Búsqueda de la calidad en los procesos de medición de los aprendizajes

Emilio Aliss

Departemento de Ciencias Exactas e Ingenierías
Universidad Católica Boliviana
Cochabamba, Bolivia
e-mail: aliss@ucbcba.edu.bo

Introducción

La evaluación basada en exámenes a preguntas abiertas, cuya corrección difícilmente puede dejar de recibir la influencia de la subjetividad del corrector, da lugar a una multiplicidad de problemas que restan calidad a la evaluación. Partiendo de un análisis sobre estos problemas, que se han puesto en evidencia en las evaluaciones tradicionales y que han sido el objeto de estudios e investigaciones, sobre todo durante la segunda mitad del siglo XX, ponemos a consideración del lector una alternativa cuidadosamente estudiada y probada en la Universidad de Lieja, en Bélgica, que combina los exámenes de selección múltiple, la taxonomía de Bloom, la estadística y la tecnología.

Lejos de proscribir los exámenes a preguntas abiertas, este artículo pretende generar una reflexión sobre la necesidad de buscar nuevos instrumentos , sin dejar de perfeccionar los actuales, para encontrar en una fecha cercana un compromiso que nos permita reducir el margen de error inherente a la medición de los procesos de Enseñanza-Aprendizaje.

Defectos de la Evaluación Tradicional

El problema de la evaluación basada en las pruebas a preguntas abiertas estriba principalmente en los elementos subjetivos que intervienen en su corrección. Los efectos identificables de la subjetividad del docente son [2]:

- **Efecto de severidad.** En un grupo de docentes, siempre podemos encontrar algunos que corrigen con más severidad que otros. El grado de severidad depende fuertemente de la personalidad y de la experiencia de cada docente. Dos docentes que corrigen la misma prueba difícilmente coincidirán al 100 % en su apreciación.
- **Efecto de tendencia central.** Un docente apuesta rara vez por las notas extremas (notas pésimas o notas excelentes). Por prudencia, tiene tendencia a asignar notas que se acercan a la media de un grupo determinado.
- **Efecto de halo.** En grupos pequeños, en los que el docente tiene

la posibilidad de conocer a todos sus estudiantes, la impresión que cada uno de ellos causa en él (por sus actitudes, su manera de vestir, su aspecto físico, etc.) tiene una influencia sobre la nota asignada.

- **Efecto de estereotipo.** El docente tiene tendencia a asignar notas buenas a los estudiantes cuyas notas anteriores son buenas y notas malas a los estudiantes cuyas notas anteriores son malas.

- **Efecto de secuencia.** Un examen corregido después de otro que ha merecido un excelente resultado será calificado con una severidad adicional. Un examen corregido después de otro que ha merecido una nota muy mala será corregido con especial benevolencia.

- **Efecto de relativización.** El docente tiene tendencia a establecer un rango de calidad, incluso entre las pruebas que presentan las mismas características.

Un estudio realizado sobre los resultados del bachillerato francés en 1967, muestra claramente la poca fiabilidad de la nota asignada por un docente después de la corrección. Un grupo de seis expertos para cada área evaluada ha corregido independientemente los exámenes. Los resultados se muestran en el cuadro 1.

Características de una Evaluación Ideal

Una evaluación ideal debería tener las siguientes características [2]:

1. *V*alidez: Los resultados deben reflejar lo que el docente quiere medir.

2. *F*idelidad: La nota del examen debe mantenerse si las condiciones de corrección son cambiadas.

3. *S*ensibilidad: La medida debe ser precisa.

4. *D*iagnosticidad: El diagnóstico preciso de las dificultades de aprendizaje, de los procesos adquiridos y de los no-adquiridos debe ser posible.

5. *P*racticabilidad: La evaluación debe ser realizable en los tiempos previstos y con los medios técnicos y humanos a disposición.

6. *E*quidad: Todos los alumnos deben ser evaluados con los mismos parámetros.

7. *C*omunicabilidad: todas las informaciones no-confidenciales deben estar a disposición de todos los involucrados en el proceso de evaluación.

