VNIVERSIDAD NACIONAL AVTÓNOMA DE MEXICO

PERFILES EDUCATIVOS

ISSN 0185-2698

Cañedo, Adrián (1992)

"LA EVALUACIÓN MATRICIAL"

en Perfiles Educativos, No. 57-58 pp. 70-79.

Centro de Estudios sobre la Universidad

iresie

Banco de Datos sobre Educación

PERFILES EDUCATIVOS
Julio • Diciembre 1992
Núm. 57-58, pp. 70-79

LA EVALUACIÓN MATRICIAL

Adrián CAÑEDO

Presenta un nuevo programa de computadora para elaborar y calificar exámenes de opción múltiple, llamado evaluación matricial. todo cuidado. El antecedente de este instrumento de evaluaciones un programa construido por el autor que ha sido probado y utilizado durante 15 años en diferentes facultades de la UNAM. Esta nueva versión tiene la ventaja de permitir elaborar los exámenes de opción múltiple con la riqueza de un examen de ensayo tradicional.

Con el incremento de la demanda estudiantil en la UNAM, algunos grupos llegaban a tener, hace 25 años, 200 o más estudiantes. En ese tiempo había un profesor que tenía un estilo muy personal de evaluar a sus alumnos: aplicaba un examen con todo el rigor que la situación requería, pero para calificar los250 exámenes se subía a una escalera y lanzaba las hojas de papel, aquellos alumnos cuyo examen caía en el primer escalón tenían 1, los que sus exámenes caían en el segundo escalón tenían 2 y así sucesivamente hasta el 10. Cuando se le decía que aquello era injusto, el respondía:

> Aquel estudiante que cree que su calificación fue injusta, que pida revisión, y entonces lo califico con todo cuidado. Si un estudiante se conforma, es que su calificación fue justa o hasta fue un regalo.[1]

Quizá el procedimiento del profesor, frente al trabajo de calificar 250 exámenes, fue tan justo como otros, pero el problema no estaba en la asignación de calificaciones. El verdadero problema estaba en que había renunciado al conocimiento del saber de sus alumnos y con ello había abandonado una parte esencial de su trabajo.

Durante el proceso de industrialización de la educación, ésta dejó de ser personal. El profesor, que carecía de herramientas, fue perdiendo eficiencia y se colmó de trabajos administrativos, y las consecuencias de ello las vemos en el paulatino deterioro que ha sufrido la calidad de la enseñanza. Hoy, la situación de la evaluación escolar no ha variado mucho, los exámenes objetivos de opción múltiple, que ocasionalmente se aplican, no han resuelto el problema ni lo resolverán. Además, son pocos los profesores que les tienen confianza, y no sin razón.[2]

En 1973 construí un programa de computadora lo más práctico y versátil posible como para que un profesor o, mejor dicho, un departamento, pudiera calificar los exámenes de opción múltiple con la computadora. El programa se usó primero en la Facultad de Medicina, y posteriormente en el CCH, para calificar el examen de admisión a la UNAM y otras facultades y escuelas.[3]

Una característica de dicho programa es que no sólo asigna calificaciones, sino que también hace un análisis de los *items.* Su uso en la División de Estudios de Posgrado de la Facultad de Psicología es una garantía de que el modelo corresponde a los supuestos teóricos del análisis de reactivos. Después de cerca de 20 años de utilizarse, hasta donde yo sé, sólo el 20 por ciento de las preguntas que componen dichos exámenes están bien formuladas. El problema no es que las preguntas no se elaboren con cuidado, pues el ritual de aquellos que se encargan de redactar los

“Bancos de preguntas” consiste en discutir hasta el último detalle cada una de las opciones. El problema radica en que la técnica misma es engañosa. Pongamos el siguiente ejemplo para construir una pregunta de cinco opciones como:

Cuando la aurícula se contrae, el movimiento es ...

Como tenemos cinco opciones y sólo dos movimientos -diástole y sístole-, debemos inventar otras opciones falsas fuera del contexto natural para obtener la pregunta completa. En general, estos inventos producen preguntas mal formuladas según los cánones teóricos del análisis de los items. Por desgracia estos casos son los más frecuentes y los que ocupan mayor tiempo, llevando a discusiones bizantinas sobre si tal opción inventada está en el contexto o no de la pregunta.

En 1987, al desarrollar una versión más sofisticada de un generador de preguntas de opción múltiple -la primera se construyó en 1975-, se me ocurrió lo que podemos ahora llamar evaluación matricial (EM).[4] Veamos que se entiende por ello.

Características de la evaluación matricial

Las preguntas que caracterizan a la evaluación matricial tienen la siguiente estructura. Un ejemplo:

Ordene las expresiones que dan solución a la ecuación: 3 (x-5) + (x/2) – 13 = 0

	[1]	[2]	[3]	[4]	[5]	[6]	[7]
6 (x-5)+2(x/2)-26=0	[...]	[...]	[...]	[...]	[...]	[...]	[...]
x=8.....	[...]	[...]	[...]	[...]	[...]	[...]	[...]
6x-30+x-26=0.....	[...]	[...]	[...]	[...]	[...]	[...]	[...]
3x-15+(x/2)=13.....	[...]	[...]	[...]	[...]	[...]	[...]	[...]
x=56/7.....	[...]	[...]	[...]	[...]	[...]	[...]	[...]
3x+(x/2)=13+15.....	[...]	[...]	[...]	[...]	[...]	[...]	[...]
7x=56.....	[...]	[...]	[...]	[...]	[...]	[...]	[...]
6x+x=26+30.....	[...]	[...]	[...]	[...]	[...]	[...]	[...]

