[Cierre de edición el 31 de diciembre del 2023]

https://doi.org/10.15359/ree.27-3.17218
https://www.revistas.una.ac.cr/index.php/educare
educare@una.ac.cr

Evaluación estandarizada de los aprendizajes en el tronco común de lenguas: Un estudio en la Universidad Autónoma de Baja California, México

Standardized Learning Assessment in Languages Core Curricula: A Study at the Autonomous University of Baja California, México

Avaliação padronizada da aprendizagem na tronco comum de línguas: um estudo na Universidade Autônoma da Baixa Califórnia, México

Jorge Gustavo Gutiérrez-Benítez
Universidad Autónoma de Baja California
Mexicali, México
gutierrez.jorge@uabc.edu.mx
https://orcid.org/0000-0003-3392-6398

Luis Alan Acuña-Gamboa
Universidad Autónoma de Chiapas
Tuxtla Gutiérrez, México
luis.gamboa@unach.mx
https://orcid.org/0000-0002-8609-4786

Recibido • Received • Recebido: 03 / 07 / 2022
Corregido • Revised • Revisado: 06 / 09 / 2023
Aceptado • Accepted • Aprovado: 15 / 11 / 2023

Resumen:
Objetivo. El estudio presenta la estructura y composición de un examen departamental estandarizado de una de las asignaturas de mayor relevancia curricular para el tronco común de Lenguas perteneciente a la Facultad de Idiomas(FI), así como los resultados de la evaluación psicométrica realizada a este mismo. **Método.** Se empleó la metodología formulada por el Instituto de Investigación y Desarrollo Educativo (Contreras Niño, 2000; Contreras Niño & Backhoff Escudero, 2004), con base en el modelo psicométrico propuesto por Nitko (1994) para elaboración de exámenes de gran escala de referencia criterial. La muestra de estudio fue de 260 estudiantes de primer semestre, procedentes de la FI a nivel Estado. Se empleó el método de análisis psicométrico basado en la teoría de respuesta al ítem, con lo que se midieron atributos como el índice de discriminación, coeficiente de discriminación y el índice de dificultad. **Resultados.** A partir del análisis psicométrico realizado al examen, se obtuvieron valores promedio que sobrepasan el estándar en cuanto al índice y coeficiente de discriminación (.32), destacan los obtenidos en la unidad temática de mayor relevancia curricular para la asignatura en cuestión. Asimismo, se observó una dificultad media en general en el examen y una distribución de la dificultad de los reactivos aceptable. **Conclusiones.** Los valores obtenidos en los atributos de confiabilidad y validez permitieron evidenciar la calidad

https://doi.org/10.15359/ree.27-3.17218
https://www.revistas.una.ac.cr/index.php/educare
educare@una.ac.cr

del instrumento, lo cual contribuye a la certeza del nivel de dominio que el estudiantado refleja en relación con el universo de conocimientos representados en la prueba, situación que tiene un alto impacto, por la relevancia de dicha asignatura.

Palabras claves: Segunda lengua; estandarización; evaluación; exámenes; morfología; psicometría.

Abstract:
Objective. The study presents the structure and components of a standardized departmental test applied to one of the subjects of outmost curricular relevance for the core curricula of the Faculty of Languages (FL). It also presents psychometric evaluation results of the evaluation performed on the test. **Method.** To conduct this standardized test, the study used the methodology developed by the Educational Research & Development Institute (Contreras Niño, 2000; Contreras Niño & Backhoff Escudero, 2004). The test is based on the proposed psychometric model by Nitko (1994), designed to construct large-scale criterion-referenced tests. The study sample consisted of 260 first-semester students from FL in the state. The psychometric analysis followed the Item Response Theory to measure features such as the discrimination index, discrimination coefficient, and difficulty index. **Results.** Average values obtained from the psychometric analysis of the test exceeded discrimination index and coefficient standards (.32), especially those obtained in the thematic unit of greatest curricular relevance for the studied subject. Also, a general average difficulty and an acceptable distribution of said difficulty in the items were observed. **Conclusions.** The obtained values for the reliability and validity features evidenced the instrument´s quality. This directly contributes to the certainty of the level of mastery that the students show in relation to the knowledge universe represented in the test; this situation has a high impact because of the relevance of this subject.

Keywords: Second language; standardization; evaluation; test; morphology; psychometry.

