DIAGNOSTICOS Y ANALISIS DE PRUEBAS DE EVALUACIÓN MEDIANTE USO DE EQUIPO Y PROGRAMAS DE COMPUTACION

Danilo García *
*Javier Gaínza***

RESUMEN

La disponibilidad de técnicas de computación ofrece posibilidades amplísimas de análisis de pruebas de evaluación, en razón del gran caudal de información que brindan y de su gran versatilidad al aplicarse a diferentes tipos de examen. La calificación computarizada ofrece no solamente la rapidez de la calificación individual del estudiante sino también una oportunidad de análisis del examen respecto al grupo examinado. La distribución de frecuencias, las estadísticas básicas, los índices de dificultad y de discriminación de cada pregunta y del examen en sí, son un buen indicador de la presencia o no de defectos en la elaboración de la prueba.

Se analiza la experiencia recogida con un grupo de 88 estudiantes de Bioquímica, durante el primer semestre de 1978. La respuesta de los profesores y de los estudiantes a la aplicación del método en cuestión fue altamente satisfactoria.

Este tipo de diagnóstico y análisis de prueba adquieren gran simplicidad utilizando dispositivos como lector óptico de marca sensible incorporada directamente a un computador y con programas de análisis estadísticos. *(Rev. Cost. Ciencia Méd. 1984;5 (Suplemento 1);*

El alto grado de desarrollo alcanzado por la metodología de la enseñanza en las últimas décadas ha hecho casi inevitable la aparición de técnicas cuantitativas para el estudio de pruebas administradas al estudiante. Tales técnicas son importantes por cuanto convierten la evaluación del estudiante en objeto de análisis riguroso y objetivo, apartándole, al menos potencialmente, de su posible dependencia de la idiosincrasia del profesor del curso. Ello indica que el proceso de enseñanza, en su etapa de medición de asimilación de conceptos, está destinado a perder cualquier resabio de empirismo, y a convertirse en una tarea tan exigente como cualquiera otra dentro del campo de la ciencia. (2)

El convencimiento de que el análisis de la tarea docente, y en especial de la evaluación de conocimientos es de vital importancia en una casa de enseñanza superior, ha llevado a los autores a comunicar en este reporte su limitada experiencia, con la esperanza de que pueda ser de utilidad a quienes se hayan planteado interrogantes o hayan encontrado problemas en el desarrollo de sus labores docentes en lo que se refiere al análisis de las pruebas o exámenes que se usan para la calificación de la labor del estudiante.

La disponibilidad de técnicas de computación ofrece posibilidades amplísimas de análisis de exámenes en razón del gran caudal de información que brindan de manera rápida y económica, y de su gran versatilidad por cuanto pueden aplicarse a diferentes tipos de examen. Es necesario señalar que lo que aquí se comenta tiene su aplicación a cursos teóricos y prácticos, que posean exámenes escritos como parte de su sistema de eva-

* Departamento de Bioquímica, Facultad de Medicina, Universidad de Costa Rica, San José, Costa Rica.
** Escuela de Ciencias de la Computación e Informática, Universidad de Costa Rica, San José, Costa Rica.

luación y que tengan un número relativamente elevado de estudiantes. Es obvio que disciplinas que incluyan actividades diferentes como instrumento de evaluación escapan a los alcances de lo que aquí se expone.

ANTECEDENTES

Uno de los autores, Dr. Jorge Danilo García, labora en el Departamento de Bioquímica de la Escuela de Medicina. Durante los años de 1974 a 1978 fue posible apreciar en toda su magnitud, la pesada tarea que la calificación de exámenes "clásicos", basados en preguntas de tipo ensayo, asociación o dibujo de fórmulas o modelos, impone al personal docente de la unidad, debido al número elevado de estudiantes, tanto de la carrera de Medicina como pertenecientes a cursos de servicio. La calificación "manual" de este tipo de exámenes consume gran cantidad de tiempo, especialmente cuando el número de estudiantes superan los 70 ó 100 por grupo. Además, debe sumarse a ésto el tiempo dedicado a atender solicitudes de revisión o apelaciones por parte de aquellos una vez que se hace entrega del examen calificado, que en ocasiones puede ser considerable. La calificación de un número elevado de este tipo de exámenes generalmente trae como consecuencia el que transcurran varios días antes de que pueda procederse a la tabulación final. Esto puede ocasionar que, aunque el docente que califica haga esfuerzos por evitarlo, los criterios de calificación para cada pregunta puedan presentar variaciones a lo largo de los días que se tarde en efectuar la calificación. Ello puede disminuir la uniformidad de criterio que se aplica a la hora de la calificación, máxime son varios los docentes que intervienen con ella.

