Diseño de un sistema de validación de reactivos con base al constructivismo

Design of an item validation system based on constructivism

Beatriz Beltrán Martínez
Benemérita Universidad Autónoma de Puebla
bbeltranmtz @gmail.com

Ana Patricia Cervantes Márquez
Benemérita Universidad Autónoma de Puebla
cervantes.patty@gmail.com

Víctor Enrique Hernández Hernández
Benemérita Universidad Autónoma de Puebla
enrikisimo90@gmail.com

Resumen

La evaluación del aprendizaje siempre ha sido tema de estudio, ésta tiene como función el valorar el nivel de conocimientos y habilidades obtenidas por el alumno, así, con el paso del tiempo se han empleado diferentes tipos de instrumentos de evaluación los cuales contienen reactivos que se elaboran en función del objetivo a medir. De manera general existen dos tipos de reactivos: los reactivos de respuesta de elección múltiple y los reactivos de pregunta abierta o ensayo. Los reactivos de respuesta abierta o ensayo son un poco más complejos de evaluar, al exigir al examinado una respuesta menos concreta y que puede variar dependiendo de la persona.

Respecto a la parte pedagógica existen diferentes corrientes que pueden considerarse en el proceso de aprendizaje-enseñanza, en este sentido, en la actualidad no solo se evalúan conocimientos, en particular

en el constructivismo se considera que se debe evaluar además capacidades, destrezas, habilidades y actitudes.

La presente investigación pretende implementar un sistema que valide reactivos con la finalidad de crear una base de datos para la automatización de exámenes, con base al constructivismo, la cual servirá de apoyo en el trabajo de aplicación de exámenes departamentales que se vienen realizando en la Facultad de Ciencias de la Computación desde el año 2001, así mismo permitirá la realización de análisis de los resultados para la toma de decisiones que permitan mejorar el proceso de aprendizaje -enseñanza en beneficio de los estudiantes.

Abstract

Evaluation of learning has always been a subject of study, the purpose of which is to assess the level of knowledge and abilities obtained by the student. Thus, different types of evaluation instruments have been employed over time, which contain items prepared according to the objective to be measured. Generally, there exist two types of items: multiple-choice items and open-question or essay items. The open-answer and essay items are slightly more complex to evaluate, upon requiring a less concrete answer from the examinee that may vary depending on the person.

Regarding the educational part, there exist different trends that may be considered in the learning-teaching process. In this sense, not only is knowledge currently being evaluated, constructivism in particular considers that capacities, skills, abilities, and attitudes should be evaluated as well.

This research expects to implement a system that validates items with the purpose of creating a database for exam automation, based on constructivism. This will serve as support for applying departmental exams that are being carried out at the School of Computer Sciences since the year 2001. Likewise, it will allow for the realization of results analysis in order to make decisions that will improve the learning-teaching process to benefit students.

Palabras clave / Key words: Validación de reactivos, constructivismo, evaluación, enseñanza-aprendizaje.

Item validation, constructivism, evaluation, learning-teaching

Introducción

En la actualidad tener herramientas de trabajo para que la evaluación cumpla con diversos estándares ha cobrado importancia en estos tiempos, en México existen dependencias que han estado trabajando en este sentido. Además existen métodos para evaluación y análisis de reactivos que facilitan la elaboración de pruebas de opción múltiple tales como las de CENEVAL o las pruebas ENLACE diseñadas y operadas por la SEP.

El CENEVAL, desde 1994, se ha encargado de evaluar en base a instrumentos de medición que se elaboran en el Centro a través de procesos estandarizados de diseño y construcción, que se apegan a las normas internacionales; en su elaboración participan numerosos cuerpos colegiados integrados por especialistas provenientes de las instituciones educativas más representativas del país y organizaciones de profesionales con reconocimiento nacional [1].

En el caso de la prueba ENLACE, ésta se diseña bajo estándares internacionales de calidad, con la asesoría de comités académicos y de un Consejo Técnico, en el que participaron expertos del Educational Testing Service (ETS) y la Universidad Complutense de Madrid. La elaboración y validación de reactivos se realiza por expertos en el diseño de pruebas y maestros con amplia experiencia docente en el nivel Educativo Medio Superior y Superior.