Resumo:
Objetivo. Este trabalho de pesquisa apresenta a estrutura e composição de uma prova departamental padronizada de uma das disciplinas de maior relevância curricular para o tronco comum de Línguas que pertencem à Faculdade de Idiomas (FI), assim como os resultados da avaliação psicométrica realizada na mesma. **Método.** Na elaboração desta prova padronizada, foi utilizada a metodologia formulada pelo Instituto de Pesquisa e Desenvolvimento Educacional (Contreras Niño, 2000; Contreras Niño & Backhoff Escudero, 2004), baseada no modelo psicométrico proposto por Nitko (1994) para a preparação de exames em larga escala. A amostra do estudo foi composta por 260 estudantes do primeiro semestre da FI a nível estadual. Foi utilizado o método de análise psicométrico baseado na Teoria de Resposta ao Item, medindo atributos como o índice de discriminação, coeficiente de discriminação e o índice de dificuldade. **Resultados.** Da análise psicométrica realizada no exame, foram obtidos valores médios que superam o padrão em termos de índice e coeficiente de discriminação (.32), destacando-se os índices superiores ao padrão obtidos na unidade temática de maior relevância curricular para as disciplinas envolvidas. **Conclusões.** Os valores obtidos nos atributos de confiabilidade e validade permitiram evidenciar a qualidade do instrumento, tendo assim certeza do nível de domínio que os estudantes refletem em relação ao universo de conhecimentos representados na prova.

Palavras-chave: Segunda língua; padronização; avalição; exames; morfologia; psicometria.

https://doi.org/10.15359/ree.27-3.17218
https://www.revistas.una.ac.cr/index.php/educare
educare@una.ac.cr

Introducción

En la búsqueda por mejorar la calidad educativa, las instituciones de educación superior (IES) han implementado exámenes estandarizados en diferentes momentos de la vida del estudiantado universitario. Varios autores y autoras (Fernández Martínez, 2013; Fernández Navas et al., 2017; Márquez Jiménez, 2014) expresan que se opta por este tipo de pruebas con la intención de contar con instrumentos de evaluación válidos y confiables que permitan formar al estudiantado con los atributos y características necesarias para responder a las demandas sociales y empresariales sobre el tipo de estudiantado egresado universitario que se desea.

Este tipo de exámenes estandarizados han sido implementados con fines distintos, por ejemplo, son de carácter normativo en los filtros de selección de aspirantes con mejor rendimiento, para su admisión a la universidad o su promoción de grado, así como para pronosticar el desempeño académico futuro del estudiantado (Hernández Madrigal et al., 2018). De igual manera, se han empleado con fines de evaluación del tipo formativo y sumativo al implementarse en pruebas ordinarias o departamentales al final de un semestre, o bien en los exámenes parciales realizados en cada unidad de aprendizaje dentro de una asignatura.

Lo anterior se convierte en un reto constante para todas las instituciones educativas, y con mayor razón para las de nivel superior, ya que es necesario identificar, de manera acertada, las capacidades, competencias, nivel de dominio, "conocimientos y habilidades de los estudiantes a fin de [adecuar] los planes, … programas y … métodos educativos para mejorar el proceso de enseñanza y aprendizaje" (Hernández et al., 2018 en Gutiérrez Benítez & Acuña Gamboa, 2020, p. 119). Esto, sin duda, expresa claramente la necesidad de que los métodos empleados en la construcción y diseño de pruebas sean de calidad y con ello mejorar significativamente el proceso evaluativo (Tristán López & Pedraza Corpus, 2017).

Con la creación del Centro Nacional para la Evaluación de la Educación Superior (CENEVAL) en 1994 y la del Instituto Nacional para la Evaluación de la Educación (INEE) en el 2002, en México se ha observado el impulso de este tipo de pruebas estandarizadas, se logró, así, la creación del Examen Nacional de Ingreso a la Educación Superior, Examen General para el Egreso de la Licenciatura y el Examen Nacional de Ingreso al Posgrado, pruebas que por años han sido el instrumento para seleccionar, promocionar y obtener distinciones académicas en el sistema educativo del país (Centro Nacional de Evaluación para la Educación Superior [CENEVAL], 2017).

Este tipo de iniciativas a nivel nacional han motivado y propiciado que instituciones como la Universidad Autónoma de Baja California (UABC) desarrollen instrumentos diseñados especialmente para responder a sus necesidades evaluativas, como lo fue el Examen de Habilidades y Conocimientos Básicos (EXHCOBA) utilizado para selección o admisión. En la actualidad, la UABC ha buscado el desarrollo y aplicación de exámenes departamentales estandarizados; por ello, en el año 2016 activó estas pruebas de manera institucional con la impartición del Diplomado de Evaluación Colegiada del Aprendizaje por parte del Instituto de

https://doi.org/10.15359/ree.27-3.17218
https://www.revistas.una.ac.cr/index.php/educare
educare@una.ac.cr

Investigación y Desarrollo Educativo (IIDE), donde se invitó a participar a la mayor cantidad de facultades para recibir formación en el diseño y construcción de estas pruebas, y con ello coadyuvar al logro de una de las metas del Plan de Desarrollo Institucional alineada a la aplicación de exámenes departamentales y de trayecto para mejorar continuamente los niveles de aprendizaje del alumnado.

