apertura
Volumen 12, número 1, pp. 118-131 | Universidad de Guadalajara
https://doi.org/10.32870/Ap.v12n1.1698

Evaluación estandarizada de los aprendizajes en la UABC: innovación desde el análisis psicométrico

Standardized evaluation of learning at UABC: Psycometric analysis innovation

Jorge Gustavo Gutiérrez Benítez* | Luis Alan Acuña Gamboa**

Recepción del artículo: 21/05/2019 | Aceptación para publicación: 20/11/2019 | Publicación: 30/3/2020

RESUMEN

En virtud del fuerte impacto que ha tenido la evaluación en la enseñanza, resultan indispensables instrumentos de evaluación del aprendizaje que sean válidos y confiables. La Facultad de Idiomas de la Universidad Autónoma de Baja California no implementa una herramienta tecnológica para el desarrollo de exámenes estandarizados y su análisis psicométrico, de ahí que este trabajo analice la aplicación de exámenes estandarizados referidos a un criterio y basados en la teoría de respuesta al ítem. La metodología empleada requirió la integración de varios comités de especialistas para producir instrumentos que guiaran la construcción del examen, como la tabla de índices de relevancia curricular y la justificación de contenidos. La calidad del examen se midió mediante un desarrollo innovador que cuantificó automáticamente los criterios psicométricos de validez y confiabilidad, como el índice de dificultad de los ítems y el índice de discriminación. Los resultados permiten observar la calidad del examen e identificar el rendimiento de cada ítem, así como el nivel de dominio de la muestra de alumnos participantes, lo que contribuye a la realización de valoraciones más exactas de su desempeño académico. El instrumento mostró utilidad tanto para la evaluación formativa como para la sumativa.

Palabras clave
Evaluación estandarizada; psicometría; metodología; calidad educativa; examen

Abstract

The evaluation has had a strong impact on teaching, which is why learning instruments that are valid and reliable are indispensable. The Faculty of Languages of the Universidad Autónoma de Baja California does not implement a technological tool for the development of standardized tests and their psychometric analysis. Therefore, this paper analyze the implementation of standardized tests referred to a criterion, based on the Item Response Theory. The used methodology required the integration of several committees of specialists to produce instruments that guided the construction of the test, such as the curricular relevance index table, the contents justification, etc. The test quality was evaluated using an innovating development that automatically quantified the psychometric criteria of validity and trustworthiness like the difficulty index of the items and the discrimination index. The results allow to observe the quality of the test identifying the yield of each item, as well as the yield and level of knowledge of the participant students sample, which contributes to the realization of more accurate assessments of their academic performance. It is concluded that the instrument showed utility for both formative and summative evaluation.

Keywords
Standardized evaluation; psychometric; methodology; educational quality; test

* Estudiante del Doctorado en Innovación en Tecnología Educativa, Universidad Autónoma de Querétaro. Técnico Académico de Tiempo Completo de la Universidad Autónoma de Baja California, México. ORCID: https://orcid.org/0000-0003-3392-6398 | ** Doctor en Estudios Regionales por la Universidad Autónoma de Chiapas. Docente investigador de la Universidad Autónoma de Chiapas, México. ORCID: https://orcid.org/0000-0002-8609-4786

INTRODUCCIÓN

Diferentes propósitos educativos requieren diversas pruebas y usos de estas. Cuando una prueba no guarda estrecha relación con sus propósitos, difícilmente pueden efectuarse inferencias válidas a partir de sus resultados. Desarrollar exámenes para la evaluación estandarizada del aprendizaje (Fernández, Alcaraz y Sola, 2017; Márquez, 2014) es una tarea delicada, ya que evaluar con un instrumento mal diseñado puede tener un impacto negativo tanto para el sustentante como para el docente; incluso medir aspectos del conocimiento distintos a los que se tenían pensados evaluar va en detrimento de este.

Para Hernández, Ramírez y Gamboa (2018), uno de los retos más desafiantes para las instituciones educativas consiste en la identificación, mediante procesos evaluativos, de las capacidades, conocimientos y habilidades de los estudiantes, a fin de que se adecuen los planes, los programas y los métodos educativos para mejorar el proceso de enseñanza y aprendizaje. Lo anterior hace evidente la importancia de los aparatos metodológicos en el desarrollo o la elaboración de exámenes, pues, a través de estos, se puede asegurar la calidad del instrumento, así como su aporte para la obtención de información válida y confiable.

En las instituciones de educación superior de los países industrializados es común que se utilicen evaluaciones diseñadas y validadas para propósitos de admisión. En Estados Unidos, por ejemplo, desde 1926 se emplea el Scholastic Aptitude Test

para ingresar a la licenciatura; desde 1949, el Graduate Record Examination, para el posgrado; y desde 1964, el Test of English as a Foreing Languaje, para acreditar el conocimiento del inglés (Tirado *et al.*, 1997).

En el caso específico de México, en 1994 se creó el Centro Nacional para la Evaluación de la Educación Superior (Ceneval) a fin de que se contará con un examen nacional de carácter indicativo que proporcionara información confiable y válida sobre los conocimientos y las habilidades que adquieren las personas como beneficiarias de los distintos programas educativos; el examen serviría como un indicador para las instituciones de educación superior que participaran. La concepción del Ceneval pretendía atender la necesidad del país,

> de contar con un examen nacional indicativo previo a la licenciatura [...] [que permitiera] a las propias instituciones, a las autoridades gubernamentales y a la sociedad en general, evaluar las habilidades y los conocimientos básicos que poseen los aspirantes al cursar estudios de licenciatura y un examen general de calidad profesional [...] [que hiciera posible] conocer la pertinencia e idoneidad de la formación académica de los nuevos profesionistas del país (Centro Nacional para la Educación Superior, 2017, p. 14).