En razón de lo expuesto anteriormente, fue bien recibida la sugerencia de someter a prueba modificaciones en el tipo de exámenes y su evaluación por parte del Lic. Guillermo Calderón del NIDES (Núcleo de Investigaciones de la Enseñanza de la Salud)*. Las modificaciones a que se alude consisten en la elaboración de pruebas basadas exclusivamente en preguntas de selección múltiple y sus variantes, que se califican mediante el uso de tarjetas para computación, previamente adecuadas para ser perforadas manualmente, de manera tal que una leve presión ejercida en el espacio correspondiente basta para dejar una perforación en la tarjeta. De esta manera el estudiante usa el cuestionario de examen y perfora la opción escogida.

* Escuela de Medicina, Universidad de Costa Rica.

MATERIALES Y PROCEDIMIENTO

Durante el año 1975 se elaboró un programa para la máquina IBM/360 escrito en lenguaje FORTRAN para calificar y analizar estadísticamente los resultados de exámenes de tipo selección múltiple.

Se aplicó principalmente en exámenes de "ubicación" a los estudiantes que ingresaban al curso MA—0101 Matemática de Ingreso, para la orientación de los estudiantes de las carreras de Matemática y de Ciencias de la Computación. Se realizaron varias versiones de programa y durante varios años se utilizó para diagnosticar diferencias entre gru-

pos de cursos masivos y evaluación estadística básica que permitía a los directores y coordinadores de sección emitir criterios apropiados de los exámenes y pruebas realizadas. Durante el año 1978 se realizó un estudio sistemático en la cátedra de Bioquímica de la Facultad de Medicina y con una versión completa del programa que corrige y da estadísticas sobre una población de unos 90 estudiantes. Esta decisión se tomó después de haber probado el programa en el curso ME-1001 Estructura y Función Normal, dictado en el segundo ciclo lectivo de 1977, a 200 estudiantes.

El programa lee las respuestas correctas, califica a cada estudiante con nota de 0 a 10, independientemente del número de preguntas, y da las siguientes estadísticas evaluadoras de la prueba:

1.—Distribución de frecuencias, en absoluto y porcentajes relativos a la calificación de 0 a 10 de los estudiantes así como el gráfico de la misma.

2.—Número de estudiantes y porcentaje de nota inferiores a 5, entre 5 y 7, y superiores a 7.

3.—Indice de dificultad de la prueba, por porcentaje de aprobados, nota media general y desviación típica.

4.—Indice de dificultad de cada pregunta, a través de frecuencias y procentajes relativos al número de estudiantes que respondieron correctamente la pregunta.

5.—Indice de discriminación de cada pregunta. Para ello el programa compara cada pregunta de los estudiantes del 27 por ciento de la población que obtuvieron mejor nota en el examen y respondieron bien a la pregunta con los estudiantes de 27 por ciento inferior que también respondieron bien a esa pregunta. La diferencia de estas cantidades divididas por el número de estudiantes que componen el 27 por ciento nos da el índice de discriminación id, tal que —1 <id < 1.

Un índice de discriminación muy bajo, menor que o.3 determina que la pregunta fue o muy fácil (respondiendo casi todos bien), o muy difícil (no respondiendo casi nadie), o se hizo fuera del contexto de la prueba. Una simple revisión de las otras estadísticas determinarán la causa del índice tan bajo.