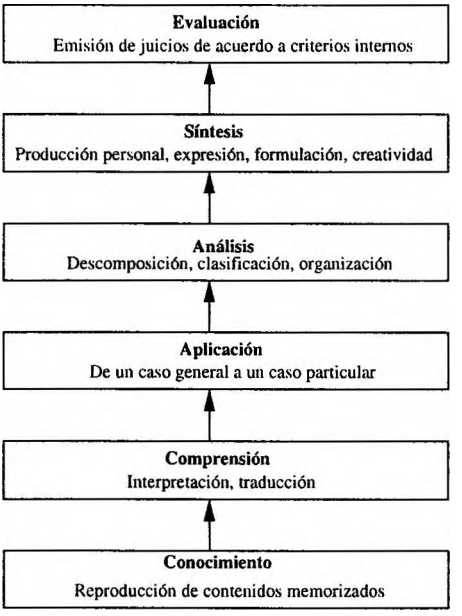
Taxonomía de Bloom

En 1948, en el Congreso de la American Psychological Association [3] y [4], en Boston, los examinadores de la educación superior centraban su atención sobre una evaluación escolar demasiado dirigida a la restitución de la memoria, con el consecuente efecto negativo sobre el aprendizaje significativo y duradero. Fue en este congreso que Benjamín S. Bloom planteó la idea de una taxonomía de objetivos cognitivos. Esta idea fue analizada y trabajada durante años, hasta desembocar en una publicación oficial, realizada en 1956. La taxonomía de Bloom

	Reprobados por los seis correctores	Aprobados por los seis correctores	Aprobados por unos y reprobados por otros
Versión latina	40 %	10 %	50 %
Composición francesa	21 %	9 %	70 %
Inglés	37 %	16 %	47 %
Matemáticas	44 %	20 %	36 %
Filosofía	9 %	10 %	81 %
Física	37 %	13 %	50 %

Cuadro 1: Resultados de las correcciones realizadas por seis diferentes correctores (en cada área) de los exámenes del bachillerato francés en 1967 [2].

establece seis niveles estrictamente jerarquizados de objetivos cognitivos:



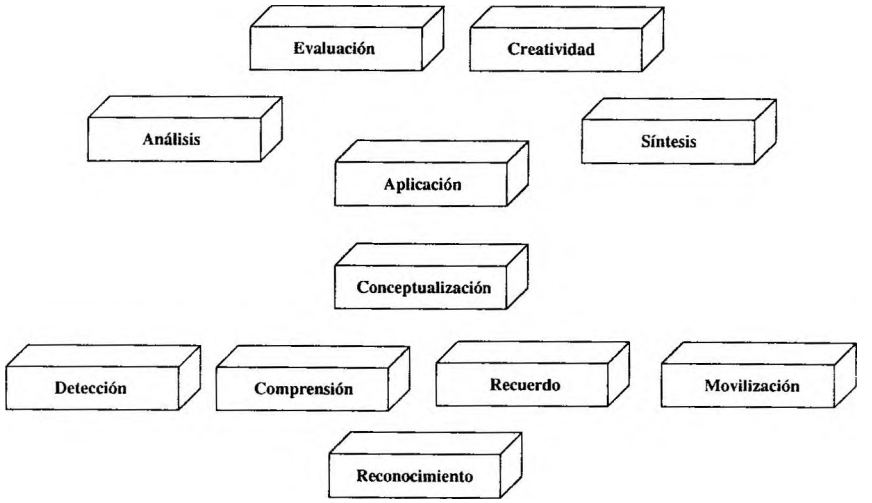
Taxonomía de Dieudonné Leclereq

Esta primera idea de Bloom dio lugar a una extensa investigación sobre la evaluación y trajo a su vez una reflexión profunda sobre los procesos de Enseñanza-Aprendizaje. Adolece, sin embargo, de dos defectos [3] y [4]:

- Limita todos aquellos procesos que no pertenecen a estos seis niveles.
- La jerarquía estricta entre los niveles carece de flexibilidad e ignora las ricas relaciones que pueden establecerse entre los diferentes niveles.