En el caso particular de la Benemérita Universidad Autónoma de Puebla (BUAP), los Lineamientos Generales de la Evaluación Colegiada del Aprendizaje por Asignatura fueron desarrollados con el fin de poder brindar elementos para la implementación de estrategias para el proceso de evaluación en concordancia con el Modelo Universitario Minerva, reafirmándose el hecho de que la educación integral requiere de acciones e instrumentos de evaluación que tiendan también hacia la integralidad del mismo proceso [2]. Es por esta razón que la elaboración de exámenes departamentales se ha vuelto una práctica recurrente en la mayoría de las unidades académicas de la BUAP; lo característico de ésta forma de evaluación es que aún no se cuenta con los criterios claros para su valoración.

En cuanto a qué se evalúa, la gran mayoría coinciden en la parte de conocimientos conceptuales de los cursos, dejando de lado lo procedimental y actitudinal-valoral; a excepción de los programas del área de la Salud y la Escuela de Artes, quienes si han tendido a evaluar los procedimientos y las actitudes. Es necesario un análisis y seguimiento de los resultados, esto con el fin de que la retroalimentación ayude en casos futuros a mejorar el proceso de evaluación. [2]

Con un sistema que ayude a evaluar y analizar reactivos, en lo particular, de pregunta abierta, las cuales se podrían incluir para crear pruebas más completas que abarquen más áreas de conocimiento, debido a que los reactivos de respuesta abierta o ensayo son complejos y exigen al examinado una respuesta que no sea concreta y que puede variar dependiendo de la persona.

Marco Teórico

Los reactivos o ítems

Los reactivos o ítems son los indicadores que se utilizan en una prueba para determinar el grado de dominio de algún tema en particular.

Existen diferentes tipos de reactivos entre los que se destacan principalmente: los de opción múltiple, falso-verdadera, aparejamiento y los reactivos de tipo ensayo.

Los reactivos de opción múltiple constan de dos partes: una proposición que se expresa en forma directa o como una oración incompleta y las otras actúan como distractores. Los reactivos de falso-verdadero sirven para evaluar conocimientos que inequívocamente son ciertos o falsos. Los de aparejamiento se elaboran con dos o más columnas de palabras, símbolos, frases u oraciones, mismas que el alumno deberá asociar o

relacionar de algún modo. Los reactivos de pregunta abierta o de ensayo consisten precisamente en una interrogativa ante la cual, el examinado deberá redactar y justificar su respuesta.

El propósito de los reactivos que constituyen las pruebas de evaluación, es obtener datos que permitan hacer inferencias acerca del conocimiento que tiene un estudiante respecto al dominio evaluativo que mide la prueba, información que es útil a los educadores para tomar decisiones tendientes a mejorar el proceso educativo. Sin embargo, uno no puede saber si una prueba representa adecuadamente el dominio evaluativo si éste no está explícitamente definido, es por tal motivo que cada uno de los reactivos dentro de la prueba tiene que pasar por un proceso de validación, mediante mediciones estadísticas y de discriminación, con el fin de establecer un estándar que al aplicarse mediante una prueba, éstos evalúen exactamente lo que la institución requiere.

Tests- Pruebas.

Son instrumentos de medición que nos permiten evaluar conocimientos, habilidades y aptitudes de una o más personas frente a determinados objetos o situaciones en diferentes áreas. Los criterios de medición de dichos test son establecidos por procesos estadísticos o clínicos.
Los test deben de estar acompañados de normas de aplicación, cálculo e interpretación de los resultados, ya que, en algunos casos, eso reduciría la ambigüedad entre los reactivos. Dependiendo de lo que se va a medir con la prueba, ya sea desempeño laboral, escolar, deportivo, etc., éstas van ligadas a un criterio que determina la empresa que evalúa.

Rúbricas.

Para la evaluación de los reactivos generalmente se utilizan rubricas, que son instrumentos que nos permiten evaluar el nivel de desempeño o alguna tarea. Son guías precisas que valoran los aprendizajes y productos realizados. Indican el logro de los objetivos curriculares y las expectativas de los docentes.

Permiten que los estudiantes identifiquen con claridad la relevancia de los contenidos y los objetivos de los trabajos académicos establecidos. [3]

Análisis de ítems.

Cuando hablamos de análisis de ítems, tenemos que tener en cuenta su objetivo principal. El realizar un continuo análisis de las pruebas hará que su calidad mejore, además de poder emplear los datos obtenidos para ser comentado con los alumnos para rescatar ciertos puntos relativos a las dificultades presentadas durante la aplicación. Esto nos da cuenta de los principales errores, ya sea por puntos difíciles o al comprender mal algún reactivo, además que se mejorará la calidad de enseñanza en las instituciones donde se apliquen dichos análisis.

