Evaluación y validación de pruebas parciales de opción múltiple de un curso universitario de primer año

Rafael Arocena[1], Cecilia Gascue[1] y Julia Leymonié[2]

[1]Instituto de Biología, Facultad de Ciencias, Universidad de la República, Uruguay. E-mail: rarocena@fcien.edu.uy [2]Unidad de Enseñanza, Facultad de Ciencias, Universidad de la República, Uruguay.

Resumen: Para evaluar las pruebas de opción múltiple (2003-2005) de un curso de Introducción a la Biología de la Universidad de la República (Uruguay) se procedió al análisis: 1) estadístico de los resultados, 2) criterial de la formulación basado en juicio de experto, 3) cuantitativo de los índices clásicos de validación y 4) de correlación entre los anteriores. El análisis criterial incluyó la formulación, la figuración en los textos y la fidelidad al programa. El análisis cuantitativo incluyó los índices de dificultad y discriminación, el coeficiente de discriminación y un nuevo índice "de discriminación relativo." Tanto los resultados de los parciales como su nivel de adecuación disminuyeron entre el primer y segundo parcial de cada año, aunque menos en 2005 debido a algunos cambios introducidos en virtud de la reflexión que este tipo de estudio genera en el colectivo docente. Los índices clásicos también mejoraron en 2005, pero en ningún año se correlacionaron con los índices de formulación, ni lo hicieron éstos con los resultados de las pruebas. Esta falta de incidencia de la formulación de las preguntas en los resultados indicaría que las formulaciones incorrectas no impiden que la mayoría de los estudiantes comprenda la pregunta.

Palabras clave: análisis criterial, juicio de experto, índices de validación, dificultad, discriminación.

Title: Evaluation and validation of multiple choice tests in a first year university course

Abstract: To evaluate the multiple choice tests (2003-2005) of an Introduction to Biology course at the University of the Republic (Uruguay) we proceeded to analyse 1) the test results statistically, 2) the formulation of the answer options by means of expert judgement, 3) quantitative classic validity indices and 4) correlations between previous results. Expert judgement included adequacy of formulation, presence in a textbook, and fidelity to the program. Quantitative analysis included the indices of difficulty and discrimination, the coefficient of discrimination and a new index of "relative discrimination". The results of midterm tests and the "adequate formulation" diminished from first to second midterm test of each course, although less in 2005 due to the reflection that this kind of study produces in the faculty. The classic indices also improved in 2005, but they were never correlated to the formulation indices, nor were the latter correlated to the test results. This lack of influence of the formulation on the results could indicate that their poor formulation did not always impede the students' understanding of the question.

Keywords: criteria analysis, expert judgment, indices of validation, difficulty, discrimination.

Introducción

La evaluación del aprendizaje es una práctica común y necesaria en todo sistema educativo. Mediante la misma, los estudiantes, sus docentes y las autoridades de la institución educativa a la que pertenecen pretenden conocer el grado de aprendizaje de los primeros. De ser posible, muchas veces también interesa conocer las dificultades, problemas, aciertos u otras razones que explican dicho grado de aprendizaje. Si bien puede cumplir distintas funciones, la evaluación en la universidad está destinada generalmente a la sola calificación del estudiante, a efectos de habilitarle o no a proseguir con sus actividades curriculares.

Si bien la evaluación debería servir para orientar y motivar al alumnado y para informar al docente sobre su práctica formativa, muchas pruebas dejan de lado la comprensión de los complejos procesos educativos (Careaga y Rodríguez, 2002). Hoy se reconoce que la evaluación no debe ser un acto final ni paralelo, sino algo imbricado en el proceso de aprendizaje. Sin embargo, las innovaciones ocurridas en la enseñanza no han llegado en igual grado a la evaluación, donde hay una gran distancia entre la práctica y sus avances teóricos (Bordas y Cabrera, 2001). Así, se sigue empleando el concepto conductista de evaluación, como medida de resultados al final de un proceso, como si nada hubiera cambiado en el diseño curricular. Pero si la evaluación no cambia, el resto tampoco lo hace, ya que la misma puede llegar a dirigir todo el proceso educativo (Casanova, 2003).

Si bien hay voces en contra de la evaluación tal cual se la conoce hoy, por considerarla discriminatoria y arbitraria (Rivas y Ruiz, 2005) o por su escasa significación educativa (Méndez Vega, 2000), el actual sistema universitario parece impensable sin una práctica de evaluación de los estudiantes. Sin embargo, no por eso debemos perder de vista los importantes efectos, en general negativos, que tiene la evaluación sobre los estudiantes. Muchas veces los mismos están más preocupados por una buena calificación que por el conocimiento (Rivas y Ruiz, 2005), lo que les conduce a una "visión estrecha e instrumental del aprendizaje" (Godoy, 1995), el que pasa a depender del modo de evaluación (Fierro y Fierro-Hernández, 2000; Casanova, 2003). Asimismo es importante la insatisfacción que las formas y resultados de la evaluación suelen producir en muchos docentes (Alonso y otros, 1993) y también entre los alumnos. Según Casanova (2003) la evaluación tiene mala imagen porque se aplica al final, destaca lo negativo, se usa para clasificar, y en ella no participa el evaluado, quien solo debe aceptar el resultado.

Una de las metas de la enseñanza superior es desarrollar la capacidad de pensar y trabajar en forma independiente. Sin embargo se cuestiona hasta qué punto la evaluación favorece el logro de la misma (Godoy, 1995). Si aceptamos que debe existir cierta coherencia entre objetivos, contenidos y evaluación del aprendizaje, es necesario que nos detengamos en estos aspectos antes de abordar un sistema de evaluación concreto. Actualmente existe un acuerdo generalizado en que la formación de personas para que

en el futuro se enfrenten con éxito a situaciones diversas y novedosas, no requiere de un cúmulo interminable de información, sino del desarrollo de habilidades intelectuales que les permita adquirir y comprender conocimientos continuamente y de manera independiente, y aplicarlos ante situaciones complejas con pensamiento crítico (Haladyna y otros., 2002). En tal sentido debemos distinguir los diferentes contenidos del aprendizaje, y por ende de la evaluación: contenidos conceptuales, procedimentales (habilidades, técnicas) y actitudinales (López e Hinojosa, 2001). O bien, conceptos, destrezas y habilidades (Haladyna y otros, 2002).

Pruebas de opción múltiple

Existen muchos tipos de pruebas de evaluación, apropiados a los diferentes objetivos y contenidos que se deseen evaluar. Cuando se pretende una alta representatividad del conjunto de conocimientos, objetividad y atender a un alto número de estudiantes, se suele acudir a las llamadas pruebas objetivas como son las de opción múltiple (POM). Estas permiten además una rápida corrección y su consiguiente devolución a los estudiantes. Pueden ser empleadas no sólo para evaluar el conocimiento fáctico, sino habilidades más complejas como la capacidad analítica (Bush, 2001; Williams y Clark, 2004), pero no así para las destrezas (Haladyna y otros., 2002; López e Hinojosa, 2001). Cuando están adecuadamente planteadas, permiten evaluar incluso la comprensión de las bases teóricas que sustentan la resolución de problemas prácticos (Berezina y Berman, 2000) y distintos niveles cognitivos, desde la memoria y comprensión hasta la síntesis y valoración u otros contenidos complejos (López e Hinojosa, 2001; Alonso y otros, 1993). Contrariamente a la opinión de Godoy (1995), no serían necesariamente incompatibles con las metas de reflexión y análisis crítico.