Las necesidades anteriores se convirtieron en lo que ahora se conoce como Examen Nacional de Ingreso a la Educación Media Superior, Examen Nacional de Ingreso a la Educación Superior, Examen Nacional de Ingreso al Posgrado y Examen General para el Egreso de la Licenciatura, mecanismos insoslayables en materia de selección, promoción y obtención de borlas académicas en el sistema educativo mexicano.

Actualmente, el Examen de Competencias Básicas (Excoba), antes denominado Exhcoba (Góngora, Rocha y Verver, 2015; Pérez, Larrazolo y Backhoff, 2015), permite incluir elementos multimedia que enriquecen la experiencia evaluativa y ofrecen resultados automáticos. Este examen es una propuesta innovadora de evaluación sobre las competencias escolares, pues se aleja del formato de opción múltiple y se acerca a formas más "auténticas o naturales" de evaluar los aprendizajes (Ferreyra y Backhoff, 2016); sin embargo, la estructura del Excoba está alineada con el currículo nacional, por lo que evalúa competencias académicas básicas que se precisan en los planes de estudio de la educación obligatoria.

La Facultad de Idiomas de la Universidad Autónoma de Baja California (UABC) no dispone de una herramienta tecnológica que ayude a construir exámenes estandarizados y que, a su vez, puedan valorarse mediante el análisis psicométrico. Si bien existen algunos *software* que permiten efectuar estos análisis, se requieren conocimientos previos por parte del usuario; además, es necesario construir archivos fuente en formatos técnicos, lo que limita su uso para personal con experiencia en estos campos del conocimiento. Lo anterior abre la posibilidad de innovar el proceso de construcción de exámenes para la evaluación en los estudiantes, y también para la medición de su calidad desde una perspectiva psicométrica de manera más eficiente.

En los siguientes apartados se detalla, en primer lugar, el método elegido para guiar el desarrollo del examen, que incluye la descripción de los instrumentos utilizados para su construcción y evaluación. Se explica la interpretación numérica para los valores de calidad que se van a medir y lo que estos implican. En el apartado de resultados se muestra la valoración técnica del análisis psicométrico, que es la que determina la calidad y confiabilidad del examen, como lo es el índice de dificultad y discriminación. Asimismo, se presentan los datos obtenidos con tres de los principales instrumentos que conducen la elaboración del examen. En el último apartado se concluye la relevancia de contar con instrumentos estandarizados de evaluación y las ventajas de tener *software* automatizados para la aplicación, la construcción y el análisis psicométrico de los exámenes.

DISEÑO

Al tomar como referencia el objetivo general de esta investigación (diseñar una propuesta innovadora de análisis psicométrico para los exámenes estandarizados de los aprendizajes en la Facultad de Idiomas de la UABC), la metodología empleada en la estructuración de este examen fue formulada en el Instituto de Investigación y Desarrollo Educativo (Contreras, 2000; Contreras y Backhoff, 2004; Contreras, Encinas y De las Fuentes, 2005) con base en el modelo psicométrico que propuso originalmente Nitko (1994) para elaborar exámenes de gran escala de referencia criterial, orientados por el currículo.

El desarrollo de estos exámenes requiere la definición de ciertas etapas propias de la metodología para elaborarlos de manera estandarizada y con calidad. Además, para las últimas dos etapas del proceso, que implican el análisis del comportamiento de los ítems, es necesario un análisis estadístico especializado que precisa un *software* diseñado ex profeso para efectuar estos cálculos.

Un elemento importante en la metodología es la integración de comités que se encarguen de la ejecución de las actividades para cada fase del desarrollo del examen. El comité diseñador prepara el diseño general de la prueba; entre sus acciones se puede citar el análisis del currículo, la elaboración de la retícula y el análisis de los índices de relevancia curricular. Otro comité es el responsable de las especificaciones para la construcción de los ítems y de detallar la estructura estándar que estos deben mostrar.

El comité de ítems tiene la encomienda de producirlos de acuerdo con las características señaladas en las especificaciones. El comité de jueceo realiza una valoración de diferentes aspectos, como la calidad técnica del ítem, la validez cognitiva, el formato y la edición, entre otros. El último comité es el analítico informático, que lleva a cabo los procesos de evaluación psicométrica de los ítems mediante el procesamiento de la información con un *software* especializado.

Otro detalle importante sobre la integración de estos comités es que se conforman por un docente especialista en la asignatura, un docente

La asignatura para la cual se diseñó el examen fue Morfología de la segunda lengua, que está seriada con Morfosintaxis, ambas ponderadas con la más alta relevancia curricular para el tronco común

experto en el currículo, un docente con años de experiencia en educación y uno con amplia práctica en evaluación. Estos comités, a su vez, pueden preparar o capacitar a otros docentes para que participen en las actividades que se desempeñan.

Con base en lo anterior, la muestra a la que se aplicó este examen se integró de 394 alumnos, quienes cursan el primer semestre del tronco común de la Licenciatura en Idiomas. La asignatura para la cual se diseñó el examen fue Morfología de la segunda lengua, que está seriada con Morfosintaxis, ambas ponderadas con la más alta relevancia curricular para el tronco común. El género de la población fue 50% masculino y 50% femenino; de estos, se eligió 33% de alumnos con bajo rendimiento (menores o iguales a 6.9), 33% de alumnos con rendimiento regular (entre 7 y 8.9) y 34% de alumnos con alto rendimiento (mayores o iguales a 9).

Las variables que se midieron para valorar la calidad del examen son las establecidas por la teoría de respuesta al ítem (Gómez, 2015; Mola *et al.*, 2013; Muñiz, 2010; Mayaute y Vázquez, 2010; Lord y Novick, citado en Kramp, 2008), la cual plantea un modelo y un conjunto de criterios que, de ser alcanzados, garantizan la precisión en la medida del dominio de los estudiantes sobre un determinado aprendizaje. Estos criterios son el índice de dificultad del ítem (*p*), el índice de discriminación, la correlación ítem-total y la relación opción-total de los distractores.

