobresaliente exigida por la normativa de la Universidad a través del oportuno cuadro de correspondencias.

En la confección del examen García Barbero (1987) resalta cinco puntos en los cuales hay que poner especial cuidado en evitar:

Que haya errores en la interpretación.
Proporcionar claves a los alumnos.
La utilización de términos ambiguos.
La complejidad de la formulación.
Los juicios de valor.

Los puntos que señala como más frecuentes en los que se cometen errores son:
Preguntas irrelevantes.
Utilización de más palabras de las necesarias.
Preguntas que parecen jeroglíficos.
Uso de adjetivos y adverbios superfluos.
Uso de adverbios de calidad y cantidad poco exactos.
Inclusión de unas opciones en otras.
Uso de nombres propios de manera indiscriminada.
Uso de cifras inexactas.
Colocación de las cifras en desorden.
Cifras que se engloban las unas en las otras.
Falta de unidades de medida.
Mezcla de negaciones entre los enunciado y las opciones.

Son errores que sin duda interfieren en la calidad evaluativa de la prueba y que nos servirá en el proceso de análisis de cada pregunta. Posteriores re-evaluaciones proporcionarán un feedback a los profesores, reforzando su calidad e introduciendo un dinamismo al "banco de preguntas" del que por lo general carece.

METODOLOGIA

La metodología a utilizar se basa en el análisis con soporte informático de la prueba tipo PEM, que ofrezca una información

individualizada (evaluación crítica por pregunta), a través de los siguientes indicadores:

- Identificación de los grupos fuertes y débiles.
- Indice de dificultad.
- Indice de discriminación.

El Soporte informático lo constituy

Lector óptico modelo ScanMark 2 Hoja de respuestas especial (Fig 1).

Ordenador: Pentium 120 con 16 memoria Ram. CDRUM incorporado.