Este tipo de preguntas se distinguen porque:

A) Son de opción múltiple
B) Se aplican como un examen de lápiz y papel
C) Su calificación es objetiva
D) Da una “imagen” del contenido

A) Pregunta de opción múltiple

Si bien la pregunta es claramente de opción múltiple, se diferencia de las así llamadas en que se responde en un cuadro de doble entrada; una matriz en lenguaje matemático, en lugar de una

respuesta entre cinco opciones, donde sólo una es verdadera.

La ventaja de esta forma matricial de respuesta múltiple es evidente. La principal está en el aumento de la información que se puede extraer de la pregunta, ya que no sólo es más de una opción la “verdadera”, sino que las relaciones entre las opciones también son una fuente de información. Una segunda ventaja es que desaparecen las opciones falsas inventadas, pues la estructura misma de la pregunta proporciona diferentes opciones “falsas” en forma natural, es decir, dentro del contexto, evitando las discusiones sobre el contexto de los distractores y ahorrando así mucho tiempo. Una tercera ventaja es que permite hacer preguntas más interesantes o con características impracticables en las preguntas de cinco opciones. La pregunta anterior sobre los pasos para encontrar la solución de una ecuación es un ejemplo. También podemos evaluar procesos, pues la solución de una ecuación es sólo el caso particular de un proceso. Igualmente ocurre con los sucesos históricos y sus consecuencias. Podemos asimismo evaluar teorías, conceptos, etc. Y la construcción puede ser más rápida y natural. La siguiente pregunta es ilustrativa.

El Surrealismo se describe como ...

	[++]	[+]	[ ]	[ ]	[-]	[--]
Una visión onírica de la realidad.....	[ ]	[ ]	[ ]	[ ]	[ ]	[ ]
Una manera de pintar.....	[ ]	[ ]	[ ]	[ ]	[ ]	[ ]
Una escuela de cine.....	[ ]	[ ]	[ ]	[ ]	[ ]	[ ]
Una expresión mágica del universo.....	[ ]	[ ]	[ ]	[ ]	[ ]	[ ]
Semejanza entre magia y realidad.....	[ ]	[ ]	[ ]	[ ]	[ ]	[ ]

B) Aplicación

Las preguntas se aplican en un examen de lápiz y papel, es decir, se pueden aplicar a grupos (teóricamente tan numerosos como se quiera), y el costo, la organización de la evaluación, etc., es equivalente a los exámenes de opción múltiple. Es más, como las preguntas resultan más interesantes, en comparación con las de cinco opciones, se espera que la fatiga disminuya y la atención se incremente, ya que no es lo mismo responder 100 preguntas que saltan de un tema a otro en forma casi aleatoria que 10 ó 15 preguntas coherentes.

C) La calificación es objetiva

Existe la creencia de que la calificación de un examen de opción múltiple es objetiva. Desafortunadamente, cuanto más rígida es, tanto más subjetiva, en algunos casos. Por ejemplo, en arte o en algunas disciplinas o métodos de reciente formación, donde existen criterios diferentes y en ocasiones opuestos. Se puede pensar en las teorías del aprendizaje o en los criterios para evaluar una obra de arte. Si en esas condiciones aplicamos un criterio rígido, nos orientamos, en la mayoría de los casos, a un solo criterio o a un solo método, sesgando así la evaluación. Esto es frecuente en la técnica de opción múltiple con cinco opciones, ya que siendo una opción verdadera y sólo una, el criterio es rígido. Para evitar esta dificultad se requieren muchas preguntas que, en general, no es posible hacer.

La evaluación matricial es una técnica novedosa que, si bien falta experimentar más y desarrollar, tiene la posibilidad de calificar objetivamente diferentes criterios y, con ello, ajustarse más a la realidad de aquello que se quiere evaluar. Un ejemplo: Existen en filosofía ciertas escuelas que

describen la realidad de una determinada forma. La pregunta:

La realidad se define como...

	[++]	[+]	[ ]	[-]	[-]
Una invención humana.....	[ ]	[ ]	[ ]	[ ]	[ ]
Lo que perciben los sentidos.....	[ ]	[ ]	[ ]	[ ]	[ ]
Lo que dice la ciencia.....	[ ]	[ ]	[ ]	[ ]	[ ]
Lo que es exterior.....	[ ]	[ ]	[ ]	[ ]	[ ]
Las partículas elementales.....	[ ]	[ ]	[ ]	[ ]	[ ]
Lo que Dios creó.....	[ ]	[ ]	[ ]	[ ]	[ ]
Lo que fluye.....	[ ]	[ ]	[ ]	[ ]	[ ]
El espacio y el tiempo.....	[ ]	[ ]	[ ]	[ ]	[ ]

Puede calificarse con diferentes criterios. Si el modelo es teísta, “Lo que Dios creo” con ++, “Lo que es exterior” con ++ o +, etc.; es decir, la respuesta a aquellas opciones que corresponden al modelo filosófico serán las correctas; si la filosofía es positivista, las opciones que le corresponden serán las correctas. Lo que se calificaría como erróneo sería la inconsistencia, las contradicciones.

La calificación de este tipo de preguntas es lógica, a diferencia de las preguntas de opción múltiple, donde un acierto es un punto que se añade al puntaje total, es decir, si el alumno responde ciertas opciones, entonces la pregunta puede calificarse, en caso contrario, vale cero. Un ejemplo mucho más sencillo:

¿Qué valor toma x en las ecuaciones siguientes?.