El índice de dificultad se mide en un rango de valores que va de 0 a 1; entre más cerca se encuentre el valor del índice a 1, la pregunta es más fácil de contestar, y viceversa. Sucede lo mismo con el índice de discriminación del ítem, ya que se mide entre 0 y 1, y la discriminación será mejor entre más cercano a 1 se encuentre este valor. La correlación ítem-total como mínimo debe situarse en un valor de .25, y entre más cerca esté al 1, el ítem guardará mayor relación con la probabilidad más alta de obtener un buen puntaje en todo el examen si este es contestado correctamente.

Instrumentos

El sustento rector en la construcción del examen es la carta descriptiva de la asignatura, documento que establece el universo de conocimientos a evaluar. Por esto, a la prueba se le conoce como examen referido a un criterio (Leyva, 2011); por criterio se entiende específicamente los conocimientos a los que se enfoca el examen. Algunos de los instrumentos empleados en el desarrollo son producto de la ejecución de etapas propias de la metodología para elaborar exámenes estandarizados, por ejemplo, la retícula, la justificación de contenido y las especificaciones de la prueba.

Uno de los instrumentos especializados indispensable para las etapas finales del desarrollo son aquellos *software* que permiten efectuar el análisis psicométrico. Los utilizados fueron ITEMAN (Cechova, Neubauer y Sedlacik, 2014; Thoe, Fook & Thah, 2008) de la compañía Assessment Systems Corporation, así como TAP, publicado por Brooks y Johanson (2003). También, se recurrió a un *software* nuevo llamado SIEXAES, que permite integrar en un mismo sistema el espacio para la construcción digital del examen y la funcionalidad de generar de modo automático el análisis psicométrico.

RESULTADOS

Uno de los productos que arroja la aplicación de la metodología y que tiene amplia utilidad para los docentes es la retícula. Este instrumento permitió apreciar de manera completa cómo los contenidos temáticos de la asignatura se relacionaban entre sí, e identificar aquellos que tenían un mayor impacto para concretar otros conocimientos.

La figura 1 ejemplifica un pequeño segmento de la retícula, en la que se puede observar mediante las flechas cómo los contenidos temáticos de la primera unidad guardan relación con los de la unidad 2, y cómo algunos de estos tienen un número más amplio de relaciones con temas de unidades anteriores o posteriores. Este tipo de contenidos originan un mayor impacto en el aprendizaje del estudiante.

El producto correspondiente a los resultados del índice de relevancia curricular (IRC) de los contenidos temáticos se muestra en la tabla 1, mientras que la jerarquización de estos, a partir del mayor IRC, se reporta en la tabla 2, la cual presenta los contenidos temáticos, así como las prácticas que debían realizarse en la asignatura.

La tabla 1 indica algunos de los criterios que se valoraron para obtener el IRC. Cada uno de estos atributos para el universo temático de la asignatura fue evaluado y se observó cómo ciertos temas registraban una relevancia tan baja que, si se omitían, prácticamente no habría ninguna afectación en el aprendizaje del estudiante. Esto resulta interesante si se centra la atención en el

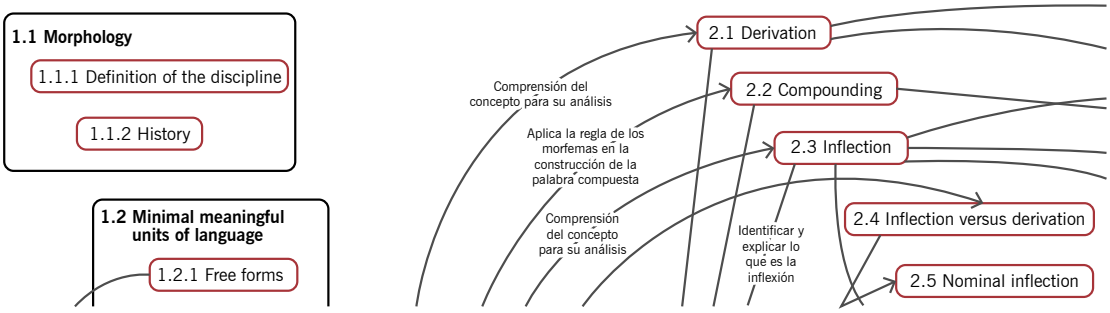


Figura 1. Ejemplo de la retícula de la materia de Morfología de la segunda lengua.
Fuente: elaboración propia.

Tabla 1. Atributos valorados para obtener el índice de relevancia curricular

Contenido	Contribución al logro de la competencia de la unidad 20%	Dosificación (cuantía de contenidos implícitos) 10%	Carga horaria (asignada o estimada para su aprendizaje) 10%	Relevancia disciplinaria 20%	Índice de relevancia curricular
1.1.1 Definition of the discipline	0.20	0.10	0.10	0.20	0.600
1.1.2 History	0.13	0.07	0.07	0.13	0.400
1.2.1 Free forms	0.20	0.10	0.07	0.13	0.533
1.2.2 Signs and morphemes	0.20	0.10	0.10	0.20	0.633

Fuente: elaboración propia.

diseño de los currículos de la asignatura, cuyo contenido temático, según los resultados mencionados, contribuye poco a lo que el estudiante debe conocer y saber hacer.