Es por esta razón que la idea original fue enriquecida por otros educadores, entre los que podemos mencionar a Gilbert de Landsheere, profesor emérito de la Universidad de Lieja y Premio Mundial de la Educación y a Dieudonné Leclercq, también profesor en la Universidad de Lieja y responsable de varios grupos de trabajo concentrados en el mejoramiento de la evaluación. Leclercq propone la siguiente taxonomía, en la cual, a pesar de existir una jerarquía entre los distintos niveles, cada nivel contiene varios procesos, entre los cuales se establecen relaciones horizontales.

Estas taxonomías, y otras existentes, nos permiten tener una visión global y completa de todos los elementos que constituyen el proceso de aprendizaje y nos permiten, por ende, apuntar a uno o varios de los componentes de este proceso. La definición de objetivos claros para



la medición, lleva también a una selección adecuada de instrumentos de medición.

Las Preguntas a Selección Múltiple (PSM)

El examen a selección múltiple (compuesto por PSM correctamente elaboradas) se presenta como una alternativa interesante a los exámenes a preguntas abiertas, para lograr las condiciones que requiere una buena medición ([3] y [4]). Para utilizar un examen a selección múltiple, sin embargo, es necesario establecer claramente las reglas de redacción de las pruebas y los parámetros de corrección que permitan evitar el peligro de que el estudiante adivine la respuesta sin conocer del tema. El siguiente ejemplo ilustra claramente este peligro [4]:

Cuando se aumentan cristales de Bestor al agua:

a. Se desprende calor

b. La temperatura de la solución se eleva

c. La solución se torna azul

d. El recipiente se calienta

Un poco de observación nos mostrará que las respuestas a, b y d tienen algo en común, de manera que la respuesta c resulta aventajada, por ser diferente. Efectivamente, en estudios realizados sobre grupos de estudiantes, esta respuesta, que es la correcta, fue la preferida para:

- El 45 % de los que respondieron a la pregunta en la experiencia de Diamond y Evans (1972)
- El 52 % de los estudiantes antes del aprendizaje y el 79 % después del aprendizaje en la experiencia de Slakter *et al.* (1972)

Las Reglas de Redacción

Para evitar problemas como el que se muestra en el párrafo anterior, varios educadores han estudiado la manera de optimizar la redacción de las PSM. Entre ellos, después de varios años de expe-

rimentación y reflexión, el profesor Dieudonné Leclercq propone 20 reglas básicas de redacción. Es posible agrupar estas reglas, de acuerdo a [1] y [4], por:

1. La adecuación a los objetivos (3 reglas)
2. El valor diagnóstico de la respuesta (3 reglas)
3. La forma (6 reglas)
4. Las soluciones propuestas (8 reglas)

Versión Mejorada del Examen a Selección Múltiple

Si las preguntas a selección múltiple están bien redactadas, es posible evaluar el conocimiento, la comprensión y la aplicación (los tres primeros niveles de la Taxonomía de Bloom), pero no es posible evaluar la capacidad de análisis, de síntesis, ni la capacidad metacognitiva (último nivel de la taxonomía de Bloom). Para evaluar los niveles superiores, es necesario:

- Aumentar a las respuestas tradicionales de una PSM las llamadas repuestas generales, que permiten medir la capacidad de análisis.
- Establecer rangos de seguridad que desarrollen en el estudiante la capacidad metacognitiva (último nivel en la taxonomía de Bloom).

Con estas dos adiciones, las PSM permiten evaluar cinco de los seis niveles establecidos por la Taxonomía de Bloom: el conocimiento, la comprensión, la aplicación, el análisis y la evaluación (capacidad metacognitiva). No se ha encontrado todavía una técnica que permita a una PSM, evaluar la capacidad de síntesis del estudiante. Esa capacidad debe ser evaluada utilizando otro instrumento diferente.