Teoría clásica del test. [4]

Para el análisis recurrimos a ciertas medidas que nos indican si la calidad de los reactivos es la requerida para ser aplicados en un test. El índice de dificultad y discriminación, la correlación de Pearson (punto biserial) y la teoría de respuesta al ítem son algunos de los elementos que ayudan a dicho análisis.

Índice de dificultad. Entendemos como índice de dificultad (P_i) a la cantidad de examinados que aciertan a un ítem (A_i), entre el resto que intentó resolverlo (N_i):

$$P_i = {}^{A_i}/_{N_i}$$

Entre más cercano es el índice a 0, más difícil es el reactivo, mientras que será más sencillo mientras más se acerque a 1, como se muestra en la Tabla I.

Tabla I. Valoración del índice de dificultad

p índice de dificultad	CALIDAD
< 0.32	Difíciles
0.33 – 0.52	Medianamente Difíciles
0.53 – 0.73	Dificultad media
0.74 – 0.86	Medianamente fáciles
> 0.86	Fáciles

Índice de discriminación (ID). Es la diferencia en la frecuencia con que un grupo de examinados de puntaje alto (P_s), y la frecuencia con la que un grupo de puntaje bajo aciertan un reactivo (P_d), entre grupo mayor (N_p), el cual se muestra en la tabla II.

$$ID = {(P_s - P_d)}/{N_p}$$

Tabla II. Valoración del índice de discriminación

ID	Calidad	Recomendaciones
> 0.39	Excelente	Conservar
0.30 – 0.39	Buena	Posibilidades de mejorar
0.20 – 0.29	Regular	Necesidad de revisar
0.00 – 0.20	Pobre	Descartar o revisar a profundidad
< -0.01	Pésima	Descartar definitivamente

Un ítem tiene una discriminación perfecta cuando todos los sujetos del grupo superior lo responden correctamente y ninguno del grupo inferior lo hace. Si el ítem por el contrario presenta una discriminación negativa debe ser descartado.

Coeficiente biserial puntual. Es una medida de que podemos usar en el análisis para conocer la validez de un reactivo con criterios externos, lo obtenemos correlacionando de manera binaria las calificaciones de un reactivo, siendo 1 la respuesta correcta y 0 la incorrecta.

$$r_{pb} = \frac{(\overline{Y}_p - \overline{Y})\sqrt{n_t\, n_p / [(n_t - n_p)(n_p - 1)]}}{S_t}$$

Donde:

r_{pb}: Coeficiente biserial puntual.

n_t: Cantidad total de examinados.

n_p: Cantidad de examinados que resuelven correctamente el reactivo.

$\overline{Y}_p$: La media de las calificaciones de un examen de quienes aciertan el reactivo.

$\overline{Y}$: La media de todas las calificaciones del examen.

S_t: La desviación estándar de todas las calificaciones del examen.

Cuanto más elevada sea la correlación entre el reactivo y la calificación, más preciso será el reactivo como predictor de la calificación. Consideramos como una correlación alta cuando ésta es mayor o igual a 0.2. Cuando un reactivo tiene una correlación menor, se sugiere someterlo a revisión o descartarlo.

Las corrientes pedagógicas

En cuanto las diferentes corrientes pedagógicas utilizadas durante el proceso de evaluación, se pueden considerar: el conductismo, la cognitiva, el constructivismo, el modelo o enfoque por competencias, entre otros. En este trabajo nos enfocamos en el constructivismo debido a que dentro del área de las ciencias,

existe un interés porque el alumno desarrolle métodos propios para resolver un problema con las herramientas que se le proporcionan y no solo que simplemente obtenga la información.

Constructivismo

El constructivismo es una teoría que equipara al aprendizaje con la creación de significados a partir de experiencias [4].

Los constructivistas no niegan la existencia del mundo real, pero sostienen que lo que conocemos de él nace de la propia interpretación de nuestras experiencias. Los humanos crean significados, no los adquieren. Dado que de cualquier experiencia pueden derivarse muchos significados posibles, no podemos pretender lograr un significado predeterminado y "correcto". Los estudiantes no transfieren el conocimiento del mundo externo hacia su memoria; más bien construyen interpretaciones personales del mundo basándose en las experiencias e interacciones individuales.

Tanto el estudiante como los factores ambientales son imprescindibles para el constructivismo, así como también lo es la interacción específica entre estas dos variables que crean el conocimiento.