Debido a que el propósito de estos exámenes es evaluar el dominio de un determinado conocimiento que los estudiantes poseen, se utiliza una cantidad más grande de ítems de contenidos temáticos que posean un mayor IRC, porque este tipo de temas requieren la integración de una mayor cantidad de conocimientos y, por ende, supone un grado de dominio más alto. Estos fueron algunos de los resultados obtenidos en las primeras etapas de la metodología; sin embargo, para las siguientes etapas los resultados fueron más técnicos, ya que implicaron la valoración de mayor peso en el desarrollo del examen y la medición de la calidad a través del análisis psicométrico.

El análisis psicométrico que se derivó de la aplicación del pilotaje mostró en el ítem los diferentes criterios de calidad que se deseaban medir. En total, 63 ítems se aplicaron para el examen de Morfología de la segunda lengua; para cada uno se emitió un juicio a fin de determinar si se aceptaban o no, y las razones de este dictamen. Las estadísticas descriptivas generales del examen se presentan en la tabla 2.

Entre los datos importantes de la tabla 2 se encuentra el valor medio de *p* o el índice de dificultad del examen; este valor se ubicó en .638, lo que indica que el examen es de dificultad media con una leve tendencia a ser más fácil que regular. También destaca que, en promedio, se aciertan 40 de las 63 preguntas y el logro de una puntuación cercana a la perfecta al alcanzar 61 aciertos.

Ahora bien, la tabla 3 detalla los resultados de algunos ítems. Se observa, por ejemplo, el ítem 25, cuya dificultad es muy baja, es decir, es muy fácil de responder; lo anterior se concluye porque el valor *p* del ítem es de .789. Como se mencionó en el apartado anterior, entre más cercano a 1 se encuentre este valor, más fácil de responder es el ítem o la pregunta.

Tabla 2. Estadísticas descriptivas del examen

Ítems	Media de aciertos	Desviación estándar	Puntaje mínimo	Puntaje máximo	Media del valor p	Media de rpBis	Alpha
63	40.211	7.242	27	61	0.638	0.212	0.796

Fuente: elaboración propia.

Tabla 3. Análisis psicométrico de los ítems

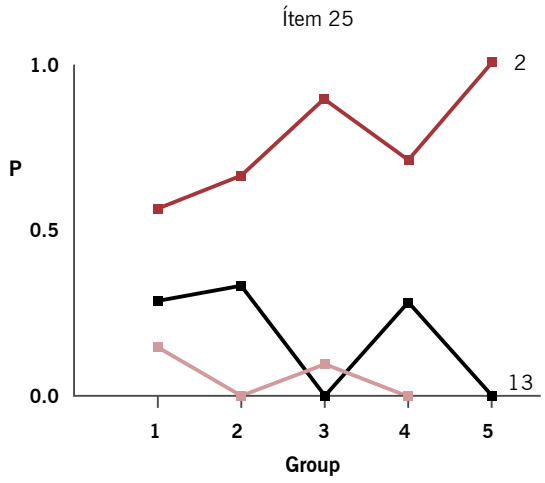
Análisis	IT25	IT26	IT28	IT48
Dificultad (*p*)	Ítem fácil con p=.789	Dificultad difícil con .395	Ítem muy difícil, con una *p*=.289	Ítem muy difícil con .21
Discriminación	Discriminación regular con .26	Discriminación muy buena .53	Discriminación buena con .33	Discriminación negativa con .16
rpBis	Buena con .264	Muy buena con .432	Baja con .132	Negativa con .271
Distractor	Los distractores funcionaron bien; se obtuvieron diferencias de -.173 y -.091.	Los distractores funcionaron bien; se obtuvieron diferencias de -.173 y -.355.	Los distractores funcionaron bien, se obtuvieron diferencias de -.064 y -.264.	Los distractores fueron elegidos, pero en igual proporción por el grupo alto y bajo

Fuente: elaboración propia.

El poder discriminatorio del ítem es regular, con un valor obtenido de .26; además, el contestarlo correctamente tiende a guardar relación con los buenos puntajes en todo el examen (rpBis de .264).

Otro factor que se observa en el ítem 25 son los distractores o las respuestas incorrectas. Estos fueron elegidos en todos los casos y con mayor proporción por los grupos bajos (alumnos con escaso rendimiento, promedios menores o iguales a 6.9) o los que obtuvieron menor nota al final del examen. Para saber si un ítem discrimina bien, o no, los valores alcanzados de rpBis para la respuesta correcta deben ser mayores de .2; las respuestas incorrectas o distractores deben ser negativos, en caso de ser positivos, deben contar con valores menores que los de la respuesta correcta. Aunado a esto, también se debe considerar que todos los distractores tienen que ser elegidos.

El ítem 25 es el ideal que se busca en un examen, ya que cumple correctamente con los estándares de calidad para todos los criterios de valoración mencionados. La gráfica ilustra el comportamiento del ítem.



Gráfica. Comportamiento del ítem 25.
Fuente: elaboración propia.

Los resultados mostrados se obtuvieron a partir del análisis psicométrico realizado con el *software* SIEXAES, el cual generó automáticamente los reportes técnicos con los estadísticos descriptivos del examen en general

En la gráfica, las tres respuestas posibles se representan con una línea de color rojo, una negra y una rosa. El eje *x* u horizontal simboliza a los grupos bajo, medio y alto, mientras que el *y* o vertical, a la población. La línea roja corresponde a la respuesta correcta y, como vemos en la gráfica, esta fue elegida en mayor proporción por el grupo alto, que estaría situado entre las categorías 4 y 5; su elección fue menor por el grupo bajo, que se ubica entre las categorías 1 y 2. La pregunta presentó una dificultad muy baja, 30 de las 38 personas lo acertaron, pero aun así el ítem mantuvo una mayor proporción de aciertos para el grupo alto.

