# Resultados del Trabajo en el CCH sobre el Sistema que Califica y Analiza Exámenes de Opción Múltiple

## INTRODUCCION

Este trabajo pretende dar una interpretación de los alcances y resultados que proporciona el sistema que califica y analiza exámenes de opción múltiple. Debido a que una computadora tiene la posibilidad de realizar millones de comparaciones, asignaciones, operaciones aritméticas, algebraicas y de trabajar con grandes volúmenes de información en tiempos cortos, el programa permite diagnosticar y calificar una pregunta, no sólo por dos o tres parámetros (dificultad, equilibrio temático, etc.) sino por aquellos que desde un punto de vista teórico tienen importancia. Así una pregunta puede ser analizada por el proceso psicológico que se realiza para contestarla, su peso relativo en un objetivo, el peso que tiene éste en el curso, el tipo de evaluación que se va a realizar, la dificultad y el sesgo de cada opción, etc.

La metodología aquí expuesta permite calificar exámenes más acordes con las necesidades del curso y proporcionar a su vez instrumentos más cercanos a los principios de una evaluación por criterios, a diferencia de una evaluación por normas, que es la que predomina en la práctica, por las restricciones que tiene el método manual o la interpretación "manual" de los resultados de la computadora.

Una ventaja adicional del método expuesto es la posibilidad de abordar el problema de la validez del instrumento, contando con un marco de referencia hacia el programa del curso y el uso de múltiples criterios en la selección de reactivos que constituyen la estructura del examen. A lo largo del presente trabajo se expondrá cómo ello es posible. Este se ha dividido en tres partes:

*a)* El marco teórico de referencia, donde se da un panorama somero de lo que justifica y se entiende por evaluación escolar.
*b)* La descripción de lo que se entiende por un examen de opción múltiple.
*c)* La descripción del procedimiento de calificación del mismo.

Por último, se han eliminado del texto, para hacerlo más accesible, las fórmulas estadísticas y tecnicismos de programación.

## Marco Teórico

Es notorio en un grupo pequeño (máximo 8 alumnos) observar cómo un profesor no tiene dificultad para practicar la evaluación escolar. Intuitivamente maneja la información necesaria y adecuada para que dentro del proceso de enseñanza-aprendizaje se valore, corrija y ayude a cada alumno miembro del grupo. Cuando el grupo crece (en grupos de 20 a 25 alumnos ya se hace notable) el profesor pierde esa capacidad: la información necesaria y adecuada es demasiado extensa y compleja para que la retenga y utilice.

En el primer caso un profesor trabaja con individuos, en el segundo con un grupo. Cada alumno ha perdido su individualidad y así la capacidad de establecer una comunicación personal con el maestro, de tal forma que a éste no le es posible establecer criterios de evaluación y de enseñanza-aprendizaje en forma individualizada.

En grupos aún mayores (de 30 o más alumnos) el profesor recurre a conferencias y recomendaciones generales y utiliza cada vez menos cualquier tipo de evaluación individualizada, debido a la gran cantidad de tiempo y trabajo que le llevaría hacerlo. En estas condiciones utiliza como principal evaluación sistemática el "examen final", una tradición que tanto a él como al jefe de departamento correspondiente les da información en aciertos o categorías como "NA", "B", "S" y "MB", las que resultan necesarias para poder asignar el número de créditos que cada alumno tiene que cubrir en el currículum de un ciclo escolar.

A un director o un funcionario en la administración escolar, la evaluación le interesa como "Acreditación", donde los factores socioeconómicos y demográficos son relevantes y las necesidades de información versan sobre cuestiones tales como: ¿Cuántos alumnos terminan el ciclo escolar? ¿Cuál es la demanda de ingreso? ¿Qué población permiten las instalaciones?, y las proyecciones de éstas y otras cuestiones similares en los años siguientes. Para ello se asignan grandes recursos para realizar exámenes de selección y acreditación, cuyas distribuciones de aciertos son profusamente utilizadas. En este contexto, un funcionario ya no maneja información sobre grupos y los estudiantes como individuos no existen; son una población cuyos datos académicos son abstracciones entre las demás abstracciones socioeconómicas.