Entre las críticas a las POM, se menciona que pueden ser contestadas correctamente por azar. Sin embargo, la incidencia del azar en el resultado puede ser perfectamente contemplada y minimizada. Esto se logra mediante un alto número de preguntas y de opciones, así como con la asignación de un puntaje negativo a las respuestas incorrectas. Las críticas de que no requieren que los evaluados sepan escribir o disertar, o que sólo valen para proposiciones indiscutibles y no para el pensamiento crítico al no permitir que el estudiante demuestre libremente su conocimiento y razonamiento, no consideran que no son estos sus objetivos, los que deberían ser evaluados por otros métodos. Más aún, lo que se pretende evaluar con estas pruebas es el conocimiento independientemente de su forma de comunicación. A las POM también se les critica la dificultad de diseñar buenas preguntas además del alto número de las mismas (Bush, 2001). El número elevado de preguntas es en realidad una ventaja de estas pruebas al permitir atender a la diversidad del público evaluado, aspecto no contemplado en general por la mayoría de los otros tipos de pruebas (López e Hinojosa, 2001).

A la falsa oposición entre los planteamientos del positivismo que sostienen que el único conocimiento válido es el científico sin tomar en cuenta el entorno, y los más naturalistas o cualitativos que sí lo toman en cuenta, es preciso encontrar un equilibrio entre ambas (Méndez Vega, 2000). Tal equilibrio no sólo pasa por la combinación de diversos

instrumentos de evaluación, sino por la continua valoración de los mismos. Varios autores mencionan entre las distintas fases del diseño de las pruebas (Alonso y otros, 1993; Buchweitz, 1996) al análisis de los resultados, que debe incluir la validación cuantitativa de la información proporcionada por la prueba, es decir la evaluación de la evaluación.

Un instrumento de medición es válido si mide lo que pretende medir. Cuando se evalúa si el nivel de conocimiento de un grupo de estudiantes es suficiente para aprobar la materia, resulta esencial que el instrumento de medición realmente esté midiendo el conocimiento de los estudiantes. En este trabajo se plantea la importancia de evaluar la calidad de las pruebas de opción múltiple planteadas a los estudiantes, de manera de poder validar su eficacia para medir el conocimiento. Para ello se toma como modelo el curso de Introducción a la Biología de la Facultad de Ciencias, Universidad de la República (Uruguay).

Históricamente se han obtenido bajos resultados en las pruebas parciales de este curso. En promedio, de 2003 a 2005, los estudiantes alcanzaron entre el 42 y el 64% de puntaje total. Este fenómeno llevó a que tanto docentes como estudiantes formularan diversos cuestionamientos sobre la forma y contenido de las preguntas planteadas. A efectos de saber si este bajo rendimiento obedecía a una inadecuada formulación de las pruebas, se procedió a realizar el presente análisis de las mismas.

Metodología

El curso Introducción a la Biología

El curso Introducción a la Biología se imparte en el primer semestre de las licenciaturas de Biología y Bioquímica. Abarca prácticamente todas las áreas de la biología con un propósito integrador y jerarquizador de los principales conceptos que luego se irán desarrollando a lo largo de la carrera de Biología. El programa comprende seis módulos de entre cuatro y seis clases cada uno. Las casi 30 clases teóricas son impartidas en dos clases semanales, por distintos docentes especialistas en cada tema. A su vez, cada módulo cuenta con un coordinador y existe un coordinador general. Además de las clases teóricas, el curso consta de Grupos de Discusión, en que semanalmente grupos de alrededor de 30 estudiantes discuten un artículo de divulgación científica referido al tema correspondiente del programa. Durante los tres años aquí analizados, la matrícula inicial estuvo en poco más de 600 estudiantes por año.

Estructura de los parciales

Durante cada año 2003, 2004 y 2005 se aplicaron dos parciales de opción múltiple. El primero, promediando el curso, incluyó 20 preguntas sobre esa primera parte, y el segundo al finalizar el curso, 40 preguntas en 2003 y 2004. Diez de estas 40 preguntas correspondían a la primera mitad del curso y las restantes 30 a la segunda, de modo que entre ambos parciales se formulaban 30 preguntas de cada mitad. De este modo el segundo parcial tenía un carácter globalizador, lo que permitía la exoneración del examen según los reglamentos vigentes. En 2005, a la vista de las evaluaciones realizadas, el sistema se modificó y ambos

parciales tuvieron 20 preguntas, correspondientes a cada mitad del curso en forma independiente, eliminándose la exoneración del examen.

Al menos tres preguntas de cada clase fueron elaboradas por el docente respectivo siguiendo una serie de criterios previamente establecidos. Luego eran revisadas, corregidas y una o dos de ellas seleccionadas primero por los coordinadores del módulo junto con el coordinador general. En todos los parciales se prepararon dos series con las mismas preguntas presentadas en distinto orden, al igual que las opciones, a efectos de dificultar que los estudiantes se copiaran las respuestas. En todos los casos cada pregunta incluía 5 opciones, de las cuales sólo una era la acertada. A cada pregunta bien contestada se le asignaron 5 puntos, mal contestada –1 y sin contestar 0 punto. Los estudiantes tuvieron 1 hora y media para resolver la prueba.

Análisis de las pruebas

El análisis de las pruebas parciales de opción múltiple consta de cuatro partes: 1) un análisis estadístico del desempeño de los estudiantes en las pruebas, 2) un análisis criterial de la formulación de las opciones basado en juicio de experto, es decir en la opinión fundada del responsable del curso, 3) un análisis cuantitativo clásico de los índices usuales de validación, y 4) las correlaciones entre los resultados anteriores.

Desempeño de los estudiantes en los parciales

La distribución de los resultados de los seis parciales analizados así como sus varianzas y sus medias fueron comparadas entre los estudiantes de ambas licenciaturas, entre años y entre el parcial I y II de cada año mediante los estadísticos F de Fischer y t de Student (Snedecor y Cochran, 1967). Los resultados de ambos parciales fueron correlacionados entre sí para cada año y licenciatura, y los parámetros obtenidos (coeficiente R^2 y pendiente b) fueron comparados empleando el programa Sigma Plot.

Formulación de las pruebas

Se emplearon tres criterios diferentes para evaluar la formulación de las preguntas. Estos criterios fueron aplicados a cada una de las 5 opciones, adjudicándole 1 punto si cumplía con el criterio y 0 puntos si no lo hacía. De este modo cada pregunta recibió un puntaje de 0 a 5 según cuántas de sus opciones cumplían con el criterio en cuestión. El primer criterio fue la "adecuación" de cada opción atendiendo específicamente a su claridad, pertinencia, nivel de especificidad en relación a los objetivos del curso, vaguedad, opciones contrapuestas que invalidaran al resto, etc. Además del número de opciones adecuadas, se registró por qué motivo una opción se consideraba inadecuada.

Un segundo criterio fue la "figuración" del concepto en uno de los libros de texto recomendados (Campbell y otros, 2001). Si bien no existe un único libro de referencia, éste fue el recomendado a los docentes para que se guiaran en la confección del temario de sus clases, y se lo toma simplemente como ejemplo de la bibliografía más común. Un tercer y último criterio empleado fue la "fidelidad" al temario, es decir que el tema estuviera claramente incluido en el temario del curso. De este modo cualquier estudiante, incluso sin asistir a las clases teóricas, que no son obligatorias, se podría haber preparado adecuadamente para las pruebas.

Por último, se sumaron los resultados de los tres análisis criteriales para obtener una valoración total de la formulación de cada pregunta.