Los resultados mostrados se obtuvieron a partir del análisis psicométrico realizado con el *software* SIEXAES, el cual generó automáticamente los reportes técnicos con los estadísticos descriptivos del examen en general –como ya se mostró–, y los criterios de confiabilidad y validez que se establecieron como indicadores de calidad, el índice de dificultad de cada ítem, el poder discriminatorio, la correlación ítem-total y el funcionamiento de cada distractor. Las gráficas se elaboraron con base en el análisis efectuado con el *software* ITEMAN.

DISCUSIONES

Después de la aplicación de la metodología para la construcción del examen estandarizado y la realización del análisis psicométrico, se identificó con claridad cómo ciertos reactivos en el examen presentaban indicadores de calidad aceptables, mientras que otros aún necesitaban ser corregidos. El proceso de diseñar y construir el examen aportó información suficiente para determinar si el instrumento era válido y confiable, pero no solo mediante el análisis del examen, sino también la herramienta tecnológica utilizada para su construcción, aplicación y posterior análisis. Este último rubro permitirá, en breve, dar respuesta al objetivo principal de la investigación: el diseño de una propuesta tecnológica para el análisis psicométrico de exámenes estandarizados.

La intención de la herramienta propuesta es eliminar las dificultades que representa para un docente llevar a cabo un análisis psicométrico de un examen, porque el *software* que comúnmente se emplea con estos fines requiere de la experiencia por parte de los usuarios, tanto en su manejo como en la interpretación de los resultados, ya que los reportes que se generan son técnicos.

La integración de los distintos comités que participaron en el diseño del examen resultó un gran aporte para el trabajo, pues en estos se sustentó, con mayor precisión, cada uno de los elementos de la propuesta aquí presentada

De esta manera, después de aplicar cada una de las fases propuestas en la metodología, se obtuvo una serie de productos que permitieron conocer mejor el universo de conocimientos que se tenían que evaluar en el examen, así como identificar, de una manera gráfica y cuantificable, la importancia de cada uno de los contenidos temáticos de la asignatura específica.

Por otro lado, la integración de los distintos comités que participaron en el diseño del examen fue un gran aporte para el trabajo, pues en estos se sustentó, con mayor precisión, cada uno de los elementos de la propuesta aquí presentada. Los comités se constituyeron a partir de la experiencia de los docentes participantes en los campos del diseño general de la prueba, las especificaciones, los ítems, el jueceo y el aparato analítico informático, con lo que se obtuvieron juicios más objetivos del examen que enriquecieron las etapas de justificación de contenidos, la elaboración de especificaciones de los ítems, la evaluación de estos, entre otras etapas.

Aún existen elementos que deben ser mejorados, por ejemplo, se advirtió que de los 63 ítems que integran el examen, en doce existen deficiencias; de estos, ocho corresponden al índice de discriminación, por lo que hay que analizar las respuestas que esos ítems consideran. De igual forma, todavía deben producirse más ítems para el examen a fin de generar diferentes versiones de este.

También es importante dar seguimiento a las observaciones realizadas durante la elaboración de la tabla con los índices de relevancia curricular y la retícula sobre aquellos contenidos temáticos que aportan poco al aprendizaje de los estudiantes, ya

que este tipo de situaciones repercuten de modo directo en la calidad del diseño del currículo de la asignatura. Asimismo, se debe dar continuidad a las sugerencias y valoraciones que los alumnos y maestros participantes formularon sobre el *software* SIEXAES para mejorar la calidad de la herramienta y cumplir de forma más eficiente con los criterios de usabilidad evaluados, como el tamaño de los objetos, los colores, los botones de navegación, entre otros elementos.

Al retomar el aspecto educativo de este trabajo, se optó por utilizar exámenes referidos a un criterio, ya que la intención principal del examen es explorar el nivel de dominio del estudiante sobre un universo de conocimientos claramente delimitado, es decir, medir las habilidades y los conocimientos que este domina, o no, en un campo específico del conocimiento. Como se mencionó, el criterio para guiar el desarrollo del examen es el currículo de la asignatura en cuestión, el cual especifica los contenidos temáticos a evaluar y su secuencia. Todas las decisiones de lo que se evaluará y la forma en que se hará están orientadas por lo establecido en el currículo. Por lo anterior, se tiene la ventaja de describir con precisión y claridad lo que se intenta medir.

A través de este tipo de exámenes, se obtienen puntuaciones individuales asociadas solo al desempeño del estudiante, y no en función del desempeño del grupo al que el estudiante pertenece, como supone la evaluación normativa. Además, las puntuaciones particulares de cada alumno ofrecen la posibilidad de diseñar estrategias individuales de mejora, situación que resulta complicada con los exámenes normativos.

El punto principal de la evaluación criterial es la representatividad del elemento respecto al universo de medida, entendida en el aspecto educativo como congruencia entre el ítem y el objetivo; así, el estudiante puede tener la certeza de que lo que se le evalúa es en realidad lo que aprendió y no otro aspecto del conocimiento ajeno a su aprendizaje, mientras que para el docente significa la seguridad de que el puntaje obtenido en el examen refleja exactamente el aprendizaje del alumno en relación con lo enseñado en clases y conforme a los contenidos temáticos determinados en el currículo.

En otro momento, fue necesario para la investigación recuperar la opinión de los docentes y alumnos participantes en el pilotaje del examen, para lo cual se aplicó una encuesta de satisfacción a una muestra con base en el criterio de la conveniencia; es decir, que de manera voluntaria quisieran emitir sus comentarios y percepciones acerca de este examen. Se obtuvieron estos comentarios:

> Estudiante 1: Me gustó el formato de opción múltiple, facilita responder el examen.
>
> Estudiante 2: Las preguntas fueron claras y sin tantos rodeos, como esas que son capciosas.