Como puede observarse, en la medida que la población crece, la evaluación se diferencia cada vez más del proceso de valoración que intuitivamente tiene un profesor ante un grupo pequeño. Hay que señalar que, al evaluar por medio de un "examen final", se da preponderancia a la información que tiene un alumno sobre la materia y no al conocimiento que de ella tenga, ya que analizando la relación entre información y conocimiento resulta que no hay conocimiento sin información, pero puede darse el caso de haber información sin conocimiento. En términos estrictamente pedagógicos, los aciertos o las categorías como "B", que produce un "examen final", son sólo una apreciación vaga y sin mucha ayuda práctica para un profesor, ya que a éste le es más útil saber, por ejemplo, si sus

alumnos pueden manejar logaritmos, que saber si el promedio es "S". En resumen, y en esta primera aproximación, se podrá decir que en teoría la evaluación escolar estudia la técnica y los instrumentos que permiten medir y describir el proceso de enseñanza-aprendizaje dentro de los cuatro niveles de información descritos, a saber: el alumno como individuo, el grupo, los grupos y la población, así como las relaciones entre ellos dentro de un programa dado. En la práctica, en la medida en que la población crece, disminuye y se distorsiona la información académica.

Un segundo enfoque sobre la evaluación escolar lo podemos llamar: "la aproximación epistemológica" y en términos generales es el siguiente.

El conocimiento humano puede clasificarse en dos grandes conjuntos: el conjunto del conocimiento no verbal o privado y conjunto del conocimiento público. Al primero pertenece, por ejemplo, el gustar cierto tipo de música; en cambio el conjunto de conocimiento público está constituido por proposiciones comunicables y verificables. Cabe señalar que dichas proposiciones pueden ser falsas o verdaderas, pueden ser vagas o precisas, etc. Por ejemplo, el conocimiento científico es un subconjunto de todas las proposiciones públicas que son comunicables y verificables y el método científico es el procedimiento mediante el cual puede decirse cuándo una proposición resulta verdadera en un determinado experimento, y así se sabe si pertenece o no al subconjunto de proposiciones que constituyen el conocimiento científico.

Con este enfoque, el conocimiento escolar que es posible de evaluar, se define como un subconjunto del conocimiento público; parte de él, a su vez, es un subconjunto del conocimiento científico. Las técnicas o instrumentos de evaluación generalmente hacen hincapié en la intersección entre el subconjunto del conocimiento científico y el subconjunto del conocimiento escolar. En otras palabras, las técnicas actuales se limitan a aquellas proposiciones que, tanto por su influencia histórica como por su veracidad, componen el conjunto de conocimientos que delimitan las áreas lingüísticas de los programas escolares.

Existen dentro del conjunto general del conocimiento escolar, aspectos no verbales. Un ejemplo evidente es la mímica, o el "Rapport" en una entrevista psiquiátrica. Sin embargo, hay que reconocer que las técnicas de evaluación actuales no miden o describen estos aspectos en grupos grandes, y es un campo que aún está por desarrollarse. El que no se evalúen, no les resta importancia, pero exceden los límites de las técnicas aquí descritas.

La evaluación escolar en su aproximación epistemológica puede definirse como el conjunto de técnicas que permiten medir y describir el conocimiento escolar evaluable en una disciplina que incluya entre sus objetivos de aprendizaje la información contenida en los seis conjuntos siguientes:

*a) Términos técnicos.* Es el conjunto de palabras con que una disciplina determinada denota los objetos de estudio, por ejemplo, en Química las palabras "Valencia", "Molécula", "Sodio", etc. son términos técnicos.

La evaluación de este conjunto consistiría en indagar si un alumno los puede reconocer adecuadamente, los puede agrupar en subconjuntos, puede discriminar la pertenencia o no de un término a un subconjunto, etc.

*b) Conceptos.* Es el conjunto de definiciones o propiedades que una determinada disciplina atribuye a los términos técnicos. Así en Química las propiedades de metal, conductividad, etc. que tiene la plata, o la definición de molécula como la unión relativamente permanente de al menos dos átomos, son elementos del conjunto CONCEPTOS.

La evaluación de este conjunto consistiría en indagar si un alumno puede establecer relaciones entre las propiedades de un concepto, puede asociar el término técnico apropiado a una definición, puede discriminar propiedades que tiene o no un término, etc.

*c) Leyes.* Es el conjunto de relaciones que una disciplina establece entre los términos técnicos para describir el comportamiento de uno o varios fenómenos. La ley de Newton, según la cual fuerza es igual a la masa por la aceleración, es un buen ejemplo. Algunos aspectos de este conjunto pueden evaluarse, cuando se indaga sobre la identificación, aplicación, ejemplificación, de una ley o un fenómeno que se aplique: sin embargo, la comprensión y generalización de una ley no pueden ser evaluados.

*d) Teorías.* Es el conjunto de explicaciones que permiten describir el comportamiento de un conjunto de leyes dentro de un todo. A nuestro juicio con exámenes de opción múltiple no es posible evaluar teorías; algunos aspectos de ellas serían evaluables con la técnica de diferencias semánticas.

*e) Métodos.* Es el conjunto de procedimientos o algoritmos que un área de estudio recomienda para resolver un problema. Con excepción de algunos métodos matemáticos o científicos bien definidos, cuya solución puede verificarse y así suponer que el método empleado fue correcto, con preguntas de opción múltiple es imposible evaluar el proceso en su conjunto.