Después de cada corrida se localizaron y eliminaron del examen las preguntas por algún motivo defectuosas. Esto se hizo principalmente analizando el índice de discriminación.

En todos estos análisis se utilizaron pruebas de selección múltiple, una respuesta correcta de cuatro opciones. A fin de evitar la nota mínima obtenible simplemente al azar y con el afán de hacer pensar más al estudiante al momento de decidir por la respuesta correcta, se advirtió a los estudiantes de responder sólo a las preguntas sobre las que estuvieran seguros. Por cada cuatro respuestas malas se eliminó una buena en la nota general. Esto mejora el índice de discriminación de cada pregunta, al tratar de disminuir un factor de “ruido” El estudiante tiene opción de arriesgar hasta tres preguntas en todo el examen respondiendo sin mucha seguridad.

Mediante sistema de procesamiento “batch”, o en lote, se leyeron tarjetas con las respuestas de los estudiantes en la computadora IBM/30 ubicada anteriormente y en lenguaje FORTRAN, se obtuvieron los resultados. Las tarjetas leídas eran tarjetas especiales, que los mismos estudiantes perforaban en el momento del examen. Se utilizaron tarjetas pre-perforadas que mediante un impulso con la punta de un lápiz produce un hueco rectangular, similar al de una perforadora de tarjetas. Como procedimientos práctico el estudiante marcaba con lápiz la opción correcta y luego, al final del examen, procedía a realizar las perforaciones manualmente.

Se trabajo así durante todo el primer ciclo de 1978.

RESULTADOS Y DISCUSION

Esta experiencia de evaluación de exámenes y uso de pruebas de selección múltiple susceptibles de análisis mediante programas de computación, se efectuó en el departamento de Bioquímica en el curso BQ-0330 Bioquímica para estudiantes de Microbiología durante el primer ciclo de 1978. Este es un curso introductorio y de carácter general, con cierto énfasis en algunos tópicos que revisten interés para el futuro profesional. Como tal, tiene cierta orientación hacia la relación de aspectos teóricos de Bioquímica con la Química Clínica. Posee un total de 6 horas teóricas por semana y se dicta de manera colegiada. Durante el período a que se hace alusión, tuvo una matrícula inicial de 94 estudiantes y una final de 88 estudiantes. Las lecciones se dictaron de la manera habitual, utilizando fundamentalmente diapositivas como elementos de apoyo visual para facilitar al estudiante la comprensión de ilustraciones moleculares en tres dimensiones y otros conceptos claves.

La evaluación del desempeño del estudiante consistió en la administración de tres exámenes parciales y un examen final, constando ellos de preguntas de elección múltiple, de cuatro opciones cada una. Cada uno de los exámenes parciales constaba de 50 preguntas con un valor de 2 puntos cada una, dando así un total de 100 puntos por examen. El mismo criterio se siguió en la elaboración del examen final.

Los exámenes se elaboraron tratando de establecer equilibrio entre preguntas cuyo nivel de exigencia al estudiante fuera relativamente bajo, como lo es el caso de preguntas que involucran solamente memorización, y preguntas con niveles de exigencia más altos, que requieran del estudiante relación y análisis de conceptos previamente memorizados.

La elaboración de un examen de este tipo es una tarea difícil cuando el profesor trata de ajustarse estrictamente a las exigencias establecidas por los pedagogos para tal fin. Dicha tarea, empero, se compensa con creces en razón de los enormes beneficios que la calificación computarizada proporciona en lo que se refiere al análisis del examen y a la rapidez de obtención de la calificación. (1)