Si bien este análisis tiene un fuerte componente de subjetividad, el mismo fue aplicado por la misma persona en todos los casos, lo que posibilita la comparación de los resultados aquí presentes entre sí. Además, esa persona –el primer autor de este trabajo-, fue el coordinador general del curso durante los años que aquí se analizan.

Índices de validación

Para este análisis se estudiaron los resultados de entre 264 y 446 estudiantes. Se analizaron los cuatro parciales de 2004 y 2005. En primer lugar se analizaron por separado los parciales de los estudiantes de Biología de los de Bioquímica, así como los pertenecientes a las dos series, obteniéndose entonces cuatro subgrupos en cada parcial.

La distribución de las opciones escogidas en cada pregunta fue comparada gráficamente entre subgrupos y luego entre series, antes de analizar la distribución para el total de estudiantes. En todos los casos se evaluó cada pregunta calculando su índice de dificultad, índice de discriminación y coeficiente de discriminación según Alonso y otros (1993) y Backhoff y otros. (2000). Asimismo, se propone y calcula un nuevo índice de discriminación, que llamamos índice de discriminación relativo.

a) Índice de dificultad

El índice de dificultad (ρ) de una pregunta se calcula como la proporción de personas que la responden correctamente respecto al total de estudiantes que hicieron la prueba (Backhoff y otros, 2000). Cuanto mayor es esta proporción, menor la dificultad de la pregunta. También se evaluó si la opción más escogida es la correcta, y si las otras han sido escogidas en grado suficiente para ser consideradas como distractores adecuados. Este índice permite conocer la adecuación de cada pregunta para evaluar, y discrimina cuáles, por muy fáciles o muy difíciles, deben ser eliminadas o modificadas porque no cumplen su función (Alonso y otros, 1993).

b) Índice de discriminación

El índice de discriminación (D) de cada pregunta fue calculado según:

$$D = (G_S - G_I)/N$$

donde G_S y G_I equivalen al número de aciertos en el 27% de personas con las puntuaciones más altas y más bajas respectivamente, y N equivale al número de aciertos más numeroso de esos dos grupos. Si una pregunta posee un alto índice de discriminación significa que discrimina a los alumnos que obtuvieron una alta calificación en toda la prueba de los que obtuvieron una mala calificación. Los primeros tienen alta probabilidad de haberla contestado bien mientras que los segundos es más probable que la hayan contestado equivocadamente (Backhoff y otros., 2000). El índice D toma valores entre -1 y +1, los que adquiere cuando sólo los peores y mejores alumnos respectivamente contestan bien la pregunta, y el valor 0 se adquiere cuando todos los alumnos la contestan por igual.

c) Coeficiente de discriminación biserial

El coeficiente de discriminación biserial ® determina el grado en que cada pregunta mide las mismas competencias que toda la prueba. Es la correlación de Pearson entre las calificaciones totales de la prueba y las de la pregunta, cuando ésta se dicotomiza en respuestas correctas e incorrectas (Henrysson, 1971 citado por Backhoff y otros., 2000). En el presente análisis se usó el puntaje 5, 0 y –1 para las correctas, no contestadas e incorrectas respectivamente, el mismo que se usó para la corrección de las pruebas. Este coeficiente corresponde al índice de homogeneidad de Alonso y otros (1993) pues se refiere a si la prueba evalúa un núcleo de conocimientos integrados, que es uno de los objetivos del curso.

d) Índice de discriminación relativo

Se propone el índice de discriminación relativo R, que a diferencia del anterior D, considera el número de personas que contestaron acertadamente la pregunta en relación al número de personas que integran los grupos del 27% inferior y superior. En este caso, en vez de dividir (G_S – G_I) sobre el número de aciertos más numeroso (generalmente G_S), se divide entre el total de integrantes de cada grupo (igual en ambos grupos, y para todas las preguntas). También varía entre –1 y +1, valores que adopta cuando todos los miembros del grupo inferior o superior respectivamente, contestan correcto. Este índice aporta la ventaja de indicar la proporción de respuestas correctas –al igual que el índice de dificultad- y permitir la comparación entre preguntas independientemente de cuántos estudiantes las contestaron.

Tanto los promedios de estos cuatro parámetros como sus valores para cada pregunta fueron comparados entre subgrupos mediante la prueba no paramétrica de Q para valores sospechosos. La prueba consiste en calcular la razón Q entre la diferencia del valor sospechoso y el valor más próximo sobre el rango de los valores. Este valor es comparado con valores tabulados para diferentes cantidades de mediciones al 90% de confianza y de esta manera comprobar si hay diferencias significativas entre los grupos analizados (Harris, 1995).

Relaciones entre los distintos resultados

Por último, y a efectos de saber si el desempeño de los estudiantes se relaciona con el planteamiento de las pruebas, se realizaron correlaciones entre los índices clásicos y los indicadores de la formulación de las pruebas, así como de los índices clásicos entre sí.

Resultados

Desempeño en los parciales

Se observó una disminución de 7.5% en promedio en el número de estudiantes que se presentaron al segundo parcial en los tres años analizados (Tabla 1). Esta disminución fue en aumento año tras año, del 5 al 9% para el conjunto de ambas licenciaturas. Sin embargo, al analizar éstas por separado se observó que sólo los estudiantes de Bioquímica muestran esta tendencia, donde la disminución aumenta

del 3 al 13%. Estos estudiantes siempre fueron algo más de la mitad de los estudiantes de Biología.

También la media de los resultados de las pruebas parciales disminuyó entre los parciales I y II en los tres años analizados (Tabla 1). En ambas licenciaturas tal disminución fue de 60 a 46 % en el promedio de los tres años. Pero esta disminución fue descendiendo año tras año para los estudiantes de Biología y en el último año también para los de Bioquímica.

Año	2003		2004		2005		MEDIA	
Parcial	I	II	I	II	I	II	I	II
Puntaje total	100	200	100	200	100	100		
Biología								
Número de estudiantes	321	300	255	230	279	261	285	264
Media (puntaje)	59	83	58	87	64	53	60	46
Mediana (puntaje)	60	84	58	85	65	53	61	46
Moda (puntaje)	82	71	58	72	94	47	78	40
Desvió estándar (puntaje)	18	33	18	31	21	19	19	17
Coef. Variación (%)	31	40	31	36	33	37	32	38
Q1 (25%) (puntaje)	47	60	46	64	51	40	48	55
Q3 (75%) (puntaje)	73	104	71	104	79	69	74	93
Bioquímica								
Número de estudiantes	166	161	150	143	169	147	162	150
Media (puntaje)	57	87	59	83	64	52	60	46
Mediana (puntaje)	58	89	59	80	64	52	60	46
Moda (puntaje)	70	100	76	62	88	76	78	52
Desvió estándar (puntaje)	18	32	19	33	21	20	19	17
Coef. Variación (%)	32	36	32	40	33	38	32	38
Q1 (25%) (puntaje)	46	68	47	61	50	37	48	55
Q3 (75%) (puntaje)	70	107	72	104	83	69	75	93
Biología y Bioquímica								
Número de Estudiantes	487	461	405	373	448	408	447	414
Media (puntaje)	58	85	58	85	64	53	60	46
Desvió estándar (puntaje)	18	33	18	32	21	20	19	17
Coef. Variación (%)	31	39	32	37	33	37	32	38
Mínimo (puntaje)	4	4	7	-4	-8	-2	1	-1
Máximo (puntaje)	100	178	100	172	100	100	100	150

Tabla 1.- Estadísticos de los resultados obtenidos en los seis parciales de los tres años analizados. Se presentan los resultados separados por carrera y en conjunto.