*f) Hábitos de estudio.* Será el conjunto de conductas realizadas por un estudiante en el proceso de enseñanza-aprendizaje. La técnica de opción múltiple sólo es aplicable como escalas de actitudes, las que infieren la conducta, pero no la evalúan directamente.

Tanto el enfoque poblacional como el epistemológico están relacionados. El primero especifica a quiénes y en qué contexto se lleva a cabo la evaluación, y así define las diferentes evaluaciones como son: de diagnóstico, de acreditación, de selección, etc. El segundo enfoque delimita los contenidos que pueden ser evaluados y determina las técnicas necesarias para lograrlo.

La evaluación propiamente dicha es un procedimiento o algoritmo que permite establecer las relaciones entre ambos dominios. Llamaríamos al esquema completo MARCO TEORICO.

A continuación se utiliza el marco teórico para definir la técnica de exámenes de opción múltiple que son ampliamente utilizados, lo que permitirá además señalar algunos aspectos en que se ocupa el campo de la evaluación.

## Exámenes de opción múltiple

Entenderemos por examen de opción múltiple un instrumento de evaluación que nos permite medir y describir la amplitud y precisión que un alumno, un grupo, los grupos o una población tienen en el manejo lingüístico de los términos técnicos, conceptos, y algunas propiedades del conjunto "leyes" que utiliza una determinada área del conocimiento escolar en los parámetros de un curso escolar. Un instrumento así está constituido por distintos tipos de reac-

tivos que pueden clasificarse en la forma siguiente:

*1) Preguntas de reconocimiento.* Son preguntas que discriminan de un conjunto de términos técnicos las propiedades o la definición, que un alumno supone que es una correspondencia correcta, utilizando para ello básicamente la función psíquica de reconocimiento. Este tipo de preguntas son útiles para evaluar si un alumno, un grupo o un conjunto de grupos reconocen el significado de un término. Su estructura usualmente es del tipo:

PREGUNTA. Una partícula sub-atómica con masa 1 y sin carga es:

*a)* Electrón
*b)* Protón
*c)* Neutrón

*2) Antónimos.* Son preguntas que permiten discriminar las relaciones semánticas, es decir, distintos modos de nombrar un mismo término técnico o su contrario. También son prácticas en traducciones de idiomas.

Como en la primera pregunta, el alumno utiliza la función de reconocimiento, permitiendo un conjunto de este tipo de preguntas conocer la "riqueza verbal" que un grupo o grupos puedan tener en un área específica de conocimiento. Son preguntas del tipo:

PREGUNTA. Un sinónimo de caballo es:

*a)* Cuadrúpedo
*b)* Equino
*c)* Mamífero

*3) Analogías.* Son preguntas que permiten discriminar las relaciones funcionales o abstractas que tienen entre sí los términos técnicos o propiedades de ellos. Este tipo de preguntas requiere para ser respondidas el reconocimiento y estilos cognoscitivos. Un ejemplo de este tipo de pregunta es:

PREGUNTA. Electrón es a átomo como:

*a)* Lípido a Hígado
*b)* Insulina a Páncreas
*c)* Adenina a DNA

Con una adecuada redacción, la pregunta puede utilizar diferentes proposiciones (por qué, cómo, cuándo, etc.) y adecuarse así a investigar más directamente causalidad, inducción, deducción, etc.

*4) Correlación.* Cuando existe un término técnico que denota la relación común que tienen dos conjuntos de términos técnicos, las preguntas de correlación resultan ser de gran utilidad. Un ejemplo sería el término "Capital" que relaciona un conjunto de ciudades con el conjunto de estados; otro es "Etimología" que relaciona palabras con raíces. La función que el alumno utiliza para responder este tipo de preguntas es la de reconocimiento. Son útiles para evaluar el manejo del lenguaje o de una terminología que tenga que conocerse de memoria. Un ejemplo sería:

PREGUNTA. Relacione las palabras en inglés con sus correspondientes en español.

| | | |
|---|---|---|
| *a)* cat | 1) pato | ( ) |
| *b)* dog | 2) pollo | ( ) |
| *c)* horse | 3) perro | ( ) |
| *d)* chicken | 4) caballo | ( ) |
| | 5) rana | ( ) |
| | 6) gato | ( ) |

Es importante poner en la columna a correlacionar, más opciones, ya que si ambas columnas tienen igual número de opciones, las últimas alternativas están prácticamente determinadas. Este tipo de preguntas varían sus alternativas, según se van respondiendo; en nuestro ejemplo es 1/6 para la primera, 1/3 para la segunda, y así 1/4, y 1/2, para la última.