Impresora Laser Jet IIIP. Hewlet kard

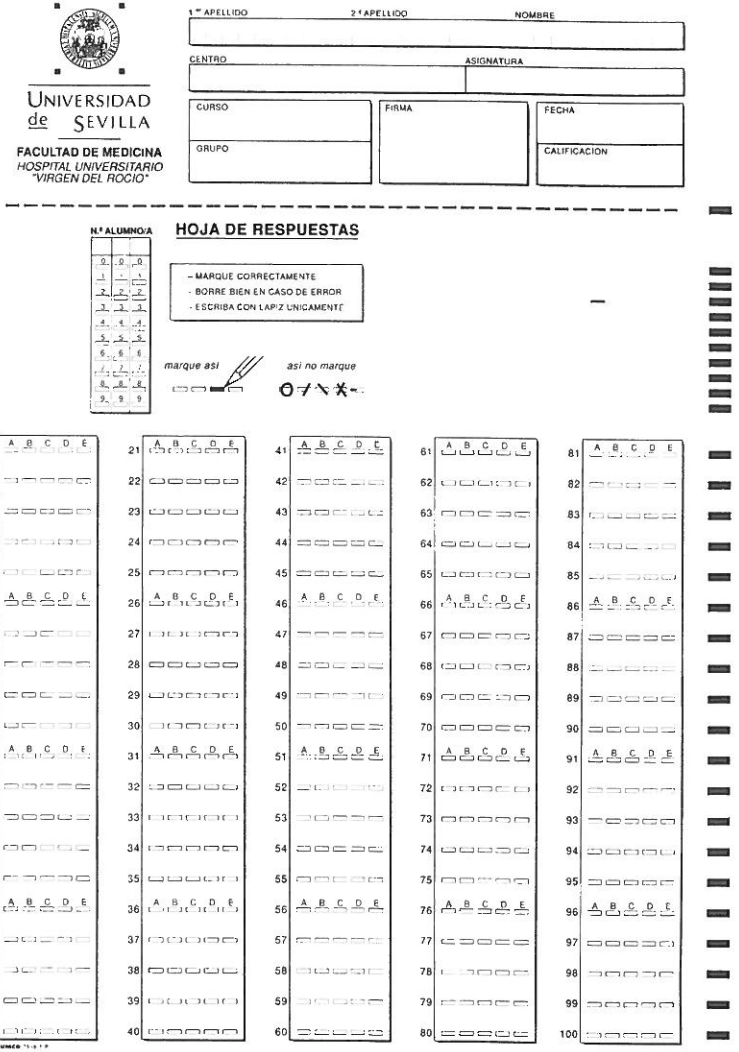

FIG. 1

IDENTIFICACIÓN DE LOS GRUPOS FUERTE Y DEBIL

Se aconseja construir los grupos fuerte y débil tomando solamente los primeros 27 por 100 (grupo fuerte) y los últimos 27 por 100 (grupo débil) del conjunto de los estudiantes clasificados por orden de puntuación. Se considera que el 27 por 100 corresponde al mejor compromiso entre dos fines deseables pero contradictorios:

Hacer los dos grupos tan grandes como sea posible.

Hacer los dos grupos tan diferentes como sea posible.

Aunque teóricamente la elección del 27 por 100 sea la mejor, no es verdaderamente preferible al 25 por 100 ó al 33 por 100. Por ello se prefiere trabajar con 1/4 ó 1/3, más que con esta cifra un poco singular del 27 por 100 (Guilbert 1994).

ÍNDICE DE DIFICULTAD:

Es un índice que nos permite determinar en qué medida una pregunta de examen es fácil o difícil. Cuanto más elevado es este índice, más fácil es la pregunta al ser mayor el porcentaje de los estudiantes de un grupo determinado que la han respondido correctamente; sería más lógico, por tanto, que se le denominara índice de facilidad o de éxito. Varía de 0 a 100, siendo 100 muy fácil y 0 muy difícil. En principio, una pregunta que tenga un índice de dificultad comprendido entre 30 y 70 por 100 es aceptable.

Para su cálculo se utiliza la fórmula siguiente:

$$\text{Indice de dificultad} = \frac{F + D}{N} \times 100$$

donde F= Número de respuestas exactas en el grupo fuerte. D = Número de respuestas exactas en el grupo débil. N = Número total de estudiantes de estos dos grupos.

Si utilizamos una prueba con un conjunto de preguntas que tengan índices repartidos entre el 30 y el 70 por 100, su índice medio se situará alrededor del 50 por 100.

Hay autores que dan valores comprendidos entre 35 y 85 por ciento. Se ha demostrado que un test con índice medio de dificultad del 50 al 60 por 100 tiene grandes probabilidades de ser fiable en lo que concierne a su consistencia interna u homogeneidad.

INDICE DE DISCRIMINACIÓN.

Indice que permite determinar en qué medida una pregunta es bastante selectiva para distinguir un grupo fuerte de un grupo débil de estudiantes. Varía de -1 a +1.

Para su cálculo se utiliza la fórmula siguiente:

$$\text{Indice de discriminación} = 2 \times \frac{F - D}{N}$$

Cuando más elevado es éste índice, mejor permite la pregunta diferenciar un grupo fuerte y otro débil de estudiantes (para una población estudiantil determinada).

Si un test se compone de preguntas con altos índices de discriminación, asegura una clasificación discriminativa de los estudiantes según su nivel de actuación. Dicho de otra manera, el test no concede ventajas a los alumnos débiles con respecto a los fuertes. Es decir, que ayuda a reconocer a los mejores estudiantes.

Es un índice muy útil para la buena constitución de un banco de preguntas.

Apoyándose en este índice, pueden emitirse los siguientes juicios.

0,35 o más	= Pregunta excelente.
0,25 a 0,34	= Pregunta buena.
0,15 a 0,24	= Pregunta límite (para revisar).
menos de 0,15	= Pregunta mala que debe ser eliminada o reexaminada.

El cronograma de nuestro proyecto pasará por las siguientes etapas:

Elaboración individual o colegiada de las preguntas.
Revisión de la pertinencia de las preguntas.
Revisión colegiada de la prueba a impartir.
Análisis de resultados con cálculos de los índices de dificultad y discriminación.
Análisis post-evaluativo con grupo de alumnos.
Selección colegiada de las preguntas aceptables. Las preguntas pasan al banco.
Re-evaluación de las preguntas cuando sean seleccionadas para futuras pruebas
Mantener el "banco de preguntas" en dinamismo de calidad (feedback).
Facilitar su desarrollo a otros profesores interesados de otras asignaturas en el ámbito de nuestro ejercicio docente.