Las medias en todos los parciales fueron muy similares a las medianas, mostrando una distribución bastante simétrica de los resultados (Figura 1). Las modas en cambio siempre difirieron de las medias, excepto en el parcial I de 2004, estando a veces por debajo y otras por encima de aquéllas, sin un patrón claro de comportamiento.

En las pruebas de 100 puntos, el desvío estándar varió entre 18 y 21 puntos, y en las de 200, entre 31 y 33 puntos (Tabla 1). El coeficiente de variación lo hizo entre 31 y 40% y aumentó siempre del parcial I al II. Los valores mínimos fluctuaron entre –8 y 7, mientras los máximos alcanzaron 100 en el primer tipo de pruebas, pero no pasaron de 178 en el segundo. El primer cuartil (Q_1) indica que el máximo puntaje obtenido por el 25% de los estudiantes con peor desempeño, osciló entre 37 y 51 puntos en las

pruebas de 100 y entre 60 y 107 en las de 200. El mínimo puntaje obtenido por el 25% de estudiantes con mejor desempeño (Q_3) osciló entre 70 y 83, y entre 104 y 107, respectivamente.

La distribución de los resultados de los parciales fue muy similar para los estudiantes de Biología y los de Bioquímica en los tres años analizados (Figura 1). También fueron similares las distribuciones obtenidas para los años 2003 y 2004. En estos casos el primer parcial se distribuyó con sesgo a la derecha, y el segundo a la izquierda, evidenciando la mayor dificultad que implicaba. En 2005 en cambio, cuando el segundo parcial sólo incluía 20 preguntas y éstas correspondían a la segunda parte del curso, la diferencia entre parciales no fue tan clara. Si bien el primero mostró una distribución similar a la de los otros años, el segundo es casi simétrico.

Los resultados obtenidos por los estudiantes de Biología y Bioquímica en los seis parciales no mostraron diferencias significativas entre ellos, tanto de varianza como de media. En los tres años analizados, el parcial II presentó menores medias ($t>3.37$, $p<0.001$) que el parcial I, incluso en 2005 cuando ambos parciales constaban de 20 preguntas. Los resultados de ambos parciales mejoraron del año 2003 al 2005 y del 2004 al 2005 ($t>3.37$, $p<0.001$), pero no fueron diferentes entre 2003 y 2004. Los puntajes de los parciales I y II estuvieron altamente correlacionados entre sí ($r > 0.419$, $p<0.001$) para ambas licenciaturas y en los tres años (Tabla 2), mostrando que los estudiantes en general mantienen un mismo desempeño a lo largo del curso.

Formulación de las pruebas

Los resultados obtenidos por parcial y año para cada uno de los criterios de adecuación, figuración, fidelidad y total se presentan en la tabla 3. El nivel de "adecuación" de las opciones en 2003 bajó del 81% del primer parcial al 66% en el segundo, aunque la primera parte del curso, presente en ambos parciales, se mantuvo casi igual, en el 80%. Este nivel también bajó en 2004, pero del 73 al 67%, también manteniéndose similar la primera parte (72%). En cambio aumentó de 77 a 93 en 2005, cuando cada uno correspondía a una parte diferente del curso. El promedio ponderado de adecuación de ambos parciales de cada año resultó 71 y 69% en 2003 y 2004 respectivamente, y subió a 85% en 2005 (datos no mostrados).

El nivel de "figuración" de conceptos en el libro en 2003 bajó de 81 a 65% entre ambos parciales, a pesar de que la primera parte del segundo, referida a los primeros temas del programa, se mantuvo en 80%. En 2004 también bajó, pero de 60 a 53%, a pesar de un pequeño ascenso a 64% en la primera parte del segundo parcial. Es decir que en ambos años, el descenso se debió a la segunda parte, al igual que con la "adecuación." Lo mismo sucedió en 2005, cuando la figuración disminuyó de 89 a 81% (Tabla 3). El promedio ponderado de figuración en ambos parciales resultó 70% en 2003, 56% en 2004 y 85% en 2005 (datos no mostrados).

Por último, el nivel de "fidelidad" al temario fue muy bajo en 2003, cuando subió de 17 a 29% en el segundo parcial, a pesar de que la primera parte bajó a 10%. De hecho, el módulo I carecía de temario, en el II éste no estaba detallado, y en todo el resto del programa, el temario era muy general. En el segundo año en cambio fue similar a los otros dos índices,

pero bajó del 71 al 56% entre parciales, debido tanto al descenso de la primera parte a 66% como a la baja "fidelidad" de la segunda (53%). En 2005 fue más alto en ambos parciales (98 y 87%). El promedio ponderado de fidelidad en ambos parciales fue 25% en 2003, 61% en 2004 y 92% en 2005 (datos no mostrados).