*5) Análisis y síntesis.* Una estructura de preguntas útiles para evaluar comprensión de lectura y análisis de párrafos se conoce genéricamente como "Análisis y Síntesis", ya que supone que el alumno utiliza tales funciones psicológicas para encontrar la respuesta. La técnica para construir este tipo de preguntas consiste en seleccionar un párrafo pequeño que encabeza a una o varias preguntas de reconocimiento del tipo:

PREGUNTAS. La idea principal del párrafo antes mencionado es:

*a)* La tercera ley de Newton
*b)* La ley de la gravitación universal
*c)* etc.

o bien:

PREGUNTA. Usted podría concluir que el autor del párrafo es:

*a)* Marxista
*b)* Idealista
*c)* Althusseriano

*6) Concatenación.* Cuando un grupo de preguntas están lógicamente relacionadas entre sí, ya que se refieren al análisis de un mismo párrafo o tema donde se preguntan sobre los distinto pasos de un razonamiento, están ligadas al reconocimiento. Una pregunta de ejemplo puede ser:

PREGUNTA 1. A un enfermo que supone tiene apendicitis, usted lo manda preferentemente a hacerse un exámen de:

*a)* sangre
*b)* coproparasitoscopio
*c)* orina

PREGUNTA 2. Si el análisis resulta positivo, usted le receta antes de operar:

*a)* morfina
*b)* penicilina
*c)* sulfadiacina

PREGUNTA 3. En qué dosis:

*a)* 2 tabletas al día durante 17 días
*b)* 3 tabletas al día por 3 semanas
*c)* 4 tabletas al día por 4 días

Como puede observarse en el ejemplo, hay una relación lógica entre el diagnóstico, el medicamento y la dosis; si el alumno se equivoca en la primera o segunda pregunta, las siguientes

aunque estén bien contestadas resultan falsas, ya que con un medicamento inadecuado, la dosis no tiene significado: es una pregunta en tres partes.

7) *Solución de problemas.* Este tipo de preguntas son adecuadas para evaluar algunos aspectos de cálculo matemático, ya que el alumno requiere hacer las operaciones antes de encontrar la respuesta. Su estructura es del tipo:

PREGUNTA. Encuentre la solución a la ecuación:
$3X + 2 = 0$

*a)* 3/2
*b)* - 2/3
*c)* 2/3

*8) Completamiento de oraciones.* Así como las preguntas de "solución de problemas" son una adaptación específica para evaluar aspectos del cálculo matemático, las preguntas de completamiento de oraciones, son una técnica para evaluar redacción y ortografía. Son preguntas donde existe una oración a la cual se ha quitado la palabra por evaluar, colocándolas en las opciones. Un ejemplo de este tipo es:

PREGUNTA. Mucha gente ve el futbol por:

*a)* Telebición
*b)* Televisión
*c)* Televición

Existen combinaciones de las anteriores preguntas: así las de "solución de problemas" pueden ser "concatenadas", preguntando las soluciones parciales de la totalidad del problema; "Análisis y Síntesis" con "completamiento de oraciones", etc. Creemos que las preguntas antes discutidas constituyen los reactivos básicos para evaluar preferentemente los aspectos lingüísticos, tanto semánticos como sintácticos entre los términos técnicos y los conceptos que maneja un área de conocimiento. Con excepción de las preguntas de solución de problemas y analogías, todas las demás utilizan la función del reconocimiento, esto es, una función en la que intervienen principalmente factores de identificación, memoria y asociación.

Si relacionamos las funciones de reconocimiento, pensamiento abstracto y cálculo matemático con el enfoque epistemológico, podemos decir que: el conjunto de términos técnicos, de conceptos, pueden evaluarse en aquellos aspectos que se refieren a su identificación y discriminación y ciertas propiedades del conjunto leyes, entre las que destacan su aplicación mecánica o algorítmica. No evalúa su interpretación, comprensión, y la toma de decisiones en problemas medianamente complejos. Los demás conjuntos (teorías, métodos, hábitos) no pueden ser evaluados eficientemente con la técnica de opción múltiple.

En el contexto del Modelo Teórico resalta que el instrumento de opción múltiple podrá evaluar si un alumno estudió un texto escolar y reconocer los términos técnicos y las relaciones adecuadas que éstos guarden con los conceptos y las leyes que ahí se enuncien. Por ejemplo, puede evaluar relaciones como si los elefantes pertenecen a la familia de los mamíferos, puede discriminar sus propiedades: reconocer a qué constelaciones pertenecen las 10 estrellas más brillantes del firmamento, hacer cálculos y la aplicación correcta de una fórmula; resolver problemas que,

por lo general, resultan hipotéticos y muy cortos, como los del tipo: "Si un tren viaja a 60 Km./h. . . etc". Por otra parte, es importante señalar aquello que el instrumento NO evalúa, como toma de decisiones, estrategias, planteamientos teóricos, demostraciones, experimentación, solución de problemas complejos, aspectos todos que bien pueden considerarse como muy importantes en el campo del conocimiento escolar.