RESULTADOS

Tras pasar por el lector óptico todas las hojas de respuestas de los alumnos con su correspondiente numero identificatorio, el programa nos ofrece los siguientes datos:
PREGUNTAS (Tabla 1).

Relación numérica de las 80 preguntas de tipo test que constituyen el examen, facilitando de cada una:

Respuestas correctas (Previamente introducido la Hoja de respuestas corresponde a la " plantilla" con las testaciones correctas).
Número de alumnos que constituye grupos fuerte y débil diseñados.
Análisis en los grupos fuerte y débil po gunta del número de contestacion cada item posible, así como de las das en blanco, denunciándonos si s duce doble contestación.
Indice de dificultad:
- Grupo fuerte.
- Grupo débil.
- Media de los dos grupos.

Indice de discriminación.

ALUMNOS (Tabla 2).

Relación por alumno del total de con ciones acertadas, erróneas y en bl denunciando igualmente si se ha p cido doble contestación en alguna gunta.
Listado de calificaciones obtenidas. el programa nos pide el criterio de cación establecido -en nuestro ca damos al aprobado (eliminación de cial) el 6 sobre diez-, automáticar nos ofrece la calificación tanto cor pecto al aprobado 5 , como al 6 indi señalando los que no han superac aprobado. Listado que al tratarse d hoja de cálculo puede ordenarse bie numero de alumnos o por orden de tuación según deseemos.
Finalmente se nos ofrece el porcenta Sobresalientes, Notables, Aprobad Suspensos.

ANÁLISIS O EVALUACIÓN CRÍTIC

Se hace sobre la base de los ín obtenidos. Guilbert (1994) nos ofrec

ANALISIS DE RESPUESTAS E INDICES DE DIFICULTAD Y DISCRIMINACION POR PREGUNTAS

PREGUNTA Nº		1	2	3	4	5	-----
RESPUESTA CORRECTA		A	C	E	C	D	---
CONSTESTACIONES GRUPO FUERTE (G.F.)	A	13	0	0	0	0	
	B	6	3	0	1	2	
	C	2	20	8	18	0	
	D	0	0	3	3	23	
	E	4	1	13	0	0	
	BLANCO	0	1	1	3	0	
CONTESTACIONES GRUPO DEBIL (G.D.)	A	6	2	1	0	0	
	B	8	7	1	4	6	
	C	1	8	5	4	2	
	D	0	0	7	1	12	
	E	5	3	8	1	0	
	BLANCO	5	5	3	15	5	
INDICE DIFICULTAD (G.F)		52	80	52	72	92	
INDICE DIFICULTAD (G.D.)		24	32	32	16	48	
INDICE DIFICULTAD MEDIO		38	56	42	44	70	
INDICE DISCRIMINACION		0,28	0,48	0,20	0,56	0,44	
TOTAL ALUMNOS G.F. O G.D.		25					

TABLA 1

ANALISIS POR ALUMNOS

Nº del Alumno		1	2	3	4	5	.....
CONTESTACIONES	CORRECTAS	73	53	66	59	43	
	ERRONEAS	4	21	8	12	14	
	EN BLANCO	3	6	6	9	23	
	DOBLE	0	0	0	0	0	
PUNTUACIONES SOBRE 5		9	5,97	8	7	4,9	
PUNTUACIONES SOBRE 6		8,75	4,88	7,50	6,25	4,04	
PUNTUACION FINAL		8,75	4,88	7,50	6,25	4,04	
ELIMINACION EXAMEN			FALLO			FALLO	

TABLA 2

gráfico de valoración de los índices de gran utilidad (Fig. 2), de reproducción autorizada (OMS).