Medición de la Capacidad de Análisis

Las cuatro respuestas llamadas generales, que se incluyen como respuestas posibles a una PSM, permiten medir la capacidad de análisis del estudiante. Estas respuestas son ([3] y [4]):

1. Ninguna de las respuestas anteriores es correcta
2. Todas las respuestas anteriores son correctas
3. Es imposible responder, porque falta información
4. El enunciado es absurdo

Si el estudiante sabe que, además de las respuestas tradicionales, una de estas cuatro es siempre posible, está obligado a realizar un análisis detallado de los conceptos ligados a la pregunta formulada.

Grados de Seguridad

Si pedimos al estudiante que declare cuan seguro está de su respuesta, estableciendo un rango claro de posibilidades de elección, estamos incluyendo los llamados grados de seguridad. Algunas de las razones que han conducido al equipo de trabajo del profesor Dieudonné Leclercq, en la Universidad de Lieja, a utilizar los grados de seguridad como uno de los elementos de la evaluación son las siguientes [2]:

1. **La incompetencia es un estado normal en la vida.** Hay muchos más dominios en los cuales somos incompetentes que aquellos en los cuales somos competentes. Es la razón que nos lleva a disminuir nuestro nivel de incompetencia.

2. **La ignorancia consciente no es peligrosa.** La conciencia de la ignorancia solo puede llevarnos a buscar el conocimiento, en cambio la ignorancia inconsciente es muy peligrosa, porque nos puede inducir a cometer errores, en algunos casos muy graves.

3. **La ignorancia disimulada es peligrosa.** El temor social a la ignorancia lleva a muchas personas a ocultar su ignorancia, a pesar de las consecuencias que esto puede tener. Una enfermera que no está segura del remedio que debe administrar y no solicita información por temor a ponerse en evidencia, puede causar daños serios a sus pacientes.

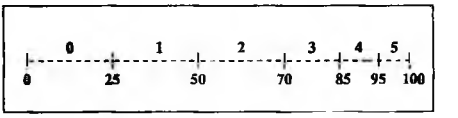
4. **La duda es el motor del conocimiento.** Son las dudas las que impulsan el progreso de la ciencia y la búsqueda del conocimiento, a todo nivel.

5. **La emisión de juicios es uno de los objetivos de la evaluación.** En la taxonomía de Bloom, la capacidad de evaluar está en la cima de la pirámide. La metacognición es un objetivo muy importante de la evaluación. Sin este objetivo logrado, un médico, por ejemplo, no podría dar un diagnóstico final, después de evaluar a su paciente. Si el estudiante es capaz de emitir juicios sobre la calidad de su propio aprendizaje y autoevaluar la seguridad con la que ha adquirido sus conocimientos, podrá mejorar constantemente y buscar la solidez intelectual.

6. **La lógica del conocimiento no es necesariamente binaria.** En muchos casos, el conocimiento parcial de la respuesta, puede llevarnos al conocimiento completo, pero es necesario tener conciencia de esta parcialidad.

7. **La auto-evaluación se aprende por experiencia personal.** La única manera de aprender a autoevaluarse es enfrentando las consecuencias de los propios actos y decisiones personales.

Después de varias experiencias realizadas en Lieja, el profesor Dieudonné Leclercq, a partir del análisis de resultados y con el apoyo de la teoría de decisiones, ha propuesto el siguiente baremo de grados de seguridad:



Donde el 0 significa que el grado de seguridad varía entre 0 y 25 % (grado de seguridad muy bajo) y el 5 se acerca a la seguridad total.

La elección del grado de seguridad permite medir la capacidad metacognitiva del estudiante, quien tiene conocimiento de las consecuencias de su elección. Estas consecuencias aparecen en el siguiente cuadro 2, en la cual RI significa respuesta incorrecta y RC significa respuesta correcta. Un estudiante que responde correctamente, por ejemplo y está inseguro de su respuesta (código 0), sólo gana 13 puntos de los 20 reservados a una respuesta correcta con una seguridad total. En cambio, un estudiante que

responde incorrectamente, pero es consciente de su inseguridad, tiene 4 puntos en lugar de 0. El castigo máximo está reservado para un estudiante que responde incorrectamente y dice estar seguro de su respuesta.