Hacer hincapié en las limitaciones que tiene el instrumento de opción múltiple, permite hacer una advertencia para su utilización generalizada e indiscriminada, ya que da la ilusión de evaluar conocimientos escolares que en la realidad no mide. Y esto nos lleva al problema de la validez, que bien puede considerarse uno de los problemas fundamentales en evaluación. Veamos en qué consiste.

En física, entre otras ciencias, existen tres tipos básicos de unidades de medida: las de peso en kilogramos, las de espacio en metros y las de tiempo en minutos, con sus divisiones en gramos, centímetros, segundos, etc., o equivalencias en yardas y quintales. A primera vista la pregunta: ¿cuántos LITROS mide una placa de madera de un metro por dos metros y por un centímetro?, resulta absurda, aunque tiene una respuesta bien definida (si un litro equivale a 1000 $cm^3$ y la placa mide 100 cm X 200 cm X 1 cm = 20000 $cm^3$. Esa cantidad equivale a 20 litros). Ahora bien, si en lugar de litros preguntamos ¿cuántos grados celsius mide? o ¿cuántos "aciertos" mide? nos preguntaremos primero si tales preguntas son válidas, ya que no contamos con una relación de equivalencia entre exámenes o temperatura con volumen.

El uso generalizado de la técnica en evaluación de exámenes de opción múltiple como único instrumento de medida nos plantea cuestiones semejantes: ¿cuántas unidades de aciertos corresponden a la comprensión de la "teoría de la Evolución"?, ¿cuántos "aciertos" mide el conocimiento escolar de la física?, o en general, ¿es válido medir el conocimiento escolar con la técnica de opción múltiple? Por desgracia, las investigaciones serias sobre este aspecto son muy difíciles de llevarse a cabo, tanto por la complejidad como por la susceptibilidad que despierta la investigación de este tipo. Nuestras observaciones informales sugieren la hipótesis de que el uso que se da actualmente al instrumento, hace que los datos sobre los que se toman decisiones no sean válidos, ya que la información que se obtiene, es tan pobre y defectuosa que difícilmente se podrá fundamentar la afirmación de que con un examen de 100 preguntas pueda decidirse si un estudiante que responde a 46 de ellas tiene conocimientos, y no aquel que responde 40. Sin embargo una discusión más detallada, sólo es posible después de considerar el procedimiento de calificación, que a continuación veremos.

### Calificación de exámenes de opción múltiple

El método más conocido y usado de calificación de exámenes de opción múltiple consiste en:

*a)* Saber cuál es la opción correcta de cada una de las preguntas del examen (Tener una "clave" de respuestas).
*b)* Hacer una correspondencia entre la opción de la "clave" y la opción correspondiente al examen contestado por el alumno.
*c)* Sumar los "aciertos", esto es, sumar las

coincidencias entre la opción "clave" y la opción "respuesta".

*d)* Graficar con un histograma las frecuencias de "puntaje" encontrado para cada uno de los alumnos.

*e)* Analizar el histograma y poner cuotas de calificación, es decir, "20 aciertos corresponde "S", "30 B", etc.

*f)* Para exámenes de "selección", se ordenan los puntajes de mayor a menor y se van admitiendo los alumnos con mayores "aciertos" hasta cubrir la cuota de capacidad de admisión.

Se conoce este método como "calificación por normas". Es un procedimiento que resulta rápido, confiable y sintético, pues con muy poca información (una gráfica, cuatro parámetros y una escala) se pueden responder preguntas como: ¿cuántos alumnos aprobaron?, ¿cuántos pueden ser admitidos?, etc., información útil para una administración escolar. Es común interpretar la escala de "aciertos" como una relación lineal de conocimiento que va de 1 a 10 o de 1 a 100, interpretando un "4" como sinónimo de "mala preparación".

Esta interpretación puede no ser necesariamente correcta, pues no toma en cuenta que:

*a)* Las preguntas de un examen no siempre son "buenas", esto es, algunas están mal redactadas, otras son triviales, etc.; sin embargo, pertenecen a la misma escala.

*b)* Los alumnos que tienen igual número de aciertos, no siempre tienen el mismo patrón de respuestas (haciendo que conocimiento e información sean distintos); por ejemplo: el alumno A puede responder las preguntas pares y el alumno B las preguntas nones si el examen es de 10 preguntas, ambos tienen 5 aciertos, pero sus patrones de respuesta no coinciden.