Propuesta

Se ha realizado un sistema que evalúa los ítems como requerimos, es decir, que evalúa y compara los ensayos redactados como respuesta por los examinados. El fin es conocer lo que la mayoría de las personas que respondieron para determinado reactivo, y conocer, en determinado momento si es lo que se esperaba.

Considerando que un reactivo cuya respuesta es abierta, no nos podemos detener a pensar que únicamente se va solicitar que se redacte un breve ensayo como respuesta.

Se parte del hecho de que los reactivos a valorar permitirán formar el banco de preguntas de los test que se utilizaran para evaluación de los alumnos. En estos reactivos no se solicita una respuesta redactada a manera de ensayo, sino que se solicita principalmente la elaboración de algoritmos, pruebas de escritorio e incluso demostraciones.

En el caso que corresponde a los reactivos mencionados anteriormente, se pensó en el manejo de rúbricas para tener el control cuantitativo de las respuestas, sin detenernos en las diferencias que los examinados pudieran tener al responder, si no en el porcentaje que fue respondido. La rúbrica, por lo general, aparece implícita en el reactivo; para otros casos, se ocupará un campo extra donde se explicará explícitamente la rúbrica correspondiente al reactivo.

Con esta información, se desarrolla un sistema en base a la Teoría clásica del test y las medidas que sugiere.

Pruebas y resultados

En una primera prueba, se utilizó como muestra los resultados obtenidos en el Segundo Examen Departamental de Metodología de la Programación sección 102. El grupo consta de 24 alumnos.
Para ejemplificar, se hace uso del siguiente reactivo, que formó parte del examen que se tomó de muestra:
"*Elabore un algoritmo que lea dos cadenas para determinar cuál es mayor lexicográficamente o en su caso, indique si son iguales. Considere que para determinar esto, las cadenas por principio, deberán ser de la misma longitud*."
Las respuestas, se evalúan en una escala de 0 a 100. La rúbrica se dividió en 3 partes, tal y como se empleó para calificar el reactivo, tal como se muestra en la tabla III.

Tabla III. Criterio de la rúbrica

33%	Comprueba si las cadenas son de la misma longitud.
33%	Utiliza iteraciones para comparar las cadenas.
33%	Determina cual cadena es mayor o si son iguales.

Una vez obtenidos los resultados se procedió a obtener los índices de discriminación, dificultad y el coeficiente biserial puntual, obteniéndose:

$$P_i = \frac{8}{24}$$

$$P_i = \frac{1}{3} = 0.33..$$

En base a lo explicado anteriormente tenemos que 0.33.. > 0.32, por lo que determinamos que el reactivo tiene la calidad de **DIFICIL**.

En cuanto al índice de discriminación, se obtuvieron los siguientes resultados:
ID=0

En base a lo anterior tenemos que el índice de discriminación es **POBRE**, por lo que se sugiere descartar el reactivo o revisarlo a profundidad.

Coeficiente Biserial Puntual:
Para calcular el CBP, necesitamos obtener el promedio de las calificaciones de los alumnos que respondieron correctamente el reactivo y el promedio total de calificaciones. Por lo que tenemos lo siguiente:

Promedio de los alumnos que acertaron al reactivo:

$$\overline{Y}_p = \frac{5+8+9+9+4+2+5+6}{8}$$

$$\overline{Y}_p = \frac{48}{8} = 6$$

Promedio de todas las calificaciones:

$$\overline{Y} = \frac{6+2+5+8+4+9+9+5+4+4+2+1.5+1+2+4.5+0.6+0+2+0+2.7+0+3.5+5+6}{24}$$

$$\overline{Y} = \frac{86.8}{24} = 3.6167$$

Posteriormente obtenemos la desviación estándar:

$$\delta = \sqrt{\frac{(Y_1 - \bar{Y})^2 + (Y_2 - \bar{Y})^2 + \cdots + (Y_n - \bar{Y})^2}{N-1}}$$

$$\delta = \sqrt{\frac{166.47333336}{23}}$$

$$\delta = \sqrt{7.23797}$$

$$\delta = \sqrt{7.23797} = 2.69035$$

Una vez obtenidos estos datos, los sustituimos en la fórmula para obtener el CBP:

$$r_{pb} = \frac{(6 - 3.6167)\sqrt{24 * 8/[(24-8)(8-1)]}}{2.69035}$$

$$r_{pb} = \frac{2.3833 * \sqrt{192/(16*7)}}{2.69035}$$

$$r_{pb} = \frac{2.3833 * \sqrt{1.71429}}{2.69035}$$

$$r_{pb} = \frac{2.3833 * 1.3093}{2.69035} = \frac{3.12045}{2.69035} = 1.15987$$

En el sistema se implementaron algoritmos para obtener lo anterior empleando determinadas consultas a la base de datos y los resultados se observan en la fig. 1.