Este trabajo se centra en la descripción de los resultados psicométricos obtenidos a partir de la aplicación de una prueba departamental estandarizada desarrollada para la asignatura de morfología en la segunda lengua, una de las dos asignaturas de mayor relevancia curricular en el tronco común de idiomas de la Facultad de Idiomas de la UABC. Para lograr lo anterior se describen las diferentes etapas que comprenden el método para el diseño de la prueba, con énfasis en la etapa de análisis de la calidad psicométrica. Así mismo se detallan los instrumentos tecnológicos empleados en dicho análisis, los criterios de confiabilidad y validez empleados y la interpretación técnica de estos. Posteriormente se muestran los resultados obtenidos y se discuten los aspectos más sobresalientes encontrados.

Referentes teóricos

Dentro de los principales aportes de los procesos evaluativos en la educación está el dar pauta, entre otras cosas, sobre el estado que presenta el alumnado en cuanto a su nivel de dominio de un determinado universo de conocimientos. Estos universos están enmarcados por las cartas descriptivas o programas de las unidades de aprendizaje de aquellas asignaturas que componen a un programa educativo. Sin embargo, algunas asignaturas tienen mayor peso curricular que otras, ya que pueden ser integradoras de conocimientos a partir de otras asignaturas, o bien ser insumos para la concreción de conocimientos futuros que insiden directamente en el logro de las competencias profesionales.

Tal es el caso de la asignatura de morfología de la segunda lengua, cuyos conocimientos son indispensables para la concreción de varias de las competencias profesionales especificadas en el perfil de egreso de la carrera de Licenciatura en Enseñanza de Lenguas de la Facultad de Idiomas en la UABC. Por lo anterior, para este tipo de asignaturas cuyo impacto en el desarollo profesional del estudiantado es mayor, se requiere contar con pruebas válidas y confiables, las cuales, a través de rigurosos procedimientos metodológicos, puedan precisar objetivamente los niveles de dominio. Para lograr esto comunmente se recurre a lo que se conoce como pruebas estandarizadas.

Evaluación estandarizada

Este tipo de pruebas son instrumentos de medición que poseen un amplio desarrollo técnico y metodológico, dotando de una capacidad para medir rasgos latentes u observables en la población, con alto grado de precisión (Tristan López & Pedraza Corpus, 2017). En este sentido, la estandarización se entiende como un proceso de sistematización de todos aquellos elementos vinculados a la recolección e interpretación de información, aplicando los mismos instrumentos o técnicas tanto para el análisis, recopilación y la interpretación; lo anterior parte

https://doi.org/10.15359/ree.27-3.17218
https://www.revistas.una.ac.cr/index.php/educare
educare@una.ac.cr

de una normativa predeterminada que rige todo el actuar del proceso en cuestión (Jornet Meliá, 2017; Oyarzún Maldonado & Soto González, 2021).

En otros espacios se ha demostrado que la principal característica de las evaluaciones estandarizadas radica en los marcos de referencia teóricos y metodológicos rigurosos (Backhoff Escudero, 2018; Fernández Navas et al, 2017). Estos marcos de referencia permiten "efectuar mediciones que dan como resultado valoraciones cuantificables de atributos asociados a la calidad de la prueba" (Gutiérrez Benítez & Acuña Gamboa, 2022, p. 327), en cuanto a la confiabilidad y su respectiva validez. En este sentido, uno de los aportes de este tipo de evaluaciones es la posibilidad de tener un mayor acercamiento a la realidad, pues permiten señalar que la variación en los resultados está en razón del sujeto evaluado o factores concretos de intervención, pero no a la calidad técnica del instrumento o su proceso de construcción (Jornet Meliá, 2017). Este tipo de pruebas, cuando se alinean con objetivos de aprendizaje, suministran la base para procesos de retroalimentación que el equipo docente puede utilizar para precisar fortalezas y debilidades curriculares, así como comprobar el logro alcanzado de manera individual en cada estudiante (Ravela, 2010).