Una pregunta de un examen de selección múltiple, con todas las modificaciones propuestas por el grupo de investigación del profesor Leclercq, y asumiendo que las 20 reglas de redacción han sido aplicadas cuidadosamente para evitar desviaciones en la evaluación, tendría la siguiente forma:

Un proyectil lanzado en la cercanía de la superficie terrestre, sobre un terreno totalmente plano, con una velocidad **V** y formando un ángulo θ con la horizontal:

1. Tendrá un movimiento cuya componente horizontal es uniforme. O
2. Tendrá un movimiento cuya componente vertical es uniformemente acelerada. O
3. Tendrá una trayectoria parabólica. O
4. Viajará el mismo tiempo durante el trayecto de subida y el de bajada. O
5. Tendrá una velocidad final de misma magnitud que la velocidad inicial. O
6. Ninguna de las respuestas anteriores es correcta. O
7. Todas las respuestas anteriores son correctas. O
8. Es imposible responder, porque falta información. O
9. El enunciado es absurdo. O

Mi grado de seguridad es:
O 0 ; O 1 ; O 2 ; O 3 ; O 4 ; O 5

Información, Entrenamiento y Gestión de Resultados

Para que este tipo de examen tenga realmente buenos resultados, además de la preparación cuidadosa de los autores de las preguntas, los estudiantes deben tener [3]:

- Una información completa y precisa sobre las consecuencias de sus respuestas (ver Cuadro 2) y la importancia y el significado de las respuestas generales.
- Un entrenamiento previo, con un examen de prueba sin valor curricular, para evitar sorpresas desagradables.

Para llevar adelante todas las tareas de información, entrenamiento y gestión de resultados de los exámenes a selección múltiple en una Universidad, de manera a garantizar el éxito del instrumento de evaluación propuesto en este artículo, es necesario contar con un Centro de Apoyo a la Evaluación. Este centro, dotado de la tecnología adecuada y con un equipo humano especializado, debe estar al servicio de toda la Universidad. La Universidad de Lieja, pionera en el desarrollo de las preguntas a selección múltiple mejoradas, cuenta con este centro, que lleva adelante, además, un trabajo de investigación que permite buscar constantemente la optimización de la evaluación.

Sistema Metodológico de Apoyo a la Realización de Tests (SMART)

La Universidad de Lieja cuenta con un centro especializado de servicio y de

Código	Zona de confianza	Centro	RI	RC
0	0 % - 25 %	12.5 %	+4	+13
1	25 % - 50 %	37.5 %	+3	+16
2	50 % - 70 %	60.0 %	+2	+17
3	70 % - 85 %	77.5 %	0	+18
4	85 % - 95 %	90.0 %	-6	+19
5	95 % - 100 %	97.5 %	-20	+20

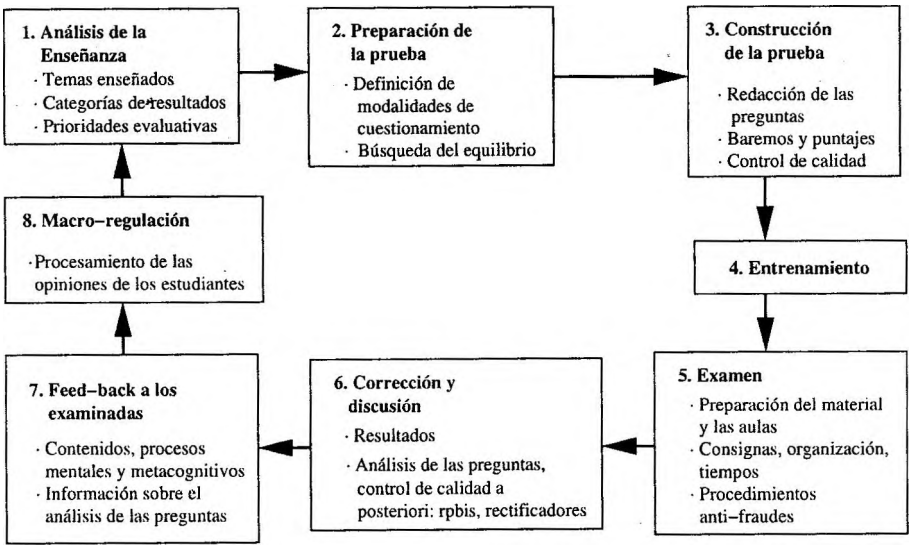
Cuadro 2: Baremo de seguridad con los puntajes asignados a las respuestas correctas (RC) e incorrectas (RI) [4].