*c)* Los puntajes encontrados, en un gran número de exámenes, se acercan más a un valor aleatorio que a medidas donde se pueda decir que hay conocimiento. Además las distribuciones son tan cercanas a las obtenidas por puro azar que uno se pregunta más si hay conocimiento o se está "midiendo" cualquier otra variable.

A este respecto hay muchas investigaciones que muestran cómo los puntajes en un examen no varían, antes y después de tomar un curso por lo que, si bien el procedimiento es práctico, existen cada vez más evidencias críticas sobre su validez.

A continuación vamos a describir otros métodos de calificación que están orientados al estudio de la validez de la escala y al contenido pedagógico de los exámenes; aunque aparentemente resultan laboriosos, las posibilidades que brinda una computadora, los hace prácticos, para obtener así escalas con criterios más acordes a aquello que llamamos el conocimiento escolar. Conocemos estos métodos como: "calificación por criterios"

### Calificación por criterios

El método de calificación por criterios se diferencia de la calificación por normas en los siguientes aspectos:

1) Utiliza diferentes "claves" que llamaremos "claves criterio", y de ellas las más importantes son:

1.1. *Criterio de relevancia.* En un examen de opción múltiple no todas las preguntas tienen la misma importancia; así una pregunta que explore la identificación de "la ley de hidrostática de Pascal" tiene más relevancia que aquella que pregunte "el siglo en que vivió Arquímedes", considerando un examen de "física". Bajo este criterio de relevancia las preguntas pueden clasificarse como:

1.1.1. Preguntas forzosas. Aquellos reactivos que a juicio del profesor y en correspondencia con el programa del curso se consideren indispensables de contestarse correctamente.

1.1.2. Preguntas importanes. Son aquellas que a juicio del profesor y en correspondencia con el curso, exploran aspectos importantes, como por ejemplo las aplicaciones de una ley, los conocimientos previos para entender algo forzoso.

1.1.3. Preguntas contextuales. Son preguntas que exploran aspectos del contexto cultural, como datos históricos, uso de sinónimos, identificación de términos técnicos no fundamentales, etc.

1.2. *Criterio temático.* En el programa de un curso escolar no todos los objetivos o, si el programa carece de ellos, no todos los "temas" tienen la misma importancia. El tema de "introducción" en general será menos importante que el de "las leyes fundamentales de la mecánica". Así el profesor puede ponderar con un criterio similar al anterior (forzosos, importantes, etc.) y asociar las preguntas del examen que corresponden a un determinado objetivo o tema. Hay que señalar que una pregunta puede pertenecer a varios objetivos o temas y habrá tantas claves como temas de exploren.

1.3. *Criterio de calificación subjetiva.* En la calificación por normas, la opción correcta tiene un único valor (1) y las opciones falsas cero. Pero una forma diferente es asignar calificaciones ponderadas a cada una de las opciones. Veamos un ejemplo:

P. ¿QUIEN DESCUBRIO AMERICA?

*a)* Zapata (-1)
*b)* Colón (1)
*c)* Cortez (0)
*d)* Pinzón (0.5)

Un procedimiento práctico es escoger 9 valores de calificación (-1, -0.75, -0.50, -0.25, 0, 0.25, 0.50, 0.75, 1) y para cada valor construir una "clave" en la que estén asociadas las opciones de las preguntas que a criterio del profesor deban calificarse con el valor correspondiente.

1.4. *Otras claves criterio.* En medicina existen preguntas cuya respuesta incorrecta "mataría al paciente", son preguntas del tipo:

P. ¿ANTE UN PARO CARDIACO, QUE HACE?:

*a)* Respiración boca a boca
*b)* Masaje al corazón
*c)* Electrocardiograma

o en ingeniería, el escoger una respuesta que haga "caer" un puente; así pueden darse criterios como "respuesta peligrosa"

"respuesta fatal", etc., que en realidad son sólo ejemplos para señalar que pueden definirse una variedad de criterios que adapten el contenido del examen a diferentes conductas, sean éstas o no académicas, siempre y cuando, como en los ejemplos no sean criterios forzados, sino que en forma lógica se adecuen al contenido del examen.

1.5. *Seguridad en la respuesta.* En las anteriores claves criterio se trataron algunos juicios que el profesor puede hacer sobre las preguntas del examen. En esta clave se pondera el juicio del alumno bajo criterios del tipo: *a)* Completamente seguro, *b)* Con cierta duda, *c)* No conoce la respuesta, *e)* La pregunta no se entiende, etc., conociendo el alumno que, si contesta el criterio "a", y su respuesta es falsa, se le penaliza con un valor; si su respuesta es la correcta, se le añade un valor, y si responde con los otros criterios se califica la pregunta sin penalizarse.