Las puntuaciones particulares de cada alumno ofrecen la posibilidad de diseñar estrategias individuales de mejora, situación que resulta complicada con los exámenes normativos

"La puntuación final del examen me hace saber si en general enseñé bien o no lo que se supone que se debe ver en la materia, pensando en que la clase se apega a lo que dice la carta descriptiva"

Estudiante 3: Siento que el examen realmente puso a prueba lo que sabía de la materia, creo que abarcó todos los temas vistos.

Docente 1: Un examen que sigue un método cuantificable para determinar si una pregunta en un examen está bien hecha o no, me da la tranquilidad de saber que si un alumno mío responde correctamente a dicha pregunta es porque realmente sabe lo que se le está preguntando y, de cierta manera, me dice que mis estrategias de aprendizaje han ayudado a que él pueda contestar correctamente.

Docente 2: La puntuación final del examen me hace saber si en general enseñé bien o no lo que se supone que se debe ver en la materia, pensando en que la clase se apega a lo que dice la carta descriptiva de la materia.

Docente 3: En un principio era escéptico de la calidad del examen, pero después de observar que cada tema se evaluó de acuerdo a la experiencia de más de un docente, que se justificó cada pregunta, que no fueron los mismos maestros los que hicieron todo el examen sino que participaron diferentes comités durante todo el proceso, y que después de eso se aplicaron cálculos para demostrar lo que debía venir y lo que no en el examen, me hizo darme cuenta que la calificación que un alumno obtuviera en el examen en realidad era la que merecía tener, que ese número significaba realmente lo que sabía de la materia.

Con los comentarios anteriores queda manifiesta la importancia de contar con exámenes estandarizados. Sin importar si existe "n" cantidad de versiones de un examen, cada una de estas evalúan de igual forma el aprendizaje del estudiante; no hay exámenes más difíciles o más fáciles. Así, el docente y el alumno tienen la seguridad de que el instrumento es válido y confiable.

CONCLUSIONES

En esta primera aproximación, el *software* arrojó datos con los cuales se pueden realizar inferencias importantes; por ejemplo, una de ellas se relaciona con la cantidad de aciertos promedio en el examen. En este sentido, el grupo 2 obtuvo una mejor cantidad de aciertos en promedio, al superar al grupo 1 por casi diez puntos. Sin embargo, en la última unidad de la asignatura, el grupo 1 alcanzó un mejor promedio (6.37) en relación con el 2 (5.53).

El caso anterior es el tipo de situaciones que las áreas de coordinación académica necesitan identificar para tomar acciones tanto preventivas como correctivas en el proceso de aprendizaje de los estudiantes. La intención de recabar estos datos busca incentivar el proceso de análisis para conocer las causas de estas variaciones en los resultados, que pueden indicar situaciones como absentismo del docente durante ese período, algunas estrategias de aprendizaje mal empleadas, entre otros factores.

Un ejemplo de esas acciones o decisiones por parte de estas áreas de coordinación pueden ser los cursos de capacitación para los docentes, observaciones o sanciones por malas prácticas o, por el contrario, reconocimientos por buen desempeño. Con los reportes técnicos y especializados que el *software* SIEXAES genera, el docente puede identificar, en cuanto al contenido temático, las deficiencias de sus estudiantes o aspectos del conocimiento en concreto que no dominan y, en consecuencia, incentiva la mejora y el desarrollo de sus competencias docentes. De igual forma, el estudiante puede observar su rendimiento con mayor detalle, identificar cuales áreas domina y cuáles no y, sobre todo, tener la seguridad de que el instrumento con el que es evaluado es confiable, que realmente se le evalúan los dominios de un conocimiento que están relacionados de manera directa con la asignatura.

Es preciso mencionar que una educación de calidad no será objeto de medición únicamente por los tipos de instrumentos de evaluación que se utilicen, ya que implica otros procesos que, de igual modo, deberían someterse a valoraciones. El examen constituye por sí solo una muestra de tareas evaluativas representativas del dominio de un contenido en específico. En este sentido, se habla, por ejemplo, de las estrategias de aprendizaje que se emplean en el curso, la pertinencia de los contenidos temáticos que se estudian, los niveles cognitivos que se alcanzan, las estrategias de enseñanza del docente, entre muchos otros elementos que, a manera de sinergia, puedan resultar en una educación de calidad.

En el aprendizaje basado en el enfoque constructivista, por ejemplo, se le da mayor importancia al proceso de enseñanza y aprendizaje que a los propios contenidos; esto implica, entonces, que se valore más lo que un estudiante debe hacer que lo que debe saber. Lo anterior puede ser la causa de que algunos contenidos temáticos de la asignatura resultaran con índices de relevancia curricular tan bajos, temas que, como se mencionó pueden ser omitidos y no tendrían mayor impacto sobre el aprendizaje del estudiante. Puede ser el caso que, al diseñar el contenido temático de la asignatura, se priorizó lo que el estudiante debe ser capaz de hacer en relación con lo que debe saber.

Al considerar la evaluación desde el enfoque cognitivista, se requiere que todos los instrumentos que se vayan a utilizar tengan como fin objetivos cognitivos, lo que implica que los mismos ítems o reactivos de un examen reflejen la relación con estos objetivos. Los resultados obtenidos en el análisis técnico mostraron aspectos relacionados con este enfoque; por ejemplo, la diferencia del dominio logrado en la unidad tres entre los dos grupos fue notoria, si se parte de la suposición de que los estudiantes de ambos grupos tuvieron la misma dosificación de horas para aprender esos temas. Una primera aproximación a identificar la causa de esa diferencia puede ser el tipo de estrategias de aprendizaje empleadas por los docentes, o bien, la irregular o inadecuada dosificación de horas, que, de nuevo, expone un posible mal diseño en el currículo.