	[1]	[2]	[3]	[4]	[5]	[6]	[7]
6(x-5)+2(x/2)-26=0	[ ]	[ ]	[ ]	[ ]	[ ]	[ ]	[ ]
x-1=0	[ ]	[ ]	[ ]	[ ]	[ ]	[ ]	[ ]
6x-4=26	[ ]	[ ]	[ ]	[ ]	[ ]	[ ]	[ ]
3x-15+(x/2)=13	[ ]	[ ]	[ ]	[ ]	[ ]	[ ]	[ ]
x=112/7-3x	[ ]	[ ]	[ ]	[ ]	[ ]	[ ]	[ ]
3x+2(x/2)=15	[ ]	[ ]	[ ]	[ ]	[ ]	[ ]	[ ]

Si un estudiante no responde la ecuación x-1=10, no ser le toma en cuenta la pregunta global, ya que si esa trivialidad como x-1=0 no la conoce, no sabe resolver ecuaciones. A esta manera de calificación la llamamos lógica, ya que es del tipo “si tal condición, entonces...”

D) Produce una “imagen” del conocimiento

Una de las características más importantes de las preguntas de tipo matricial es que las opciones están encadenadas unas con otras. La primera pregunta que presentamos, donde se marcaban los pasos para resolver una ecuación, es un buen ejemplo, pues hace evidente cómo cada opción está relacionada con las otras.

Pero también podríamos concatenar las preguntas, así, un examen con unas cuantas preguntas puede explorar todo un tema. Por ejemplo, una pregunta sobre solución de ecuaciones de primer grado; dos o tres sobre los pasos en la solución de una ecuación, y dos preguntas más sobre las reglas de los paréntesis y los signos, es decir, cinco preguntas bastarían para darnos una "imagen" sobre el conocimiento de las ecuaciones de primer grado, lo que difícilmente se lograría con un examen de opción múltiple.

Si uno reflexiona sobre este ejemplo, notará que esta técnica matricial que concatena preguntas se asemeja a la evaluación con preguntas de ensayo. Pero con una gran diferencia, aquí las preguntas están bien definidas, es decir, cada opción, cada respuesta y los criterios de calificación pueden ser desmenuzados y analizados en detalle. Además, con la ventaja de que pueden medirse y así obtener índices de dificultad y confiabilidad. Puede calificarse mediante computadora, permitiendo así hacerlo en grupos numerosos. En cambio, las preguntas de ensayo, si bien pueden formar un cuestionario también coherente, su definición, su lectura y calificación son difíciles, por no decir inaplicables, en poblaciones numerosas. Aún más, como los experimentos lo han demostrado, la confiabilidad es muy baja, y esto se debe, entre otras cosas, al criterio subjetivo de calificación.

Algunos profesores piensan que el criterio subjetivo de calificación es despreciable, que si el profesor conoce su materia, su criterio necesariamente es válido y correcto. En un trabajo realizado por la Sra. Deescoudres, del Instituto J. J. Rousseau, donde cita a E. Claparede, se eligieron 24 jueces para calificar con + ó – la respuesta que dio un niño a la pregunta ¿Qué diferencia hay entre una mosca y una mariposa? La respuesta fue "es más grande"; 12 jueces la calificaron con + y los otros 12 con -. En ese mismo estudio, un juez calificó a una misma prueba con 154 preguntas semejantes a la anterior –el *test* de Binnet-Simon- con un puntaje de 106, mientras que otro, lemas severo, dio un puntaje de 54. Como bien señala Claparede, "... la moraleja de todo ello es que es preciso tanto como sea posible elaborar *tests* que no dejen más que el menor papel posible a la interpretación subjetiva."[5]

Estos datos no sólo muestran que la evaluación escolar con métodos subjetivos es poco confiable; señalan, además un punto que usualmente se olvida: una evaluación defectuosa conlleva un alto costo social y económico. En la educación industrializada, los altos índices de deserción escolar –a pesar de los exámenes de selección-, la baja calidad de la enseñanza y el hecho de que un profesor no pueda explicar algo porque sus alumnos desconocen los fundamentos, son ejemplos de este alto costo. Por otro lado, un punto importante que por desgracia parece que no se ha considerado es el hecho de que una evaluación adecuada representa el único camino para que los experimentos sobre nuevos métodos de enseñanza, y en general cualquier investigación pedagógica, deje de ser una moda y se convierta en un verdadero método científico. Es en este sentido que la evaluación matricial adquiere valor, pues es un método de evaluación objetivo que tiene alguna semejanza con el contexto de las pruebas de ensayo que los profesores aprecian por muchas razones.

La estructura de una pregunta matricial

Una pregunta matricial se divide en tres secciones: la pregunta, la escala y las opciones. Para explicar mejor cada una de las secciones vamos a construir preguntas como ejemplos; preguntas sobre los movimientos del corazón, donde éstos estén incluidos dentro de la pregunta, en la escala o dentro de las opciones.

Sístole dentro de la pregunta

Se dice que el corazón está en SÍSTOLE cuando:

	[Ventrículo]	[Aurícula]
Entra la sangre al.....	[ ]	[ ]
Sale la sangre del.....	[ ]	[ ]
Sale al pulmón del.....	[ ]	[ ]
Entra al pulmón al.....	[ ]	[ ]

La finalidad de esta cuestión es preguntar todos aquellos sucesos relevantes que ocurren cuando hay sístole.

Sístole dentro de la escala

Marque las correspondencias siguientes:

	[Diástole]	[Sístole]	[Aurícula]	[Ventrículo]
Entra sangre.....	[ ]	[ ]	[ ]	[ ]
Sale sangre.....	[ ]	[ ]	[ ]	[ ]
Se contrae.....	[ ]	[ ]	[ ]	[ ]
Se expande.....	[ ]	[ ]	[ ]	[ ]

El propósito de esta pregunta es hacer resaltar que dentro de la escala también los movimientos del corazón están quizá más claros que en la pregunta anterior.