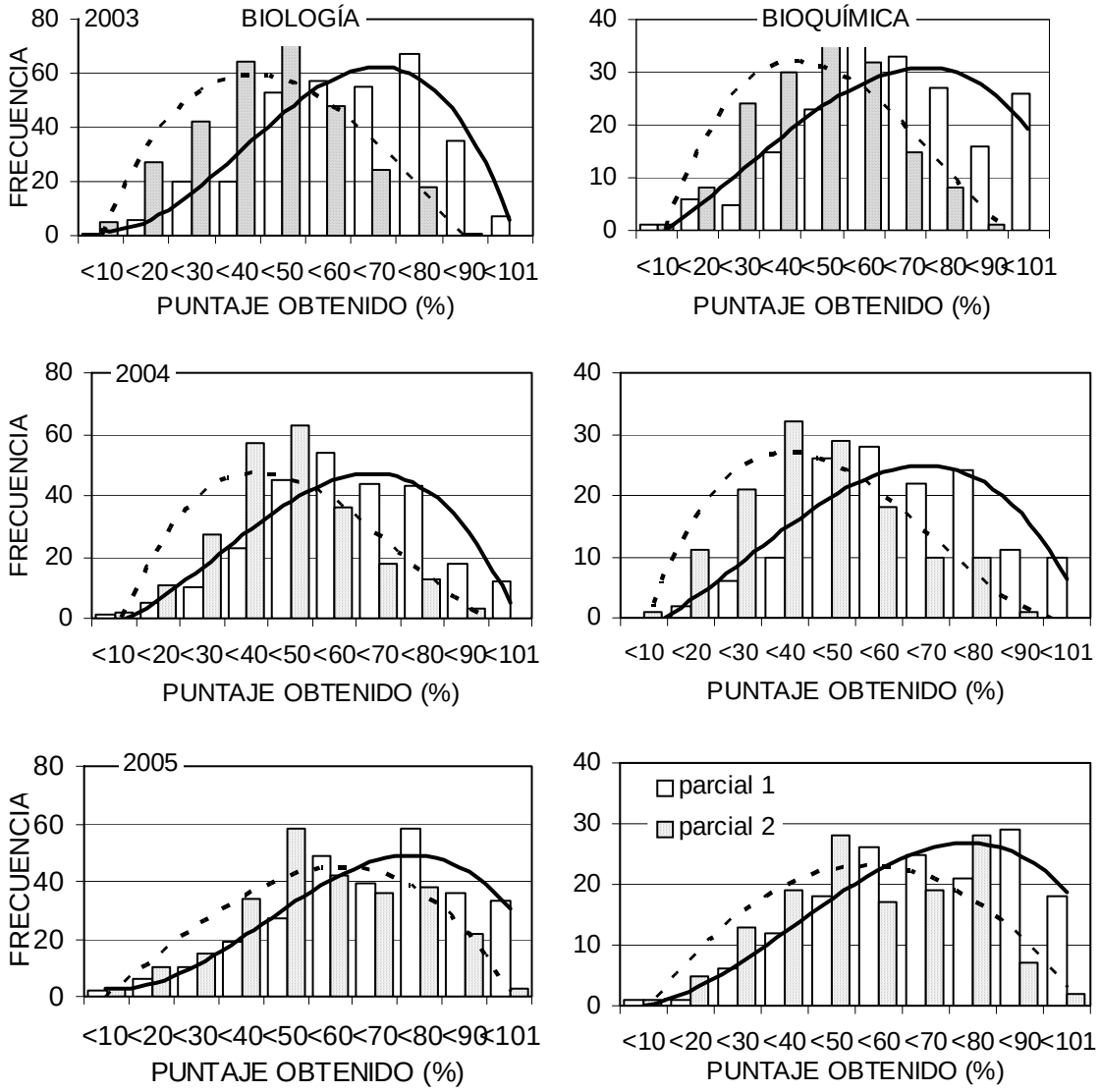

Figura 1. Histograma de distribución de frecuencias de los resultados de los parciales de estudiantes de Biología (izquierda) y Bioquímica (derecha) de 2003, 2004 y 2005 (de arriba a abajo). Primer parcial: barras claras y línea de tendencia continua. Segundo parcial: barras sombreadas y línea de tendencia punteada.

Año	Biología	Bioquímica
2003	0,6635	0,5464
2004	0,5405	0,6171
2005	0,5296	0,4190

Tabla 2.- Coeficientes de correlación de Pearson entre los resultados de los parciales I y II para estudiantes de Biología y Bioquímica durante 2003-2005.

Índices de validación

Promedios de los índices por parcial

La comparación de los promedios por parcial de los índices de validación no arrojó diferencias entre licenciaturas ni entre series. Por lo tanto se procedió al análisis de los resultados del conjunto de estudiantes que rindieron cada prueba. Estos índices resultaron similares entre todos los parciales ($t<1$, $p>>0.2$) (Figura 2).

	2003			2004			2005		
Módulos	**Adec.**	**Fig.**	**Fid.**	**Adec.**	**Fig.**	**Fid.**	**Adec.**	**Fig.**	**Fid.**
Parcial 1									
% Total	81	81	17	73	60	71	77	89	98
Parcial 2									
% 1ª parte	80	80	10	72	64	66			
% 2ª parte	62	60	36	65	50	53	93	81	87
% Total	66	65	29	67	53	56			

Tabla 3.- Análisis de validez de las pruebas parciales del curso Introducción a la Biología. Porcentaje de opciones con adecuada formulación (Adec.), figuración en el libro (Fig.) y fidelidad al programa (Fid.) totales por parcial y por partes del segundo parcial si corresponde.

La mayor parte de los problemas detectados en los dos primeros años en las opciones catalogadas como inadecuadas se referían al grado de especificidad excesiva de las preguntas (Tabla 4). En 2003 el segundo lugar de los problemas lo ocupó el uso de términos muy técnicos, incluyendo en esta categoría preguntas que involucran directa o indirectamente definiciones convencionales y arbitrarias. Este problema disminuyó al año siguiente, cuando su lugar lo ocupan las formulaciones confusas o que llevan a confusión, lo que a su vez fue el primer motivo de inadecuación en 2005. Ese año disminuyeron o desaparecieron la mayor parte de los problemas de formulación. Entre otros problemas menos frecuentes, pero que también inducen a confusión al estudiante, está la mezcla de temas de diferentes módulos (Tabla 4).

Valores de los índices para cada pregunta

En el parcial I de 2004, las tres preguntas más accesibles (mayor □) presentaron los menores índices de discriminación, tanto absoluta (D) como relativa (R) pero coeficientes de discriminación (r) intermedios. Asimismo, las tres preguntas menos accesibles fueron las de mayor discriminación absoluta D, pero no relativa R, y nuevamente coeficientes r intermedios. El coeficiente de discriminación del primer parcial de ambos años resultó significativo ($p<0.01$) para todas las preguntas, lo que no sucedió en algunas preguntas de los segundos parciales.

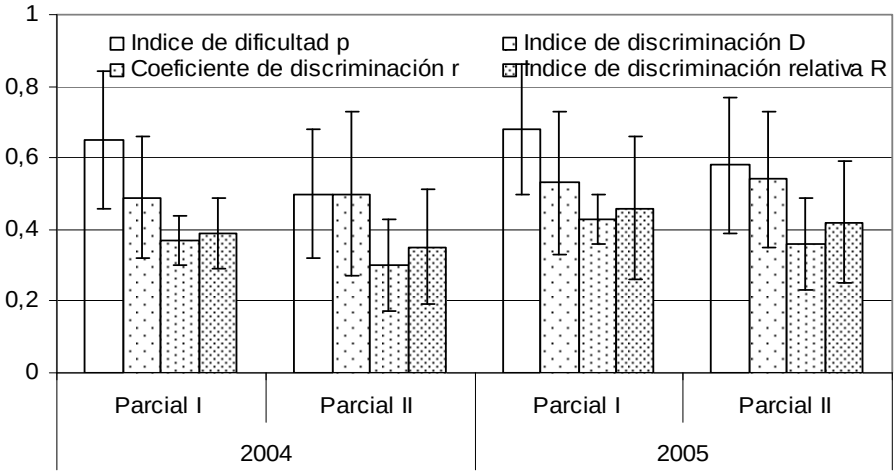

Figura 2.- Promedio y desvío estándar de los 4 índices en cada parcial.

Diferencias entre licenciaturas y entre series

A diferencia de lo que ocurrió con los promedios, al comparar los índices clásicos para cada pregunta, se observaron algunas diferencias entre los cuatro subgrupos formados por estudiantes de ambas licenciaturas y ambas series. En el parcial I de 2004, una pregunta sobre el experimento de Miller del origen de la vida y otra acerca del gen eucariota, presentaron en uno de los subgrupos de estudiantes de bioquímica, un índice de dificultad menor y mayor respectivamente que en los otros tres subgrupos, así como un coeficiente de discriminación no significativo ($p>0.01$). Sin bien casi todas las preguntas presentaron valores no significativos de este coeficiente en uno o más de los subgrupos, sólo una lo hizo en los cuatro subgrupos.

	2003	2004	2005
Conceptos demasiado específicos	23	43	4
Inclusión de términos inusuales o preguntas sobre definiciones	18	11	3
Formulaciones confusas o que inducen a confusión	8	25	14
Introducción de conceptos de otra parte del temario	7	2	6
Conceptos no biológicos	5	4	
Opciones opuestas	7	2	
Afirmaciones muy obvias	1	1	
Conceptos o enfoques opuestos o diferentes a los del libro	1	2	1
Afirmaciones vagas		3	
Afirmaciones subjetivas		3	
Errores de lenguaje			4
TOTAL	70	96	32

Tabla 4.- Frecuencia de los diferentes problemas detectados en las opciones.

El ordenamiento de las preguntas por el índice de dificultad fue diferente entre los cuatro subgrupos de ambos parciales de 2005. Sin embargo, en el segundo parcial, la pregunta más "difícil" en todos los casos fue la referida a

la energía en los ecosistemas (ρ=0.09-0.15), y la más "fácil" la referida a la sinapsis neuronal (ρ=0.89-0.99). Una pregunta de macro-evolución –también de las más difíciles- y la ya mencionada sobre energía en ecosistemas tuvieron coeficientes de discriminación no significativos en 3 de los cuatro subgrupos.