La investigación, al considerar estos primeros resultados y a pesar de que aún falta calibrar algunos aspectos del examen, ofreció información relevante: identificó el nivel de dominio que poseen de la asignatura en cuestión, reveló que para un contenido temático en particular existió una

El estudiante puede observar su rendimiento con mayor detalle, identificar cuales áreas domina y cuáles no y, sobre todo, tener la seguridad de que el instrumento con el que es evaluado es confiable

diferencia marcada entre un grupo y otro, y señaló posibles deficiencias en el currículo de la asignatura por incluir temas que no tenían relevancia curricular o secuencias temáticas mal empleadas.

El examen, además de aportar a la evaluación formativa de los estudiantes, también lo hace para la evaluación sumativa; por ejemplo, puede ayudar a predecir el éxito de un estudiante en cursos posteriores referidos a la asignatura, sobre todo cuando obtiene niveles de dominio altos en contenidos temáticos que están conectados con aprendizajes posteriores y, de cierta manera, la garantía del estudiante de que, al realizar su examen, posee un determinado dominio.

Es un hecho conocido que para una institución educativa es importante obtener información del rendimiento que tienen sus estudiantes, estimar el aprendizaje logrado y comparar esos logros con las metas establecidas. Por esta razón, es indispensable contar con instrumentos de evaluación del aprendizaje que sean válidos y confiables; es decir, que estén correctamente diseñados y que ofrezcan una seguridad sobre lo evaluado en términos de conocimientos y habilidades para los cuales se planteó.

REFERENCIAS BIBLIOGRÁFICAS

Brooks, Gordon & Johanson, George. (2003). TAP: Test Analysis Program. *Applied Psychological Measurement*, *27*(4), 303-304. https://doi.org/10.1177/0146621603027004007.

Cechova, Ivana; Neubauer, Jiri & Sedlacik, Marek. (2014). Computer-adaptive testing: Item analysis and statistics for effective testing, in European Conference on e-Learning. Academic Conferences International Limited.

Centro Nacional de Evaluación para la Educación Superior. (2017). Origen y evolución del Ceneval. Recuperado de: http://www.ceneval.edu.mx/documents/20182/49855/OrigenEvolucionCeneval_2018/f8406659-7d28-4960-9ec1-964d90c76e4c

Contreras Niño, Luis Ángel. (2000). Desarrollo y pilotaje de un examen de español para la educación primaria en Baja California (tesis de maestría). Instituto de Investigación y Desarrollo Educativo: Universidad Autónoma de Baja California, Campus Ensenada. Recuperado de: http://iide.ens.uabc.mx/images/pdf/tesis/MCE/Tesis%20MCE%20Luis-Angel-Contreras-Nino.pdf

Contreras Niño, Luis Ángel y Backhoff Escudero, Eduardo. (2004). Metodología para elaborar exámenes criteriales alineados al currículo, en Sandra Castañeda (ed.), *Educación aprendizaje y cognición, teoría en la práctica*. México: Manual Moderno.

Contreras Niño, Luis Ángel; Encinas Bringas, José Álvaro y De las Fuentes Lara, Maximiliano. (2005). Evaluación colegiada del aprendizaje en la Universidad Autónoma de Baja California: el caso del examen de Matemáticas I de la Facultad de Ingeniería Mexicali, en *Memoria del VIII Congreso Nacional de Investigación Educativa*. México: COMIE.

Fernández Navas, Manuel; Alcaraz Salarirche, Noelia y Sola Fernández, Miguel. (2017). Evaluación y pruebas estandarizadas: una reflexión sobre el sentido, utilidad y efectos de estas pruebas en el campo educativo. *Revista Iberoamericana de Evaluación Educativa, 10*(1), 51-67. https://doi.org/10.15366/riee2017.10.1.003

Ferreyra Martínez, María Fabiana y Backhoff Escudero, Eduardo. (2016). Validez del generador automático de ítems del examen de competencias básicas (Excoba). *Relieve-Revista Electrónica de Investigación y Evaluación Educativa*, *22*(1), 1-16. https://doi.org/10.7203/relieve.22.1.8048

Gómez Rada, Carlos Alberto. (2015). Diseño, construcción y validación de un instrumento que evalúa clima organizacional en empresas colombianas, desde la teoría de respuesta al ítem. *Acta Colombiana de Psicología*, (11), 97-113. Recuperado de: https://editorial.ucatolica.edu.co/index.php/acta-colombiana-psicologia/article/view/482

Góngora Ortega, Javier; Rocha Hernández, Tania Marlene; Verver, Ingrid Verenice. (2015). La prueba Exhcoba como predictora para la deserción y reprobación en medicina. *Revista de la Escuela de Medicina "Dr. José Sierra Flores" Universidad del Noreste*, *29*(1), 16-24. Recuperado de: http://www.une.edu.mx/Resources/RevistaMedicina/2015/Vol29No1.pdf#page=16

Hernández Madrigal, Mónica; Ramírez Flores, Élfego y Gamboa Cerda, Silvia. (2018). La implementación de una evaluación estandarizada en una institución de educación superior. *Innovación Educativa*, *18*(76), 149-170. Recuperado de: http://www.scielo.org.mx/scielo.php?script=sci_arttext&pid=S1665-26732018000100149&lng=es&tlng=es

Kramp Denegri, Uwe. (2008). Equivalencia entre los modelos de análisis factorial de los ítems y teoría de respuesta a los ítems en la

evaluación de las propiedades psicométricas de los instrumentos de medición psicológica. *Revista Peruana de Psicometría*, *1*(1). Recuperado de: https://www.academia.edu/3673226/Equivalencia_entre_los_modelos_de_an%C3%A1lisis_factorial_de_los_%C3%ADtems_y_teor%C3%ADa_de_respuesta_a_los_%C3%ADtems_en_la_Evaluaci%C3%B3n_de_las_propiedades_Psicom%C3%A9tricas_de_los_instrumentos_de_Medici%C3%B3n_psicol%C3%B3gica