2. *Procedimientos de calificación.* Existen dos procedimientos 2.1) El cuantitativo y 2.2) El intrínseco. Vamos a describirlos brevemente:

.1. *Procedimiento de calificación cuantitativo.* En la clave criterio "probabilidad subjetiva" y en la de "seguridad de respuesta" se plantearon algunos valores diferentes a 1 para calificar una pregunta. En otras claves criterio se puede, a su vez, asignar valores como: preguntas forzosas = 2, importantes = 1.5, contextuales = 1, también los "temas" importantes = 1.5, etc. La calificación del alumno, por lo tanto, no será una suma de unidades, como la calificación por normas, sino la suma algebraica de los valores que toman cada uno de los criterios definidos para ese examen.

Como puede verse, el concepto "aciertos" que utiliza la calificación por normas ya no es aplicable; los números son "criterios" en una escala que puede tomar valores negativos, cero o positivos y que, basándose en la definición de examen, puede interpretarse como una escala donde los números mayores representan un "mejor" uso del lenguaje técnico de la materia que se explora. En cuanto a las asignaciones del tipo "NA", "S", etc. el criterio poblacional sigue siendo válido.

¿Qué ventajas concretas tiene este procedimiento sobre el procedimiento tradicional de calificar por normas? A nuestro juicio la mayor ventaja es que permite abordar el problema de la validez de las calificaciones. Así un análisis de las correlaciones entre los criterios y la calificación global permite obtener una mejor escala; por ejemplo, es de esperarse que aquellos alumnos que tengan las mayores calificaciones, respondan bien las preguntas "forzosas" o los temas "importantes" y las correlaciones entre ellos deberán ser significativas y positivas. Si esto no sucede, puede decirse con razón que las calificaciones no son válidas Una segunda ventaja es que permite generar información más acorde con las necesidades pedagógicas del curso. Por ejemplo: el que un profesor sepa las distribuciones de respuesta de las preguntas "forzosas" o de cada "tema", etc. le da una información mayor sobre la preparación académica de su grupo, que la que obtiene con una distri-

bución de aciertos, donde están mezcladas preguntas triviales con importantes.

Por último hay que advertir que si bien este procedimiento de calificación tiene múltiples posibilidades y parece ser más objetivo, se encuentra en una etapa de experimentación y habrá que validar si realmente constituye una mejor medida del conocimiento escolar, como teóricamente está justificado.

2.2. *Proceso de calificación intrínseco.* Es conocido, aunque no siempre aceptado, que algunas preguntas de los exámenes de opción múltiple son defectuosas y están mal construidas, por lo que sería injusto utilizarlas para fines de calificación. Definimos el proceso de calificación intrínseca como:

*a)* Localizar las preguntas defectuosas en el examen.
*b)* Eliminarlas y calificar por criterios como si no estuviera en el examen.
*c)* Proporcionar las razones por las que fueron eliminadas.
*d)* Señalar aquellos criterios que, por haberse excluido muchas preguntas, se consideren defectuosamente evaluados.

El saber cuándo una pregunta debe ser eliminada, se basa en un modelo estadístico y computacional que más o menos es el siguiente:

Una pregunta de opción múltiple puede ser analizada por su coincidencia entre un modelo teórico (una opción correcta y varias opciones como distractores) y las frecuencias de respuesta a las diferentes opciones. Con este enfoque el grupo de estudiantes que responde el examen, puede ser dividido en tres grupos: *a)* Aquellos que responden la respuesta correcta, *b)* Aquellos que responden alguna de las opciones falsas o distractores y, *c)* Aquellos que no responden la pregunta.

*a)* La población que escoge la respuesta correcta, debe en teoría ser la que obtenga mejor puntaje de aciertos, por lo que se espera una correlación positiva entre escoger la opción y el puntaje. A esta relación se la conoce como índice de discriminación.

Una segunda característica es la dificultad, que se define como la proporción uno menos el número de estudiantes que escogen la opción correcta entre la población total. En teoría cuando la pregunta es muy difícil (con un índice de dificultad de .80 o más), se tiene que tener cuidado, ya que si a esa proporción (sólo 20% o menos de alumnos la responden) se le añade el error de escogerla por puro azar, funciona como una opción falsa, esto es un distractor, y no la opción correcta. Inversamente, cuando el índice indica que la pregunta es muy fácil (más del 80% la responde) también hay que revisarla, ya que no está seleccionando aquellos alumnos que "conocen" de los que no. A toda la población se le está regalando un punto prácticamente. Se recomienda que en un examen no exista más del 5% de estas preguntas.