Análisis de la selección de opciones

En los cuatro parciales analizados se observó gráficamente que no existían diferencias relevantes en la distribución de opciones entre las licenciaturas ni entre las series, por lo que también aquí se procedió al análisis del total de estudiantes. Sólo en una pregunta del parcial I de 2004 la opción más escogida (por el 36% de los estudiantes) no fue la correcta (escogida por el 27%).

			adec.	figur.	fidel.	total	ρ	D	r
2004	Parcial I	ρ	-0,017	0,310	0,083	0,070			
		D	0,088	-0,112	-0,048	-0,028	-0,880***		
		r	0,097	0,264	-0,081	0,122	-0,009	0,432	
		R	0,062	0,054	0,068	0,074	-0,513*	0,812***	0,630**
	Parcial II	ρ	0,078	0,073	0,228	0,177			
		D	0,007	-0,336	-0,376	-0,324	-0,270		
		r	0,080	-0,256	-0,184	-0,164	0,202	0,818***	
		R	0,165	-0,269	-0,200	-0,139	0,192	0,811***	0,972***
2005	Parcial I	ρ	-0,174	0,097	0,317	-0,014			
		D	0,142	-0,138	-0,297	-0,031	-0,946***		
		r	0,011	-0,137	-0,109	-0,094	-0,387	0,648	
		R	0,066	-0,102	-0,232	-0,056	-0,855***	0,968***	0,779***
	Parcial II	ρ	0,358	-0,004	-0,030	0,072			
		D	-0,169	-0,096	0,145	-0,009	-0,492*		
		r	0,101	-0,232	0,088	-0,039	0,247	0,668**	
		R	0,011	-0,140	0,144	0,015	0,031	0,827***	0,943***

Tabla 5.- Coeficientes de correlación de Pearson entre los criterios de formulación (adecuación: adec., figuración en el libro: figur. y fidelidad al programa: fidel.) y los índices de dificultad (ρ) y discriminación (D) y coeficiente de discriminación biserial (r) y relativo (R). ***p<0,001; **p<0,01; *p<0,05

Relaciones entre los distintos resultados

Las correlaciones realizadas para los cuatro parciales de 2004 y 2005 entre cada uno de los tres índices de formulación y los cuatro índices clásicos resultaron todas no significativas (Tabla 5). En cambio, varias correlaciones de estos últimos entre sí resultaron significativos. En los cuatro parciales analizados las medidas de discriminación r y R estuvieron altamente correlacionados, así como el índice de discriminación D y el índice de discriminación relativa R. En los dos parciales I también lo estuvieron los índices de dificultad ρ y discriminación D, y en los dos parciales II, el índice D y el coeficiente r de correlación. En solo una ocasión se correlacionaron el índice de dificultad ρ y el índice de discriminación relativo R.

Tampoco se encontró correlación entre la formulación de las seis pruebas (2003-2005) medida como porcentaje de opciones adecuadas o número de preguntas adecuadas y los resultados expresados como calificación promedio o porcentaje de estudiantes que obtuvieran más de 50% del valor de la prueba (Figura 3).

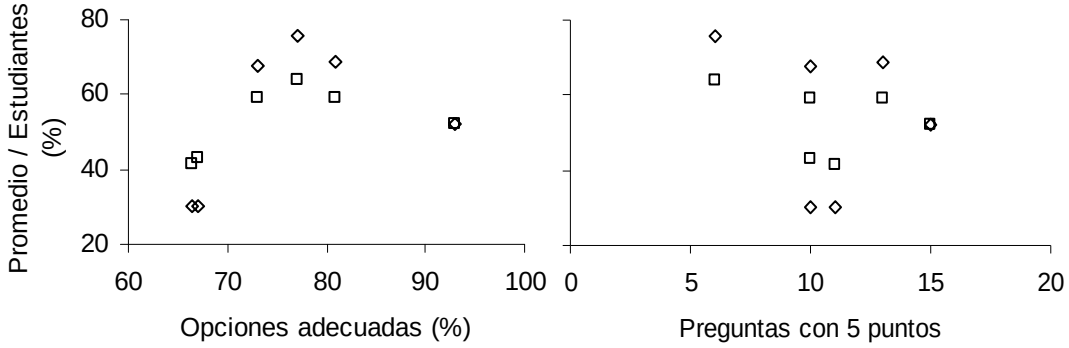

Figura 3.- Correlaciones entre la formulación de las 6 pruebas 2003-2005 como porcentaje de opciones adecuadas (izquierda) o número de preguntas adecuadas (derecha) y los resultados como calificación promedio (cuadrados) o porcentaje de estudiantes que obtuvieran más del 50% del valor de la prueba (rombos).

Discusión

El presente estudio comprende un alto número de pruebas individuales (2583), lo que brinda un tamaño de muestra apropiado para su análisis estadístico (N>373). La evaluación de los resultados de los parciales II de los dos primeros años permitió realizar correcciones en los del tercer año –menos preguntas y sólo de la segunda parte del curso-, que se vieron reflejadas en una mejor distribución de los resultados.

Los análisis realizados no son suficientes para encontrar las causas del aumento sucesivo de la deserción entre los parciales I y II, lo que sucede sólo entre los estudiantes de Bioquímica. Esto parece contrapuesto a la igualdad de los resultados entre años y entre carreras. Si bien las razones de este hecho están fuera del alcance del presente trabajo, los resultados nos permiten descartar que las causas se relacionen con un diferente rendimiento académico. La alta correlación entre los parciales I y II en ambas carreras, lleva también a descartar que posibles diferencias en su formulación expliquen el menor rendimiento en la segunda. Descartadas estas razones se deberán buscar las mismas por el lado de la motivación. El segundo parcial se refiere a los niveles de organización biológica supracelular, y por lo tanto más alejados del interés de la bioquímica. Esto podría generar una predisposición en contra de tales temas por parte de los estudiantes, que los lleve a descuidar su preparación.

Los análisis en cambio, permitieron avanzar de manera importante en la explicación del descenso –cada vez menor en Biología-, en los resultados entre ambos parciales. Junto a este descenso en el resultado promedio se da un aumento del coeficiente de variación, medida de la heterogeneidad de los desempeños, a pesar del menor número de estudiantes. En particular, ningún estudiante alcanzó una calificación siquiera cercana a la máxima en

los parciales II (200 puntos). En tales instancias los mejores estudiantes apenas lograron superar el 50% de la calificación máxima. Sin embargo, la distribución casi simétrica de los resultados muestra una población estudiantil en general equilibrada entre los de mejor y peor desempeño, con un predominio de aquéllos con habilidades y dificultades intermedias.

Si bien la formulación de ambos parciales no explica el peor desempeño en el parcial II de los estudiantes de una carrera respecto a la otra, sí puede ser una causa del pobre desempeño de todos los estudiantes en general, especialmente en el parcial II. Para verificar esta hipótesis se desarrolló y aplicó un sistema de análisis de la formulación de las pruebas. Mediante dicho análisis se verificó a) que la formulación de los parciales II era menos adecuada que la de los parciales I en 2003 y 2004, b) que la diferencia radicaba en la segunda parte de la prueba, donde se evaluaban los conceptos nuevos respecto a la primer prueba y c) que se mejoró notoriamente con la corrección introducida en 2005. Esta corrección consistió en no introducir nuevamente preguntas sobre la primer parte, ya evaluada, del curso, con lo que además se redujo el número de preguntas, quedando ambos parciales de igual extensión.