Los índices de dificultad y de discr nación, cuyo cálculo y valoración ya he comentado, se enriquecen al añadirl

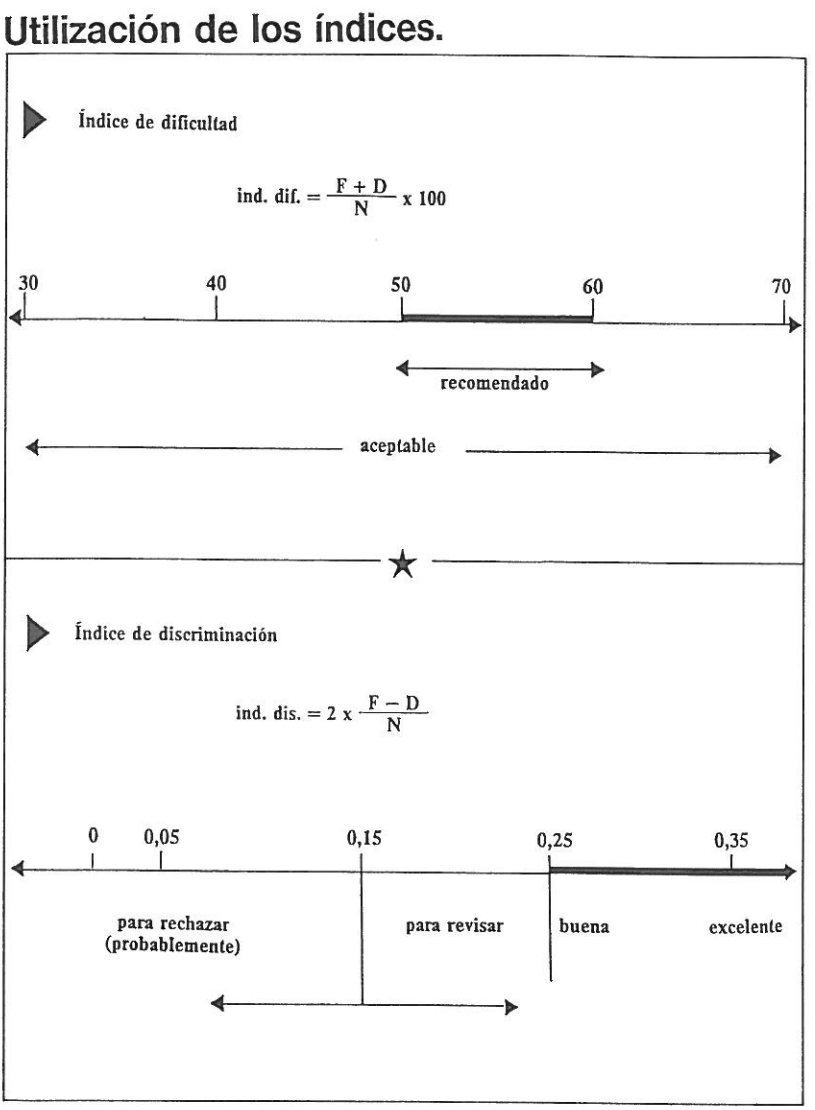

FIG. 2

número de respuestas de los distintos ítems por pregunta, lo que se efectúa en una hoja evaluativa de preguntas, diseñada al efecto (Fig. 3). Este análisis nos permite evidenciar aquellas preguntas conflictivas que han inducido a error a los alumnos, las mal for-

BANCO DE PREGUNTAS PEDIATRIA 5º CURSO
FICHA ANALISIS PREGUNTA

PROFESOR: GRUPO TEMATICO: TEMA Nº:

OBJETIVO EVALUATIVO:

PREGUNTA:

(*) R E S P U E S T A C O R R E C T A

FECHA EX. CONVOCATª	GRUPO	TAMAÑO GRUPO	A	B	C	D	E	S/C	INDICE DIFIC. (**)	INDICE DISCRIM (***)
	FUERTE									
	DEBIL									
	FUERTE									
	DEBIL									
	FUERTE									
	DEBIL									
	FUERTE									
	DEBIL									

OBSERVACIONES:

* RESPUESTA CORRECTA ** RECOMENDADO: 50-60%. ACEPTABLE: 30-70% (OTROS 30-85%)
*** > 0.35= EXCELENTE; 0.25-0,34 = BUENA; 0.15-0.24=REVISAR; < 0.15=ELIMINAR

FIG. 3

muladas o confeccionadas. En la misma hoja se prevén futuras evaluaciones en una estrategia de feedback ya referida.