En el caso que nos ocupa, y como se explica en otra sección de este reporte, la computadora puede proporcionar información que permite verificar si, en realidad, la información que procesa como proveniente del estudiante es acertada, en otras palabras, protege al estudiante contra posibles errores en el mismo programa. Además, para cada pregunta da el porcentaje de estudiantes que contestaron cada una de las opciones, lo cual permite al profesor, como lo hizo en el curso que nos ocupa, formarse una idea preliminar acerca de cada pregunta, tomando como regla empírica el que una pregunta que conteste correctamente más del 80 por ciento de los estudiantes, o debe ser sospechosa por falta de discriminación o responde a conceptos en los cuales se insistió exhaustivamente y por lo tanto debe aceptarse sin reservas. Por el contrario, si una pregunta fue contestada de manera incorrecta por más del 80 por ciento del grupo, se puede pensar en la eliminación de esa pregunta del examen. En el caso de exámenes calificados según el método tradicional, eliminar preguntas puede representar la inversión de un número apreciable de horas hombre dedicadas a cálculos aritméticos que podrían ser utilizadas de mejor manera. Utilizando programas de computación, tal inversión desaparece totalmente. La gráfica de distribución de frecuencias se obtuvo asimismo para cada examen. Aparece ilustrada la correspondiente al examen final del curso. Como se podrá apreciar, la ausencia de tendencias en la curva es un buen indicador de ausencia de defectos en la elaboración del examen.

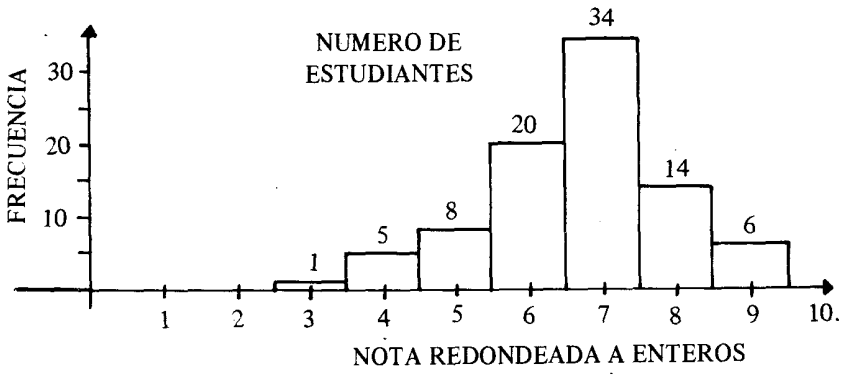

FIG. 1: EXAMEN FINAL BQ-0330 BIOQUIMICA, PRIMER CICLO 1978

También para cada pregunta se obtuvo un índice de discriminación, según se ha definido anteriormente, lo cual cuantifica objetivamente la capacidad de discriminación de cada pregunta. Habiéndose establecido el índice de discriminación aceptable, puede procederse a la eliminación de preguntas y recalificación de examen que, según nuestra experiencia, favoreció a los estudiantes con mejores calificaciones y mejoró la simetría de la curva de distribución de frecuencias. Por último, y aunque en nuestro caso no se utilizó, la capacidad del programa para dar la posición de cada estudiante con respecto a la media, en términos de desviaciones estándar, podría haber sido de gran utilidad a los estudiantes a lo largo del curso, para conocer su situación real con respecto al resto del grupo. Los exámenes se administraron los días sábados por la tarde con la finalidad de lograr un máximo de concentración por parte de los estudiantes en el examen, sin que otras lecciones o actividades del día se constituyeran en factor de presión que pudiera afectar su desempeño, y el tiempo permitido fue de tres horas, suficiente en todos los casos. Las tarjetas perforadas por los estudiantes y los cuestionarios fueron recogidos y las tarjetas sometidas a proceso a través de la lectura de tarjetas de la máquina IBM/360. El tiempo requerido para la lectura y calificación de examen, junto con el requerido para proporcionar toda la información estadística a que se ha hecho mención pudo ser en aquel entonces varios minutos. Se procedió luego al estudio de las calificaciones obtenidas por los estudiantes, y al análisis de la información accesoria proporcionada al profesor por medio de un programa de computación. Tomando en cuenta los diversos factores, se procedió entonces a determinar cuáles preguntas no reunían los requisitos mínimos establecidos a fin de proceder a su eliminación y ulterior recalificación del examen. Por razón de tipo de pregunta escogida, cada estudiante tienen una posibilidad real de contestar un 25 por ciento de su examen dando respuestas aleatorias. La tendencia de estudiante de responder al azar, cuando se enfrenta a una pregunta que va más allá de sus capacidades, se trató de compensar de manera tal que el estudiante sufriera penalización de un punto por cada cuatro respuestas erróneas. De este manera se trató de estimular al estudiante a no contestar aquellas preguntas acerca de cuya respuesta no estuviera razonablemente seguro.