Si bien este criterio se basa en una apreciación personal y por lo tanto subjetiva de la formulación de las preguntas y sus opciones, el procedimiento reúne ciertas características que lo hacen apropiado para la valoración, especialmente para una comparación de las distintas pruebas. En primer lugar, la evaluación es hecha por la misma persona –en este caso el primer autor y coordinador general del curso-, y al mismo tiempo, evitando así las diferencias en las condiciones personales y ambientales que pudieran influir en el juicio. Este se encuentra basado en criterios previos, emanados de los recaudos que deben tomarse al formular las POM según los especialistas. Por último, la valoración se realizó estableciendo las razones que llevaron a la misma, obligando a una reflexión sobre la estructura y el contenido de cada ítem.

Un criterio mucho más objetivo es el de la figuración de los temas en la bibliografía, excepto por la selección de ésta, que se limitó al libro recomendado a los docentes. Tanto en este caso como en el anterior, la disminución registrada hacia los parciales II en 2003 y 2004 está en consonancia con la también disminución del desempeño en los mismos, y posiblemente sea una de las explicaciones de ésta.

El aumento año a año de la figuración de los temas de las preguntas en el programa puede tener dos explicaciones no excluyentes. Por un lado fue mejorando la elaboración del programa conforme se incorporaban en el mismo las correcciones introducidas cada año y un mayor grado de detalle en el contenido de las clases. Pero también es posible que la formulación de las pruebas fuera teniendo cada vez más en cuenta los contenidos del programa, lo que por otra parte sólo fue posible en tanto se dispusiera con un programa más completo, lo que vuelve a remitirnos a la primera razón.

En los dos primeros años analizados, la causa más frecuente de inadecuación de las preguntas fue su alta especificidad. El objetivo de un curso general o introductorio consiste justamente en la incorporación de conceptos generales. Sin embargo, muchos docentes –que además son especialistas en los temas que imparten-, suelen propender al

planteamiento de interrogantes muy específicas, que aunque alejadas de este objetivo central son más fáciles de formular, además de numerosas y variadas, ampliando así el espectro de posibilidades. La segunda causa durante el primer año está muy relacionada con ésta, ya que los términos muy técnicos y las definiciones convencionales son necesidades que surgen de la especificidad y detalle de los asuntos tratados. Al segundo y tercer año se imponen en su lugar los problemas de redacción clara y correcta. Estos problemas están asociados a un pensamiento más complejo y difícil de redactar, pero a la vez más abarcativo y menos concreto, que intenta superar el recurso de la pregunta fácil sobre detalles que difícilmente el estudiante retenga.

Se han detectado otros problemas que no necesariamente inciden negativamente en el desempeño de los estudiantes en las pruebas, sino que por el contrario pueden favorecerlo o simplemente no afectarlo en ningún sentido. La inclusión de temas o conceptos no biológicos se refiere a las preguntas tradicionalmente enmarcadas en otras disciplinas como física, química y astronomía. Aunque relacionadas en mayor o menor grado con problemas biológicos, no tratan de los problemas que estudia la biología. Las opciones opuestas son aquellas que necesariamente incluyen la respuesta correcta en una de ellas, permitiendo que el estudiante descarte las otras opciones. En otros casos se detectaron afirmaciones que eran obviamente correctas o incorrectas, facilitando así su elección y favoreciendo al estudiante. Por el contrario, las afirmaciones muy vagas, subjetivas o claramente diferentes a lo expresado en los libros recomendados, dificultan el desempeño perjudicando a los estudiantes.

Si bien la causa relativa a la mezcla de temas de diferentes módulos puede no ser considerado como problema, sino como un intento de integración de los diferentes temas, no es así en este caso debido a la estructura de los parciales. Estos presentan un ordenamiento de las preguntas que sigue el ordenamiento del temario, así como una cantidad equitativa de preguntas por módulo. En consecuencia, la mezcla de temas y su cambio de ordenamiento puede provocar confusión en un estudiante que está esperando lo contrario. Si bien es frecuente la inclusión de preguntas fuera del programa o de los objetivos o preguntas muy difíciles u obvias, o con fallas técnicas en su elaboración (Buchweitz, 1996), es preciso realizar todos los esfuerzos posibles a fin de evitarlo.

El análisis cuantitativo clásico tampoco arrojó diferencias entre las dos carreras ni entre las dos series utilizadas. Este hecho, además de validar los resultados, muestra que la parte de formulación de las pruebas que se refiere al orden en que se presentan las preguntas y las opciones (series), no repercute en el rendimiento obtenido.

Los resultados de este análisis concuerdan con los antes mencionados: mejores en 2005 que en 2004 y mayor dificultad en el parcial II de ambos años. Pero este análisis es totalmente objetivo, basado en los resultados de las pruebas de todos los estudiantes, lo que confirma la validez del análisis subjetivo de la formulación de las pruebas.

En general las preguntas más accesibles (mayor ρ) son las menos discriminantes y viceversa, como era de esperar. Sin embargo, hay diferencias en el comportamiento de los distintos índices según el parcial

analizado. La concordancia del índice de dificultad con la discriminación absoluta, que sólo considera el 27% de estudiantes con puntuaciones más altas y más bajas, pero no con la discriminación biserial, que se basa en las pruebas de todos los estudiantes, indica que son los extremos de la distribución los que imponen el grado de dificultad. Por otra parte, la falta de concordancia entre dificultad y discriminación relativa, que considera el número de personas que contestaron acertadamente la pregunta y por tanto varía con la misma en vez del número constante de integrantes del 27% inferior y superior, indica que la incidencia de los extremos varía con las preguntas. De hecho, el resultado de algunas preguntas tampoco se correlaciona con el resultado total de los parciales II.

El mayor grado de dificultad de las preguntas puede obedecer a distintas razones. Cuando un alto porcentaje de estudiantes no la contesta se presume que existe un amplio desconocimiento del tema. En cambio, cuando las respuestas se distribuyen equitativamente entre todas las opciones, es de suponer que el tema fue abordado pero de manera incorrecta, dando origen a las más diversas interpretaciones. Por último, si existe una opción errónea que es ampliamente seleccionada, el abordaje –ya sea por parte de los docentes o los estudiantes, o incluso los libros-, adoleció de un error sesgado hacia un falso mensaje, siempre que la formulación haya sido correcta. También puede darse una combinación de situaciones, cuyo conocimiento y análisis permiten indagar en las causas de los errores e introducir las correcciones que correspondan.

En el parcial II de 2004 fueron tres las preguntas donde la opción más escogida no fue la correcta, tal vez debido a su mala formulación. Los coeficientes de discriminación no fueron significativos en estos casos mostrando que ni siquiera los mejores estudiantes las contestaron correctamente. En 2005 sólo una pregunta tuvo valores de discriminación bajos, en correspondencia con los mayores índices de dificultad (la más fácil). Si bien en general las preguntas más fáciles discriminan menos, no existe correlación entre ambas propiedades, excepto allí donde además las más difíciles discriminan más, y esto debido al efecto de los puntos extremos en cualquier correlación.

Si bien los subgrupos formados por las dos licenciaturas y las dos series en general no muestran diferencias en el desempeño de los índices, sí lo hicieron en un caso. Esta y otras diferencias de este subgrupo pueden estar relacionadas a su menor tamaño que, aunque no significativo, lo hace menos representativo que los otros del conjunto del estudiantado.

Aunque los cuatro subgrupos presenten similares valores medios en los parámetros, siempre existen diferencias entre ellos cuando se analizan las preguntas individualmente. De aquí también surge la necesidad de contar con un número alto de preguntas a efectos de contemplar estas diferencias, y por lo tanto de tener que recurrir al formato de múltiple opción.