Asimismo el análisis mencionado detecta las preguntas muy faciles o que todos los alumnos conocían, circunstancia que suele darse al proveerse éstos paulatinamente con parte de los bancos de preguntas; también ésto queda denunciado en nuestro seguimiento. Por supuesto hay preguntas que, aunque teóricamente rechazables por el resultado de sus índices, mantenemos ante la importancia de las mismas y para reforzar su captación por los alumnos.

Otro punto que nos permite valorar este análisis es la presencia de "cebos" inadecuados, al no ser elegidos por ningún alumno. Cuando las opciones están bien planteadas, suele haber una distribución uniforme de respuestas entre los cebos.

UTILIDAD APRECIADA EN EL PROCESO

El principal beneficio apreciado es el de reforzar en nuestros exámenes tipo PEM su carácter formativo, al permitir un progresivo incremento en la calidad de las cuestiones planteadas, potenciando las revisiones colegiadas tanto en el proceso de elaboración como en el de evaluación posterior a su aplicación.

Asimismo, las tutorías de revisión de exámenes se han visto igualmente enriquecidas, al disponer de múltiples datos que nos permiten profundizar en el proceso formativo de cada alumno.

Por supuesto, no podemos olvidar que la prueba de test motivo de análisis (PEM) nos es de utilidad en la identificación de aspectos como son la originalidad, la expresión verbal, la escritura y la capacidad de ción ante problemas reales, que req otros procedimientos de evaluación. tenerse en cuenta estos otros aspectos, mos encontrarnos con profesionales dos de conocimientos y faltos de la esenciales habilidades técnicas o de nicación, imprescindibles para realiz eficacia la función que les reclama la dad, para la que se precisa una forn consonante con las necesidades de sa la misma.

No olvidemos que la competencia sional se define como "el grado en q sujeto puede utilizar sus conocimi aptitudes, actitudes y buen juicio asoc a su profesión, para poder desempeña manera eficaz en todas las situacione corresponden al campo de la prac Michavila (1998) en el número de C Universitaria del 4 de Mayo publica nuestra Universidad, escribe un artícu relación a los exámenes en el que con la necesidad de búsquedas de métodos nativos o complementarios, que der puestas a los requerimientos de enrique mejorar los procedimientos de evalu de objetivos docentes. En dicha em nos hemos implicado.

REFERENCIAS

BEDOYA, J.M., DE URALDE, M.F.L.(1 *Pruebas para valoración del alumn Obstetricia y Ginecología*. Sevilla, 1 caciones de la Universidad de Sevilla.

CLIMENT BARBERÁ, J.M. (1985). Evalu de los saberes médicos mediante pr objetivas (I). *Tribuna Médica*, 1065, 9-

CLIMENT BARBERÁ, J.M. (1985). Evalu de los saberes médicos mediante pr objetivas (y II). *Tribuna Médica*, 106 21.

GARCÍA BABERO, M., ALFONSO ROCA, M.T., CANCILLO SALAS, J., CASTEJÓN ORTEGA, J.V.(1995). *Planificación educativa en Ciencias de la Salud.* Barcelona, Masson.

GALLO VALLEJO, F.J.(1998). La evaluación en la formación. *Boletín informativo. Oficina de Servicio a la Investigación. Consejería de Salud. Junta de Andalucía*, 31, 1-4.

GONZÁLEZ HACHERO, J.(1988). Evaluación de la enseñanza práctica. *Anales Españoles de Pediatría*, 29 (S33), 200-205.

GUILBERT, J.J.(1994). *Guía Pedagógica para el personal de salud.* Organización Mundial de la Salud / Universidad de Valladolid .

VELASCO, A.(1972). Evaluación y calificación de los estudiantes. *Archivo de la Facultad de Medicina de Madrid*, 22, 267-287.