investigación denominado Sistema Metodológico de Apoyo a la realización de Tests, formado por un equipo de pedagogos, psicólogos, técnicos e informáticos que dedican su tiempo al mejoramiento de la evaluación. Además de llevar adelante investigaciones sobre el mejoramiento de la evaluación, el equipo del SMART presta servicios a toda la Universidad de Lieja en el campo de la evaluación. Todos los docentes que desean aplicar este tipo de pruebas reciben todo el asesoramiento y apoyo necesarios para llevar a buen término cada examen. El servicio del SMART abarca:

- La formación adecuada para la elaboración de las preguntas a selección múltiple a todos los docentes que deseen tomar este tipo de tests.
- La impresión de los exámenes en el formato adecuado, una vez que el docente ha elaborado sus preguntas.
- La corrección de las pruebas, utilizando el lector óptico.
- La elaboración de un informe con todos los resultados y las anomalías encontradas durante la corrección.
- Un análisis estadístico de los resultados, que permite situar la calidad de aprendizaje de toda la clase.
- Un análisis de la calidad de todas y cada una de las preguntas, utilizando indicadores estadísticos bien establecidos, con un informe posterior al docente, indicando las preguntas que podrían estar en tela de juicio, si es que existen.
- Correcciones posteriores de los resultados, si el docente lo desea, a partir de la re-ponderación de las preguntas de mala calidad (si es que existen).
- Documentación de todos los resultados de las pruebas, para fines legales y cualquier reclamo de estudiantes.

La Espiral de Calidad del SMART

El equipo de trabajo del SMART ha establecido un procedimiento que le permite mejorar la evaluación en cada una de las materias a las cuales brinda servicio. Este procedimiento, que consta de ocho etapas, se denomina "Espiral de calidad". Con cada examen, se recorren cuidadosamente las ocho etapas, que se



constituyen en filtros de error, que permiten optimizar la evaluación, al terminar cada ciclo. Las ocho etapas de la espiral de calidad están esquematizadas arriba [2]. Jean-Luc Gilles, director del SMART, ha desarrollado en su tesis doctoral procedimientos estadísticos para realizar un control de calidad de las preguntas del examen. Este control de calidad se lleva a cabo en la 6ta etapa de la espiral de calidad.

La experiencia ganada por el SMART en el campo de la evaluación es muy valiosa para inspirar la búsqueda de calidad en los procesos de evaluación educativa. El equipo del SMART está listo para emprender un proyecto con la Universidad Católica Boliviana, que permita optimizar los procesos de medición de aprendizajes en Cochabamba.

Referencias

[1] Pascal Detroz. Technologie des évaluations pédagogiques. Módulo del Curso en la Universidad de Lieja,dictado por Pascal Detroz, 2003.

[2] Jean-Luc Gilles. *Qualité spectrale des tests standardisés universitaires.* Tesis Doctoral, Université de Liège, Faculté de Psychologie et des Sciences de l'Education, 2002.

[3] Jean-Luc Gilles. Technologie des évaluations pédagogiques. Módulo del Curso en la Universidad de Lieja, dictado por Jean-Luc Gilles,tutor del curso, 2003.

[4] Dieudonné Leclercq. Edumétrie et docimologie. Université de Liège, Faculté de Psychologie et des Sciences de l'Education, 1999.