Sístole como opción

El funcionamiento del corazón se describe como:

	[Paso 1]	[Paso 2]	[Paso 3]	[Paso 4]
Diástole.....	[ ]	[ ]	[ ]	[ ]
Sístole.....	[ ]	[ ]	[ ]	[ ]
Se contrae Aurícula.....	[ ]	[ ]	[ ]	[ ]
Se contrae Ventrículo.....	[ ]	[ ]	[ ]	[ ]
Se expande Ventrículo.....	[ ]	[ ]	[ ]	[ ]

El objetivo aquí es destacar la descripción del proceso, donde el movimiento de la sístole es un componente.

Yo no soy médico, pero un especialista podría hacer una selección de opciones más rica y precisa que la que se proporciona, sin embargo, los ejemplos anteriores dan una idea de la riqueza de información que se puede tener, en comparación con una pregunta de cinco opciones, donde la mayor parte del esfuerzo para construirla está en inventar distractores para engañar.

Una aplicación experimental

La técnica de evaluación matricial puede ser atractiva teóricamente, pero al pretender ser objetiva requiere que se respondan preguntas como: ¿Tiene confiabilidad?, ¿qué criterio de calificación es mejor?, ¿discrimina a los estudiantes? Con el objetivo de responder estas preguntas se aplicó un examen sobre conocimiento de álgebra elemental (véase anexo) a dos poblaciones. Una de estudiantes de la Facultad de Ciencias y otra de la Facultad de Filosofía y Letras, de la UNAM. Ambas facultades, según la tradición, tienen poblaciones diferentes; en una les gustan y tienen facilidad para las matemáticas, en la otra les disgustan y carecen de facilidad para ellas. ¿Discriminaría el instrumento ambas poblaciones? los resultados los discutiremos más adelante, ahora vamos a definir formalmente qué entendemos por confiabilidad, discriminación, etcétera.

Definiciones y modelo teórico

Como bien señalan Diaconis y Efron,

> La teoría estadística se dirige en último término a una cuestión fundamental: ¿Cómo se puede saber qué es lo verdadero? Supongamos que una cierta magnitud se haya medido 15 veces, obteniéndose 15 valores que difieren moderadamente entre sí. ¿Cuál será la aproximación óptima al valor verdadero? los métodos de la estadística permiten responder tal cuestión e incluso dar una estimación cuantitativa de la confianza que tal estimación merece.[6] Pocos párrafos resultan tan claros y tocan el verdadero sentido de la estadística. Ésta es la idea directriz que usamos en este trabajo.

Confiabilidad[7]

¿Qué entendemos por confiabilidad? En los modelos que se han diseñado para las pruebas de opción múltiple, la confiabilidad se mide a través de la fórmula de Spearman-Brown para una división en dos mitades del *test*. Esta confiabilidad se basa en el supuesto de que si los items pares y los *items* nones están altamente correlacionados, como es una división aleatoria, cada *item* contribuye de la misma manera a la calificación total. En otras palabras, cada pregunta es consistente -se puede confiar en ella-, pues las respuestas correspondientes son “malas”para los alumnos "ignorantes" y "buenas" para los que las “conocen”. Esto supone además que las dos mitades son, en cierto grado, semejantes paralelas-, es decir, su media y desviación, el sesgo y las ínter correlaciones entre *items* son independientes.

Otra medida de confiabilidad es la K-R20. Kuder y Richardson trataron de mejorar los insatisfactorios resultados que se obtenían con los métodos Split-Half. La fórmula más exacta de las que propusieron fue la número20, y es la que comúnmente se utiliza. Supongamos el examen ideal con confiabilidad perfecta. ¿Qué significa eso en el modelo K-R20? Significa que la varianza -las diferencias en las medidas debidas a errores- sería cero, y que la medida de covarianzas sería 1. En otras palabras, supone que los *items* estarían fuertemente correlacionados entre sí, de tal forma que cada uno se comportara de manera semejante a cualquier otro, así como el centímetro tres es semejante -por no decir idéntico- a cualquier otro centímetro que usemos en nuestra regla. Supone además que los errores de medida son muy pequeños o nulos.

¿Serán aplicables estos modelos a la evaluación matricial?, ¿influye en ello el que exista una estructura -las opciones concatenadas y las preguntas concatenadas? ¿Se cumplen los supuestos teóricos? Para responder estas cuestiones, nuestra idea es la siguiente; tomemos una pregunta

cualquiera:

¿Qué valor toma x en las ecuaciones siguientes?:

	[1] [2] [3] [4] [5] [6] [7]
6(x-5)+2(x/2)-26=0	[...] [...] [...] [...] [...] [...] [...]
x-1=0	[...] [...] [...] [...] [...] [...] [...]
6x-4=26	[...] [...] [...] [...] [...] [...] [...]
3x-15+(x/2)=13	[...] [...] [...] [...] [...] [...] [...]
x=112/7-3x	[...] [...] [...] [...] [...] [...] [...]
3x+2(x/2)=15	

Podemos suponer que cada opción representa una pregunta de opción múltiple como la siguiente:

¿Qué valor toma x en la ecuación 6X-4=26

A) 1 B) 2 C) 3... G) 7

Donde en este caso particular la opción E) 5 sería la opción verdadera y las otras seis serían distractores. Al hacer esta equivalencia, daríamos por sentado que el modelo de confiabilidad sería aplicable. Pero ¿se sostiene la equivalencia en preguntas concatenadas? Más concretamente, para preguntas como:

Ordena las expresiones que dan solución a la ecuación: 3(x-5) + (x/2) – 13 = 0

	[1] [2] [3] [4] [5] [6] [7]
6(x-5)+2(x/2)-26=0	[...] [...] [...] [...] [...] [...] [...]
x=8	[...] [...] [...] [...] [...] [...] [...]
6x-30+x-26=0	[...] [...] [...] [...] [...] [...] [...]
3x-15+(x/2)=13	[...] [...] [...] [...] [...] [...] [...]
x=56/7	[...] [...] [...] [...] [...] [...] [...]
3x+(x/2)=13+15	[...] [...] [...] [...] [...] [...] [...]
7x=56	[...] [...] [...] [...] [...] [...] [...]
6x+x=26+30	[...] [...] [...] [...] [...] [...] [...]