*b)* La población que responde las opciones

falsas, se comporta al azar, ya que ninguno conoce la respuesta correcta. Si la pregunta es de cinco opciones (una correcta y cuatro falsas), se supone que las opciones tendrán una probabilidad teórica de 1/4 de ser escogidas, más-menos un intervalo de confianza, esto es una banda de error que dependiendo de la población determina los valores del error estándar. Así por ejemplo, si en un examen de 1,600 alumnos 600 responden la opción correcta, se espera que los 1,000 alumnos restantes se distribuyan con valores cercanos a 250 en cada opción más-menos un intervalo de confianza que en este caso sería de aproximadamente 35 alumnos así (IC = 215 a 285).

Esto significa que, si la frecuencia obtenida en una opción excede o es menor que los valores del intervalo, su balance no es el teóricamente esperado, haciendo que la pregunta, en lugar de 5 opciones, sea de 4 o 3.

Otro supuesto es que los alumnos que respondan las opciones falsas, deben obtener menores puntajes, por lo que la discriminación debe ser al revés de como se espera en la opción correcta, esto es, una correlación negativa, o al menos, ninguna correlación.

*c)* Por último, está el grupo que no contesta la pregunta. Se ha observado que, cuando una pregunta resulta difícil de entender, en general no se contesta, así que si un 7.74% o más del total de alumnos constituyen este grupo, se tiene que revisar la pregunta, ya que puede muy bien no ser clara en su redacción.

Como puede observarse, tanto las correlaciones como las desviaciones y las fórmulas para obtener los intervalos de confianza requieren un sinnúmero de cálculos, que en la práctica sin la ayuda de una computadora sería imposible realizar; además que ello requiere cierto manejo de estadística que un profesor no necesariamente está dispuesto a aprender, sobre todo en aquellas áreas ajenas al conocimiento matemático, por lo que se construyó un cuadro con el que interpretando los índices anteriores, pueda un profesor fácilmente saber qué preguntas tuvieron problemas y cuáles no. Para explicarlo, veamos primero las reglas de asignación.

2.2.1. *Reglas de asignación.* Como ya se dijo, una pregunta puede ser analizada bajo tres parámetros: dificultad, discriminación y balance, los cuales a su vez pueden calificarse con un signo (+), cuando están cargados a un extremo de "abundancia"; con un signo (−), cuando están cargados al extremo de "carencia" y con un signo (0), cuando están dentro del intervalo de confianza que el modelo estadístico señala. Ahora bien, a partir del análisis de todas las combinaciones, se puede construir un cuadro que ubica a los reactivos como "óptimos" (cuando cumplen todos los supuestos del modelo), "buenos" (cuando algunos parámetros tienen fluctuaciones no importantes), "mejorables" (cuando algún pará-

metro puede afectar, aunque no necesariamente, la pregunta) y "defectuosos" (cuando los parámetros señalan desviaciones extremas que afectan con seguridad la totalidad de la pregunta). Existe una última categoría, "NO SE CONTESTO", donde se ubican preguntas que más del 7.74% de la población de alumnos no las contesta.

2.2.2. *Uso del cuadro.* Aquellos profesores que quieran reducir las categorías y saber solamente cuáles preguntas son "buenas" (B), cuáles requieren "revisarse" (R), "corregirse" (C), el cuadro adjunto señala la relación entre estas categorías y los casilleros correspondientes.

Por reactivos "buenos" (B) entendemos aquellas preguntas que obtienen valores aceptados dentro de las bandas, en sus parámetros de dificultad, balance y discriminación.

Por reactivos que requieren "revisarse" (R) entendemos aquellas preguntas que, si bien tienen defectos en cuanto a sus parámetros, estos no son tan graves o relevantes como para corregir la pregunta. En general señalan preguntas en las que una opción no está funcionando correctamente: sin embargo cualquier profesor que redacte preguntas de opción múltiple sabe la frecuencia con que se presentan opciones "forzadas". Si al revisar la pregunta no se encuentran mejores opciones, la pregunta puede ser utilizada.

Por último por reactivos que requieren "corregirse" (C), entendemos aquellos que sí conviene cambiar, ya que sus opciones o la redacción misma de la pregunta tiene defectos relevantes en los parámetros que el modelo observa. Estos desequilibrios exagerados afectan el error probable admitido para la opción correcta; en otras palabras, si tenemos una pregunta de cinco opciones y el 50% de los alumnos responde la opción correcta, es probable que de ellos un 5 o un 10% máximo la escogieran por adivinación; pero si la pregunta requiere corregirse, nos indica que más de un 10% escogen la opción correcta por deducción o adivinación pero no por conocer con seguridad la respuesta, lo que afecta la calificación total del examen. Por ello se indica que hay que corregirla. Sólo estas preguntas se excluyen de la calificación intrínseca.

DR. ADRIAN CAÑEDO
Coordinación del CCH.