En concreto, la respuesta de los profesores y estudiantes a la aplicación del método en cuestión fue altamente satisfactoria. Los profesores obtuvieron las calificaciones de los exámenes en un tiempo mínimo, sin tener que invertir esfuerzo y tiempo alguno en cali-

ficación *per se,* abocándose únicamente al análisis de las preguntas, teniendo como meta la ulterior la recalificación del examen, si esto se hubiese manifestado cuino necesario. A su vez, lo anterior no se traduce en un simple ahorro de tiempo y esfuerzo por parte del docente. Tal y como se pudo apreciar, la preparación de este tipo de examen requiere de una gran inversión de tiempo y esfuerzo mental, ya que, en más de una ocasión, lo que parecía ser una excelente pregunta desde el punto de vista del docente hubo de ser rechazada ulteriormente, en razón del análisis estadístico que la catalogó como no representativa o fuera de contexto. Esto da pie a la proposición ya comentada en círculos docentes de la Facultad de Medicina, de establecer bancos de preguntas tipificadas que pudieran usarse de manera continua a la vez que se afina el grado de dificultad de las pruebas, y se adquiere mejor adecuación entre el tipo de estudiante y la prueba que se le administra, mediante pruebas de ubicación que se podrían efectuar al principio del ciclo lectivo. Por su parte, los estudiantes mostraron sorpresa ante este tipo de calificación, que prontamente se transformó en aceptación crítica del método. No se pudo menos que reconocer la rapidez y absoluta objetividad del programa, centrándose la interacción profesor-alumno en lo que se refiere a si una pregunta estuvo bien o mal planteada, en lugar del consabido tira y afloja en búsqueda del punto de más que algunas veces caracteriza la reacción del estudiante cuando recibe un examen de desarrollo corregido.

Es preciso notar que este tipo de diagnóstico y análisis de pruebas adquieren una gran simplicidad si podemos contar con un mecanismo del tipo de lector óptico de marca sensible incorporada directamente al computador y con programas del tipo descrito en este artículo. Creemos que la adquisición de dispositivos de este tipo no solamente son necesarios para actividades docente-administrativas como son Registro, Admisión, Becas, y otros, en una institución de Educación Superior, sino también para las tareas docentes, primordialmente en cursos muy populosos y en pruebas de diagnóstico y aptitud, y en labores de investigación especialmente en investigación social, como un recurso confiable, y de gran velocidad para alimentar la entrada de información.

ABSTRACT

Computing techniques offer great posibilities for exam analysis, due to the amount of information provided and the versatility available when facing different kinds of exams.

Computarized scoring provides not only rapid processing of individual scores, but the opportunity to analyze each score relative to group performance. Frequency distribution, basic statistics, indexes of dificulty and discrimination for each question and for the test as a whole are good indicators of the presence or absence of defects in the test.

Data collected during the first semester of 1978, pertaining to a group of 88 biochemistry students, are analyzed. The feedback from both students and professors, relative to the application of the method in question, was highly satisfactory.

This type of test diagnosis and analysis becomes greatly simplified with the introduction of facilities such as an optical scanner directly incorporated in the computer, and with the use of programs for statistical analysis.

BIBLIOGRAFIA

1.— Fodner, G.M, “Statistical Analysis of Multiple Choice Exams” *J. Chem. Ed.* 1980; 57(3): 188-190

2.— Guilbert, J.J. *“Guía Pedagógica”* Ginebra, OMS, 1976; 480-483

Impreso en la Unidad de Producción
de Material Educativo — PASCCAP
OPS/OMS — Costa Rica 1985