Podemos traducir la pregunta de opción múltiple de la siguiente manera:

¿Cuál paso en la solución de 3(x-5)+(x/2)-13=0, representa 6x-30+x-26=0

A) Paso 1 B) Paso 2 ... G) Paso 7

Como se observa, si podemos “traducir” una opción de la matriz a una pregunta de opciones múltiples, podemos aplicar entonces el modelo de confiabilidad K-R20.

Sin embargo, existe la objeción de que en este tipo de preguntas las respuestas pueden deducirse sin que haya verdadero conocimiento. En primer lugar, yo preguntaría lo siguiente: si el método de deducción no fuera conocimiento, entonces, ¿dónde estaría la ciencia?, ¿cómo podríamos encontrar información que desconocemos?, además, ¿quién puede pretender hoy ser el gran

inquisidor para definir el verdadero conocimiento? Si a lo que se refiere esa objeción es que se alteran las probabilidades de respuesta, y con ello el modelo teórico de que las opciones son equiprobables, debemos dejar en claro que eso es un modelo ideal, semejante al modelo de los gases ideales que se estudian en termodinámica, y que dicho modelo no es supuesto necesario en el modelo K-R20. La fórmula K-R20, al incluir la covarianzas, es decir, las intercorrelaciones entre las opciones multiplicadas por el error estándar, contempla que si hubiera alteración significativa en la equiprobabilidad, ésta sería ponderada por la fórmula.

Por último, si la calificación es lógica, es decir, si tiene múltiples criterios de *si..., entonces...* ¿sería aplicable? La respuesta a nuestro juicio es positiva, ya que la confiabilidad que estudiamos depende de las ínter correlaciones de las respuestas a cada opción particular y no al criterio de calificación empleado. Cuando se correlaciona una opción con el puntaje total, se investiga la discriminación de ese *items,* pero no la confiabilidad total del examen.

En resumen, el modelo K-R20 es aplicable al modelo de evaluación matricial, y con ello podemos deducir su consistencia interna, es decir, qué tan confiable es (qué confianza tendremos en) la estructura de construcción del examen.

Discriminación

¿Qué entendemos por discriminación? Aquí nos referimos a la discriminación externa, la cual tiene por objeto responder la pregunta: ¿El examen discrimina a los alumnos que tienen el conocimiento, de aquellos que carecen de él?, u otras como ¿en qué grado la discriminación es válida?, ¿qué valor debe tener un alumno con conocimientos y qué calificación el ignorante?, ¿qué confianza tienen esas calificaciones?

En discriminación externa, la técnica que se usa comúnmente es tomar el 25 por ciento de los alumnos con calificaciones más bajas y el 25 por ciento de los alumnos con calificaciones más altas y aplicar una prueba de hipótesis, por ejemplo, una prueba t. Si es significativa, responder afirmativamente a la pregunta sobre la validez externa de la prueba. Si la t es significativa al 5 por ciento de intervalo de confianza, se responde a la pregunta ¿en qué grado de discriminación es válida? Para conocer el valor se debe tener las calificaciones de aquellos que conocen y aquellos que ignoran; se va a la distribución de calificaciones y se buscan las notas inferior y superior de ambos grupos; el error estándar responde como estimador general sobre la confiabilidad de esas calificaciones, y si queremos la respuesta exacta, podemos obtener el intervalo de confianza de una calificación particular.

¿Es aplicable este método a la evaluación matricial?, ¿la calificación lógica cumple los supuestos estadísticos de la valides externa? Cuantitativamente hablando, conviene señalar que una pregunta matricial está calificada en términos de porcentajes. Así, una pregunta tendrá un valor cualquiera en una escala de 0 a 100. Esto es útil, ya que si una pregunta puede ser calificada con distintos criterios –recordemos el ejemplo de pregunta sobre la realidad-, es necesario que tengan valores proporcionales. Cuando una condición lógica del tipo *si..., entonces...* Pregunta = 0 se cumple, es decir, cuando se encuentra inconsistencia en las respuestas, esto es equivalente a tomar un criterio más de calificación, el cual está definido y da un valor proporcional determinado.

La aplicación del modelo del estadístico t puede criticarse, pues si bien es la prueba de hipótesis más sólida para escalas numéricas, es difícil cumplir los supuestos teóricos cuando se aplica un examen a 20 ó 40 estudiantes.

Por fortuna, las investigaciones relativamente recientes de Tukey, Efron y otros (1977)

permiten prescindir del supuesto de que la distribución sea gaussiana, así como también de una serie de supuestos sobre la medida. El método de Efron, que se conoce como *Bootstrap,* se describe así:

> Las muestras *Bootstrap* se engendran como sigue: Los datos del primer alumno se copian un número enorme de veces, pongamos por caso mil millones, y lo mismo se hace con los datos de los otros 20 alumnos. Las 20 000 millones de copias se mezclan concienzudamente. Se seleccionan entonces muestras al azar de tamaño 20, y se calcula, para cada muestra, el estadístico correspondiente[8].

Empleando esta técnica se logran extraer los parámetros estadísticos más próximos al valor verdadero.