El modelo de calificación empleado no toma en cuenta el grado de dificultad de cada pregunta ni el grado de error de la respuesta equivocada. Aunque se ha investigado mucho acerca de si este sistema de evaluación es apropiado, Backhoff y otros. (2001) demostraron que ponderar los errores y los aciertos según su grado de dificultad no mejora la validez de la prueba.

La falta de correlaciones significativas entre los indicadores de formulación de las pruebas y los índices con que se analizan los resultados nos muestra que la forma correcta o incorrecta de formulación de la prueba no incidió en los resultados. Si bien esto parece ir en contra del sentido común, hay algunas explicaciones posibles. Por un lado el alto número de preguntas planteadas hace que sea limitado el impacto que las malas formulaciones puedan tener en el conjunto. Por otro lado, aún estando mal formulada, una pregunta puede ser comprendida por el estudiante, que hace uso de su sentido común y de sus conocimientos –especialmente aquellos mejor preparados- para imprimirle la significación correcta más allá de los errores semánticos que pueda contener. Este resultado fue refrendado con otras correlaciones entre las seis pruebas analizadas.

El análisis de las correlaciones entre los índices clásicos muestra un resultado hasta cierto punto esperable. Las distintas medidas de discriminación de las preguntas son las más altamente correlacionadas debido a que miden aunque de distinta forma, la misma propiedad. Solamente en los parciales I se correlaciona dificultad con discriminación. Esto podría indicar que en ese caso las preguntas más dificultosas sólo son accesibles a los mejores estudiantes, mientras que en los parciales II hay una mayor incidencia de algún factor que independiza ambos aspectos. Ese factor en 2004 pudo ser el mayor número de preguntas o la inclusión en el parcial II de preguntas correspondientes al parcial I, pero en 2005 esto no sucedió, por lo que sólo quedan la diferencia temática y el diferente momento del curso (al final del mismo) como posibles explicaciones. En cuanto a los temas, el parcial II implica un pasaje a los mayores niveles de organización biológica (de sistemas de órganos a biosfera) y un abordaje de temas más novedosos para el estudiante proveniente de secundaria: biodiversidad, taxonomía, evolución y ecología. El momento del curso en que es planteada es al final del mismo, cuando se juntan las pruebas de varias materias, puede haber un mayor cansancio y el ánimo cambia ante la proximidad de la finalización de los cursos.

Martínez (2001) constata que desde principios del siglo pasado, las POM han enfrentado rechazos siempre y en todas partes, debido al desconocimiento de sus posibilidades y limitaciones por parte de sus críticos. A pesar de ello su uso ha venido en aumento al menos en Europa, Norteamérica y varios países de América Latina.

Tal vez por ser más difíciles de elaborar, se suelen plantear de manera inapropiada, dando lugar a serias críticas. Aunque las POM no puedan sustituir al profesor en su evaluación directa, ésta tampoco es suficiente para valorar objetivamente un número elevado de alumnos (Martínez, 2001). En cualquier caso, estas pruebas deberían ser complementadas con otros tipos de evaluación.

Estas evaluaciones sirven si sus resultados se aprovechan para el mejoramiento de los sistemas a evaluar. En México, por ejemplo, durante los 60s se desarrollaron POM que sometidas *a posteriori* a evaluaciones sicométricas, mostraron altos niveles de confiabilidad y de equivalencia entre distintos años (Martínez, 2001).

Conclusión

El presente trabajo muestra que las POM, gracias al alto número de preguntas que permiten plantear, contestar y corregir en tiempos razonables, son de gran utilidad y muy robustas a sus posibles deficiencias. Sin embargo es necesario que sean complementadas con otro tipo de pruebas de evaluación, tal como se hace en este curso y que sean continuamente analizadas en su formulación, desempeño y resultados.

Referencias bibliográficas

Alonso-Tapia, J.F.; Asensio, E.; Fernández, A.; Labrada y F.C. Moral (1993). Modelos y estrategias para la evaluación del conocimiento y su adquisición: un estudio piloto. *Tarbiya*, 3, 7-48.

Backhoff Escudero, E., Larrazolo Reyna, N. y M. Rosas Morales (2000). Nivel de dificultad y poder de discriminación del Examen de Habilidades y Conocimientos Básicos (EXHCOBA). *Revista Electrónica de Investigación Educativa*, 2, 1, 13.

Backhoff Escudero, E.; Tirado Segura F. y N. Larrazolo Reyna (2001). Ponderación diferencial de reactivos para mejorar la validez de una prueba de ingreso a la universidad. *Revista Electrónica de Investigación Educativa*, 3, 1, 21-33.

Berezina, M. y A. Berman (2000). Proof reading' and multiple choice tests. International Journal of Mathematical Education in Science and Technology 31, 4, 613-619.

Bordas, M.I. y F.A. Cabrera (2001). Estrategias de evaluación de los aprendizajes centradas en el proceso. *Revista Española de Pedagogía*, LIX, 218, 25-48.

Buchweitz, B. (1996). Elaboraçâo de questôes de múltipla escolha. *Estudos en avaliaçâo educacional*, 14, 105-132.

Bush, M. (2001). A Multiple Choice Test that Rewards Partial Knowledge. *Journal of Further and Higher Education* 25, 2, 157-163.

Campbell, N.A.; Mitchell, L.G. y J.B. Reece (2001). *Biología: Conceptos y relaciones*. México: Pearson Educación.

Careaga, A. y E. Rodríguez (2002). Evaluación: ¿Acuerdos o controversias? Un análisis desde la investigación educativa aplicada. *Educar* 10, 19-26.

Casanova, M.A. (2003). Evaluar ¿Para qué?, *Educar*, 13, 20-21.

Fierro, A. y C. Fierro-Hernández (2000). Formatos de examen y objetividad en las calificaciones académicas. *Revista de Educación*, 322, 291-304.

Godoy, A.S. (1995). Avaliaçâo da aprendizagem no ensino superior: estado da arte. *Didatica*, 30, 9-25.

Haladyna, T.M.; Haladyna, R. y C. Merino Soto (2002). Preparación de preguntas de opciones múltiples para medir el aprendizaje de los

estudiantes. *OEI-Revista Iberoamericana de Educación.* http://www.campus-oei.org/revista/evaluacion5.htm

Harris, D.C. (1995). *Quantitative Chemical Analysis.* New York: W. H. Freeman and Company.

López Frías, B.S. y E.M. Hinojosa Kleen (2001). *Evaluación del aprendizaje: alternativas y nuevos desarrollos*. México DF: Trillas.

Martínez Rizo, F. (2001). Evaluación educativa y pruebas estandarizadas. Elementos para enriquecer el debate, *Revista de Educación Superior*, 30, 4, 120, 71-85.

Méndez Vega, N. (2000). Algunas reflexiones conceptuales sobre la evaluación. *Revista Educación*, 24, 1, 53-60.

Rivas Gutiérrez, J. y J. Ruiz Ortega (2005). ¿Control, castigo o premio en el otorgamiento de las calificaciones? Revista Digital de Educación y Nuevas Tecnologías 29. http://contexto-educativo.com.ar/2003/5/nota-05.htm

Snedecor, G.W. y W.G. Cochran (1967). *Statistical Methods*. Iowa State Univ. Press.

Williams, R.L. y L. Clark (2004). College students' ratings of student effort, student ability and teacher input as correlates of student performance on multiple-choice exams, *Educational Research* 46, 3, 229-239.