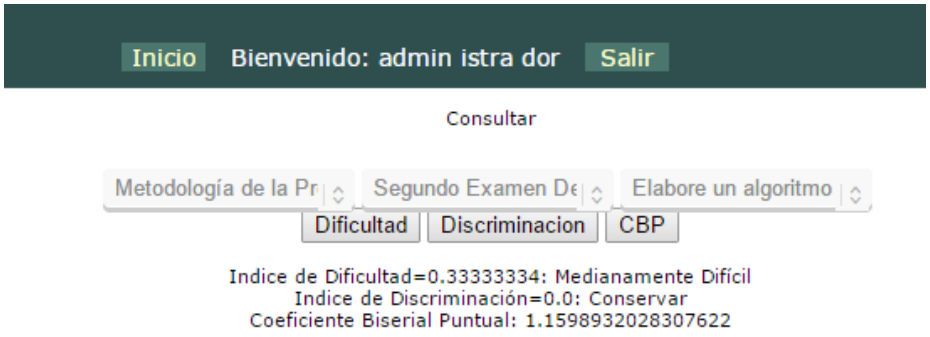

Fig. 1. Resultados obtenidos.

Conclusiones

Una vez realizando la estandarización en los reactivos, se puede apreciar que podemos evaluarlos de manera similar a los reactivos de opción múltiple, los cuales son muy utilizados por la facilidad que ofrecen al personal que aplica las pruebas. Con el uso de las rúbricas para limitar el aprendizaje esperado podemos además cuantificar los resultados, y por lo tanto, someterlos a una evaluación más controlada mediante el uso de la teoría clásica del test.

Para un campo que en su mayoría, los exámenes exigen, más que la teoría, la aplicación de los conocimientos, la aplicación de reactivos abierto o de tipo ensayo requiere que se sometan a una

evaluación similar a la utilizada con reactivos de otro tipo. Este sistema facilita realizar esta evaluación y, con los resultados obtenidos se puede examinar a detalle los elementos que se necesitan mejorar y aquellos que se pueden conservar.

El sistema fue probado en un grupo únicamente, por lo que los resultados arrojados se limitan a expresar la calidad de los reactivos en una pequeña muestra. Considerando que los exámenes son departamentales y se aplicará el mismo nivel de dificultad a una cantidad mayor de alumnos, se necesitaría comparar los resultados obtenidos en un grupo diferente o incluso implementarlo en un grupo más grande.

Mediante el uso del sistema, se pretende hacer un pilotaje para formar un banco de preguntas de alguna materia en particular. Este banco de preguntas, estará validado mediante la metodología presentada y sin que se requiera de estar realizando una evaluación y valoración manual, sirviendo para reactivos de tipo pregunta abierta o ensayo.

El sistema nos facilita el nivel de dificultad del reactivo, sin embargo, algo que sería interesante implementarle para ir más lejos, es que separe los reactivos de acuerdo a su dificultad, y además, en base al banco de reactivos, genere pruebas diferentes cuya dificultad en sus reactivos esté nivelada y, a pesar de que los exámenes sean diferentes, todos tengan la misma dificultad.

Hemos pensado que una vez el sistema pueda ser completamente utilizado para la aplicación de exámenes departamentales de una materia trabajaremos para extenderlo al grupo de materias de la misma área y después a otras materias de los diferentes programas con que cuenta nuestra Facultad.

Bibliografía

- Lineamientos Generales de la Evaluación Colegiada del Aprendizaje por Asignatura (ECAA). Documento de trabajo. Vicerrectoría de Docencia. Subdirección de Evaluación. Enero 2008.
- Gatica-Lara, F., Uribarren-Berrueta, T.. ¿Cómo elaborar una rúbrica? Pautas en educación médica. septiembre de 2012.
- Lewis R.. Test psicológicos y evaluación, Undécima edición. PEARSON EDUCACIÓN. 2003
- Bednar, A.K., Cunningham, D., Duffy, T.M., and Perry, J.D.. Theory into practice: How do we link? In G. Anglin (Ed.), Instructional Technology: Past, Present and Future. Englewood, CO: Libraries Unlimited, Inc 1991.