Otro criterio sobre la discriminación consiste en correlacionar las calificaciones con un criterio externo. Los criterios externos suelen ser otros *tests* o algunas características de la población que se definen claramente. Nosotros usamos tres preguntas de opinión, además de la prueba t de *student,* aunque, como veremos, su análisis ya no requería de la correlación.

Está claro que se necesitarán otros estadísticos tendientes a completar el análisis de reactivos, para así responder preguntas sobre lo adecuado de tal o cual opción en una pregunta, o si se recomienda substituirla. Pero responder las preguntas sobre la discriminación y confiabilidad de la evaluación matricial ya es un avance, y está por desarrollarse un modelo más completo con los estadísticos que faltan.

Análisis de resultados

El primer paso fue aplicar a un grupo de 26 alumnos de la Facultad de Filosofía y Letras un examen que contenía nueve preguntas, tres de opinión sobre matemáticas y seis sobre ecuaciones y reglas para usar los signos y paréntesis. Nuestra intención era saber si se entendían las instrucciones para su aplicación, si los tiempos para resolver el examen eran adecuados, y, además, probar los programas con datos reales y saber qué criterio de calificación era más conveniente.

Esta aplicación nos permitió comprobar que el examen, en su aplicación, era adecuado. Las instrucciones se comprendieron con facilidad; el tiempo fue de 15 minutos para entregar el primer examen y 30 minutos para el último, y el tiempo en resolverlo la mayoría del grupo fue de un promedio de 20 minutos.

Existe la posibilidad de calificar las preguntas matriciales de diferentes formas. La primera es asignar un punto a cada respuesta correcta. A este método le llamamos puntaje crudo (PC); la segunda es asignar un punto multiplicado por un valor o peso, para distinguir así las opciones más significativas, y le llamamos puntaje pesado (PP). La tercer opción es la calificación lógica, es decir, poner condiciones de *si..., entonces.* En particular, para el examen los condicionales eran muy sencillos; si el alumno respondía correctamente las opciones más elementales o las esenciales, entonces la calificación se anotaba un punto por cada respuesta correcta. En caso contrario, la pregunta completa valía cero. A esta forma de calificación la llamamos puntaje lógico (PL).

1.000	0.946	0.920
0.946	1.000	0.937
0.920	0.937	1.000

TABLA 1. Correlaciones entre métodos de calificación.

El análisis de las correlaciones de los tres métodos de calificación ensayados nos indican correlaciones altas. El puntaje lógico (PL) correlaciona con el puntaje crudo (0.920) y con el puntaje pesado (0.937). El puntaje pesado(PP) correlaciona con el puntaje crudo (PC) (0.946); con ese valor parece que no hay diferencias, siendo el puntaje crudo más sencillo.

Se realizó una segunda aplicación también con un examen sobre álgebra elemental en un grupo de 20 alumnos de la Facultad de Ciencias que cursaban el primer semestre y a un grupo de21 alumnos de la Facultad de Filosofía y Letras que también cursaban el primer semestre.

En cuanto a la confiabilidad medida con K-R20, para el grupo de Ciencias, es Rtt=.93 y para el de Filosofía Rtt=.94. Ambas muy altas para un examen de seis preguntas. Esto se debe a la coherencia de las preguntas, que en lugar de saltar de un tema a otro se concentran en un contexto global. En cuanto a la dificultad promedio que el cuestionario tiene para cada grupo, en el de Filosofía fue de 43.9 (PC) y 26.9 (PL); en el de Ciencias, 63.2 (PC) y 51.45 (PL). La t de *student* fue de 81.83 para el grupo de Ciencias, alejado en un 25 por ciento de los más altos y un 25 por ciento de los más bajos. Esta cifra nos habla de una buena discriminación.

En cuanto a la comparación entre el grupo de Ciencias y el de Filosofía, la t de *student* es de 1.27, es decir, no es significativa para el puntaje crudo, y para el puntaje lógico es de 1.74, significativa al 5 por ciento de nivel de confiabilidad.[9]

Al analizar los datos desde otro punto de vista, podemos llegar a ver gráficamente cómo los métodos de calificación discriminan. Ya hemos descrito el método de Efron, llamado *Bootstrap.*

No utilizamos 20 000 millones de datos, ni extrajimos un millón de muestras; trabajamos con doscientos mil datos, y de ellos extrajimos diez mil muestras. Para comprobar si ese método tan restringido nos era de utilidad, hicimos con el grupo de 21 estudiantes de Filosofía un análisis semejante al que Efron y Diaconis describen en su artículo. Tomamos muestras de cuatro casos, los cuales representan el número más próximo al 17 por ciento de la población. El valor del promedio que nos dio el programa usando el método restringido fue 43.49, cuando el promedio verdadero para ese grupo fue 43.90. Si bien los promedios no coinciden exactamente, la aproximación nos pareció adecuada, sólo 0.41 de diferencia, que en parte se debe al modo de redondear las cifras, y también a que las muestras extraídas son pequeñas (de cuatro casos solamente). De modo que procedimos a hacer el mismo análisis pero con muestras de 20 y 21 alumnos para los grupos de Ciencias y Filosofía respectivamente.

Aquí se presenta el examen y las preguntas que respondieron los alumnos, así como los porcentajes de respuesta a cada una de las opciones que el grupo de Ciencias dio. En las tres primeras preguntas sólo hay opiniones, juicios de valor sobre si les gustan o disgustan las matemáticas, cómo las caracterizan y si pueden resolver o no ciertos cálculos, en su mayoría ecuaciones elementales. La pregunta 4 es de relleno, pero la 5 es fundamental, a nuestro juicio. Si un alumno no conoce las reglas de los signos, no podrá resolver ecuaciones correctamente; es decir, las respuestas a las preguntas 8 y 9 del cuestionario no serán confiables. La pregunta 6 es difícil, ya que las respuestas a las opciones no se presentan directamente. La pregunta 7 es experimental; contiene opciones falsas y nos sirve para estudiar otras estadísticas tendientes a hacer el análisis de reactivos.