Leyva Barajas, Yolanda Edith. (2011). Una reseña sobre la validez de constructo de pruebas referidas a criterio. *Perfiles Educativos*, *33*(131), 131-154. Recuperado de: http://www.scielo.org.mx/scielo.php?script=sci_arttext&pid=S0185-26982011000100009

Márquez Jiménez, Alejandro. (2014). Las pruebas estandarizadas en entredicho. *Perfiles Educativos*, *36*(144), 3-9. Recuperado de: http://www.redalyc.org/articulo.oa?id=13230751001

Mayaute Escurra, Luis Miguel y Vásquez Delgado, Ana Esther. (2010). Análisis psicométrico del test de matrices progresivas avanzadas de Raven mediante el modelo de tres parámetros de la teoría de la respuesta al ítem. *Persona*, (13), 71-97. Recuperado de: http://www.redalyc.org/articulo.oa?id=147118212004

Mola, Débora Jeanette; Saavedra, Bianca Analía; Reyna, Cecilia y Belaus, Anabel. (2013). Valoración psicométrica de la Psychological Entitlement Scale desde la teoría clásica de los tests y la teoría de respuesta al ítem. *Pensamiento Psicológico*, *11*(2), 19-38. Recuperado de: https://ri.conicet.gov.ar/bitstream/handle/11336/23949/CONICET_Digital_Nro.4290b666-0f89-4d7f-8738-01898530318f_A.pdf?sequence=2

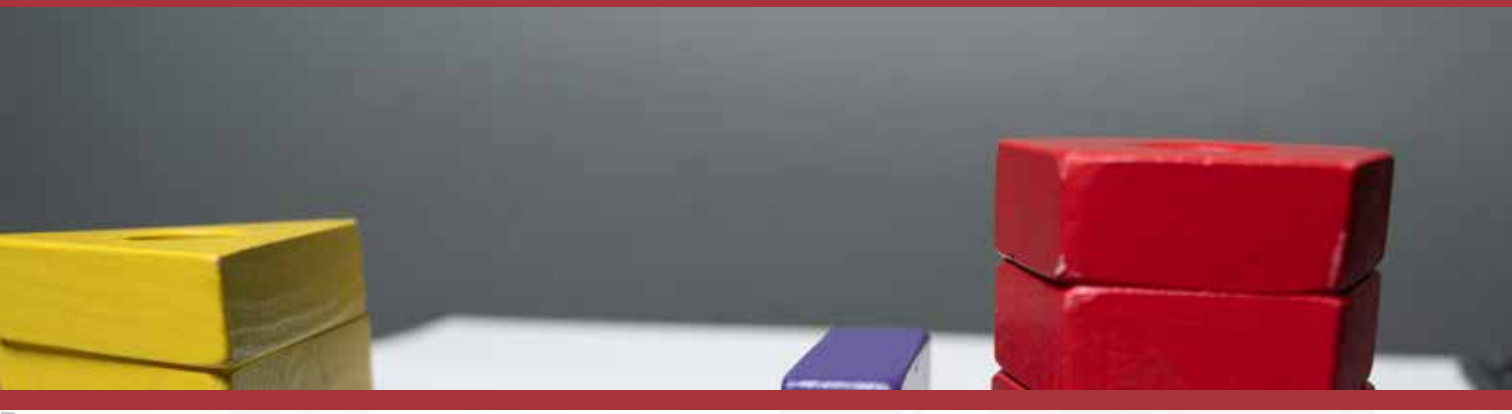
Muñiz Fernández, José. (2010). Las teorías de los tests: teoría clásica y teoría de respuesta a los ítems. *Papeles del Psicólogo*, *31*(1), 57-66. Recuperado de: http://www.redalyc.org/articulo.oa?id=77812441006

Nitko, Anthony. (1994). A model for development curriculum-driven criterion- referenced and norm-referenced examination for certification and selection of students. Paper presented in the Conference of Education, Evaluation and Assessment for the Association Studies of Educational Evaluation in Sudafrica (ASEESA). Sudáfrica. Recuperado de: https://eric.ed.gov/?id=ED377200

Pérez Morán, Juan Carlos; Larrazolo Reyna, Norma y Backhoff Escudero, Eduardo. (2015). Análisis de la estructura cognitiva del área de habilidades cuantitativas del Exhcoba mediante el modelo LLTM de Fisher. *Revista Internacional de Educación y Aprendizaje*, *3*(1), 25-38. Recuperado de: https://journals.epistemopolis.org/index.php/educacion/article/view/584

Tirado Segura, Felipe; Backhoff Escudero, Eduardo; Larrazolo Reyna, Norma y Rosas Morales, Martín. (1997). Validez predictiva del Examen de Habilidades y Conocimientos Básicos (Excoba). *Revista Mexicana de Investigación Educativa*, *2*(3), 67-84. Recuperado de: http://www.comie.org.mx/revista/v2018/rmie/index.php/nrmie/article/download/1057/1057

Thoe, Ng Khar; Fook, Fong Soon & Thah, Soon Seng. (2009). Use of ICT tool for Item Analysis of a Science Performance Test. *Journal of Educational Technology*, *9*(1), 5-15. Recuperado de: http://www.mjet-meta.com/resources/V9N1%20-%20NKT%20-%202009%20-%20ICT%20-%20Online.pdf



CÓMO CITAR ESTE ARTÍCULO:

Gutiérrez Benítez, Jorge Gustavo y Acuña Gamboa, Luis Alan. (2020). Evaluación estandarizada de los aprendizajes en la UABC: innovación desde el análisis psicométrico. *Apertura*, *12*(1), pp. 118-131. http://dx.doi.org/10.32870/Ap.v12n1.1698