Tomando esto en cuenta pudimos definir un criterio que nos permita enfocar desde otro punto de vista la discriminación. El criterio es el siguiente: Aquellos alumnos que respondan el siguiente patrón tendrán calificaciones altas. El patrón es:

Pregunta 1) responder “de acuerdo" a las opciones "tengo facilidad” y “me gustan”.

Pregunta 2) responder "de acuerdo" a las opciones "son un lenguaje" y “son exactas". (Más

tarde pondrán eso en duda).

Pregunta 3) responder con ++ o + a la opción “puedo responder 3x(4-x)=84-3xFD".

Pregunta 5) responder al menos- (b/a) a la opción b/-a y la opción -b/a con -(b/a).

Preguntas 8 y 9) deben ser correctamente respondidas, pues no sólo nos hablan de la sinceridad, sino de que ésta corresponde a una apreciación correcta.

En el grupo de Ciencias hay tres alumnos que cumplen con el patrón; aquellos cuyos números de identificación son el 18, 19 y 20, y sus calificaciones son 72, 76 y 91, respectivamente. Los otros dos alumnos con las más altas calificaciones (96 y 84) no cumplen el patrón porque responden que posiblemente tienen facilidad, es decir, pecan de tímidos, por decirlo de alguna manera, ya que todas las demás condiciones las cumplen. Es interesante constatar que, en el grupo de Filosofía, solamente dos alumnos cumplen con las condiciones de las preguntas 2, 5 y 8, que corresponden a las de matemáticas, pero sus opiniones no cumplen con el patrón definido. Ellos responden con "posiblemente" la opción "me gustan las matemáticas”, y con + la opción “puedo resolver la ecuación". Estos alumnos tienen las más altas calificaciones del grupo (73 y 57), aunque es claro que no tienen una apreciación correcta ni muestran seguridad sobre lo que conocen acerca de matemáticas.

Queremos indicar que el patrón se definió antes de conocer los resultados, sirviendo de guía en el. diseño del examen. El que señale a los mejores alumnos de cada grupo prueba que el criterio de discriminación que se empleó en el método de calificación lógica es, en lo fundamental, correcto. Muestra de ello es que la t de *student* es significativa.

En resumen, el instrumento tiene una confiabilidad más que aceptable (.93). Esto nos habla bien de que es mejor construir preguntas con un contexto coherente. En cuanto a la discriminación, también podemos dar una respuesta positiva, siendo el nivel de confiabilidad

EXAMEN

1. Responda las siguientes frases sobre matemáticas:

	[de acuerdo]	[posiblemente]	[en desacuerdo]
Tengo facilidad	35%	65%	--
No las entiendo	--	50%	50%
Me gustan	80%	15%	--
Las aborrezco	--	--	100%

2. En su opinión, las matemáticas:

	[de acuerdo]	[posiblemente]	[en desacuerdo]
Son áridas	10%	25%	65%
Son de mucha utilidad	95%	5%	--
Son aburridas	--	10%	90%
Son frías	10%	55%	35%
No son didácticas	5%	25%	70 %
Carecen de sentido	--	5%	95%
Son difíciles	20%	55%	25%
Son exactas	65%	20%	15%
Son un lenguaje	75%	25%	

3. Responda las siguientes afirmaciones:

	[++]	[+]	[+-]	[-]	[--]
Puedo resolver $3x^2+2x-7=0$	75%	10%	10%	--	--
Sé sacar raíz cuadrada	40%	10%	30%	10%	10%
Puedo resolver 2x-9=3(4x-3)-63	55%	40%	5%	--	--
Puedo resolver 8(x-1)-32=0	70%	15%	10%	5%	
Puedo desarrollar $(a+b)^2$	80%	15%	--	--	--
Sé sacar logaritmos	10%	35%	45%	5%	5%
Puedo resolver $3x(4-x)=84-3x^2$	55%	35%	5%	--	5%
Puedo desarrollar (n-1) (n+1)	65%	20%	10%	--	--
Puedo resolver $3x^2$-log 6x=15	5%	35%	40%	10%	10%

4. Marque los símbolos que comúnmente se usan en:

	Aritmética	Álgebra	Geometría	Trigonometría
2 + 3	90%	10%	--	--
ß > 120	20%	25%	30%	20%
2	30%	35%	15%	15%
COS	--	20%	25%	55%
COS	5%	15%	30%	50%
AB	10%	50%	40%	--
$3X^2$	5%	80%	5%	5%
½	95%	--	5%	--
SEN	5%	5%	15%	75%

5. Las reglas para usar los signos están dadas por:

	-(a-b)	-(a+b)	-(b/a)	b/a	-(a-b)
b/ -a	5%	--	75%	5%	5%
a (-b)	100%	--	--	--	--
-b / -a	----	15%	80%	5%	
-a -b	20%	60%	--	--	10%
-b /a	----	95%	--	--	

6. Las reglas con fracciones dan las equivalencias de:

	(a+c) /b	a/b + c/d	a/b x c/d	a/b : c/d
ac/bd	--	25%	65%	5%
a/c : b/d	--	--	25%	35%
a/d x c/b	--	--	55%	20%
a/b x d/c	--	5%	25%	40%
(ad + bc) / bd	15%	60%	5%	--
a/b + c/b	80%	5%	--	--
a/c x d/b	--	--	10%	45%

7. Ordene las ecuaciones para resolver: 3(x-2x)+2/4x-15=0

	[1]	[2]	[3]	[4]	[5]	[6]	[7]	[8]
3x-6x+2/4x=15	80%	--	--	5%	--	--	--	--
2=60x+12x²	10%	5%	20%	30%	--	--	--	--
4x(2/4x)=4x(15+3x)	5%	35%	35%	5%	--	--	--	--
12x2+60x-2=0	--	--	20%	20%	35%	--	--	--
x1=0.03	--	--	--	5%	--	25%	--	--
2/4x-3x=15	--	50%	--	10%	10%	--	--	--
x2=-5.03	--	--	--	--	5%	5%	10%	5%
x1=-0.03	--	--	5%	--	5%	5%	20%	5%
x2=4.966	--	--	--	--	--	--	10%	10%

8. Ordene las ecuaciones para resolver: 3x(x+4)=84+3x²

	[1]	[2]	[3]	[4]	[5]	[6]	[7]	[8]
x=7	--	--	--	5%	20%	50%	10%	5%
12x+3x²-3x²=84	15%	25%	60%	--	--	--	--	--
12x+3x² = 84+3x²	20%	65%	5%	--	--	--	--	--
12x=84	--	5%	20%	65%	--	--	--	--
3x(4+x) = 84+3x²	65%	5%	5%	--	--	--	--	--
x=84/12	--	--	10%	25%	65%	--	--	--

9. Qué valor toma x en las ecuaciones siguientes:

	[1]	[2]	[3]	[4]	[5]	[6]	[7]	[8]
8(x-1)-32=0	5%	--	10%	--	75%	--	--	--
3x-4=8	--	--	5%	80%	5%	--	--	--
x-1=0	90%	5%	--	--	--	--	--	--
2x-9=3(4x-3)-63	--	--	--	--	5%	25%	--	--
12x-7=2-15x	--	5%	10%	--	5%	5%	--	--
2x+3=7	--	60%	5%	15%	10%	--	--	--
3x(4-x)=84-3x²	--	--	--	--	10%	--	65%	--

cinco por ciento. Cabe recordar que se trata de alumnos recién ingresados en la Universidad. (Este examen se aplicó una semana después de iniciadas las clases.)

Conclusiones

Tanto la confiabilidad como la discriminación son adecuadas, que por ahora son los únicos estadísticos que se investigaron. En cuanto al método de calificación lógica (PL), debemos decir que requiere investigarse con mayor detalle. Aún no está bien definido el procedimiento para determinar cuáles condiciones son adecuadas y cuáles no. Ni mucho menos un modo de corregir errores. Pero lo que se observa es prometedor.

La evaluación matricial puede ser el instrumento de evaluación que, teniendo las ventajas de los exámenes de opción múltiple, en cuanto a economía, rapidez en la calificación y objetividad, tiene la riqueza de un examen de ensayo tradicional.

NOTAS

1. Era una anécdota de todos conocida en los años de estudiante en la Facultad de Psicología.

2. La evaluación es rechazada por los estudiantes, quienes la sufren, pero se oculta que también lo es por los profesores, pues algunos jefes de departamento y otros funcionarios piensan equivocadamente que con ello pueden "evaluar" al profesor. Esta idea es errónea, ya que no se evalúa al profesor sino a los alumnos. Sí se quisiera “evaluar" al profesor habría que controlar, primero, entre otras cosas, que los alumnos tengan los prerrequisitos necesarios, y hacerlo con instrumentos adecuados, instrumentos que hoy no se tienen. Para ser claros, es como pesar con una balanza medio descompuesta a los comensales, sin tomar en cuenta su metabolismo, para juzgar al cocinero. ¿No es más fácil probar la comida? Esa política absurda del jefecito por juzgar y tener control sobre el profesor, además de que atenta contra la libertad de cátedra, ha sido profundamente perniciosa para la educación, pues ha generado un rechazo hacia la investigación en evaluación, y una evaluación inadecuada ha impedido que se puedan desarrollar nuevos métodos de enseñanza, que la tecnología educacional tenga un adecuado criterio de desarrollo, etcétera.

3. El programa se desarrolló con el apoyo de la Secretaría de Educación Médica de la Facultad de Medicina, aunque su análisis y concepción se hizo en la Coordinación del CCH.

4. La primera versión se desarrolló en la Secretaría de Educación Médica, la segunda, en forma independiente, con la ayuda desinteresada del Instituto de Estudios y Servicios Educativos (IDESE), en cuyas instalaciones se desarrolló la primera versión para un Apple lle. Posteriormente se desarrolló una versión en francés para IBM-PC en las instalaciones de Ecole des Arches de Lausanne. Sus directores, Salvador Ortiz y Chistian Hofer, me ayudaron con múltiples algoritmos y sugerencias útiles. El último año se ha recibido el apoyo de la UNAM, por medio del Centro de Investigaciones y Servicios Educativos. La maestra Alicia Batllori, del CISE, fue quien aplicó los exámenes que se mencionan aquí y proporcionó múltiples sugerencias en la revisión final del manuscrito.

5. Edmund Claparede, Cómo diagnosticar las aptitudes en los escolares. Madrid, Aguilar, 1972, pp. 82-83.

6. Persi Diaconis y Bradley Efron, "Métodos estadísticos intensivos con ayuda del ordenador”, Investigación y Ciencia, junio 1977, p. 70.

7. Las ideas principales sobre confiabilidad y discriminación se tomaron de J. P. Guilfrod y B. Fruchter, Estadística aplicada a la psicología y la educación. México, McGraw-Hill, 1984; principalmente los capítulos 17 y 18.

8. ibid., p. 73.

9. Guilford, op. cit., p. 468.