Psicometría

La psicometría se conceptualiza como una disciplina de la psicología cuyo único fin es el contribuir con la creación de soluciones al problema implícito de la medición, como parte del proceso en una investigación psicológica. En el caso particular de la perspectiva teórica, incluye las teorías que abordan las medidas en el campo de acción de la psicología; permite, a partir de una descripción por categorías y mecanismos de evaluación, comprobar su utilidad y precisión, así como la investigación de nuevos métodos, teorías y modelos matemáticos que provean de mejores instrumentos de medida (Aliaga Tovar, 2007). Entre otras aspectos, la psicometría estudia dos ramas importantes, una asociada a la teoría implícita en la medición, rama caracterizada por el uso de estadística aplicada en la elaboración y análisis correspondiente de los instrumentos de medición; y por otro lado, una rama que se encarga del estudio de los usos que se le dan a las pruebas con el fin específico de medir o evaluar constructos psicológicos (Medrano & Pérez, 2019).

Entre los atributos de especial interés en esta investigación se encuentran la confiabilidad, validez de contenido y la validez de constructo de la prueba. Por un lado, la confiabilidad se relaciona con los errores generados como producto de la medición, y busca responder al problema de la certeza que refleja la puntuación del sustentante con su aprovechamiento; lo anterior implica idealmente que un instrumento confiable, aplicado en diferentes espacios de tiempo, debería arrojar mediciones consistentes (González Campos & Aspeé Chacón, 2021; Martínez Arias et al., 2014). De esta manera, la confiabilidad se encuentra ligada la validez, ya que, si bien una prueba puede ser confiable, puede no ser válida.

https://doi.org/10.15359/ree.27-3.17218
https://www.revistas.una.ac.cr/index.php/educare
educare@una.ac.cr

Formalmente la validez es un juicio evaluativo que implica una evidencia empírica y de sustento teórico que permiten avalar la suficiencia y lo apropiado de las interpretaciones con base en los puntajes obtenidos en las pruebas, va más allá de los ítems de la prueba, al contemplar el contexto en el que se desarrolla esta (Aliaga Tovar, 2007; Árraga Barrios & Sánchez Villarroel, 2012; Robles Pastor, 2018).

Atributos de confiabilidad y validez en un instrumento

Este trabajo de investigación emplea la teoría de respuesta el ítem (TRI; Aune & Attorresi, 2019) para el desarrollo del instrumento de evaluación, y con base en dicha teoría se emplean "una serie de atributos [mediante] los cuales se puede observar la validez y confiabilidad de la prueba" (Gutiérrez Benítez & Acuña Gamboa, 2022, p. 333).

La TRI, pretende suministrar, según Cortada de Kohan (2004), "[la]fundamentación probabilística al problema de medir constructos latentes (no observables) y considera al ítem como unidad básica de medición" (p. 95), provee así una forma de observar el posible comportamiento del sujeto sustentante en un ítem en particular. Esto último dota a la TRI de un marco de referencia unificado que ofrece la posibilidad de conceptualizar el sesgo a nivel de ítem.

Es así que la calidad técnica de una prueba basada en la TRI puede determinarse mediante el índice de dificultad del ítem, índice de discriminación y el coeficiente de discriminación. En este sentido, el índice de dificultad del ítem se define como la proporción de una muestra o población que responde acertadamente un ítem o pregunta en una prueba (Hurtado Mondoñedo, 2018; Medina Paredes et al., 2019) y se da en el intervalo específico de 0 a 1, cuanto más cercano a cero sea este valor significa que la pregunta es más difícil de acertar. Croker & Algina (1986, en Backhoff Escudero et al., 2000) mencionan que usualmente dicha proporción se representa como *p*.

Respecto al índice de discriminación de la prueba, generalmente representado como D, este se entiende como la propiedad que tiene un ítem para poder separar a los sujetos sustentantes que tienen una mejor puntuación final en la prueba de quienes tienen una menor puntuación (Medina Paredes et al., 2019); los valores posibles de este índice están dados en el intervalo de -1 a 1. En relación con el coeficiente de discriminación también conocido como el punto de correlación biserial (rpbis), es el atributo que permite medir con mayor certeza la discriminación de un ítem, lo anterior debido a que, como lo refieren algunos estudios (Medina Paredes et al., 2019; Pérez Tapia et al., 2008), se le considera como una forma de medida de la consistencia que tiene un ítem con toda la prueba en su conjunto. Este permite observar la correlación existente entre los puntajes obtenidos por los sujetos en un ítem especifico con el puntaje obtenido en su totalidad en la prueba, conocer el poder predictivo que tiene el ítem de la relación acierto/error con la calificación total obtenida en una prueba, así como identificar si el estudiantado con mejor desempeño es aquel que logra las respuestas correctas (Backhoff Escudero et al., 2000; Molina et al., 2015 en Aguilar-Salinas & de las Fuentes Lara, 2023).

https://doi.org/10.15359/ree.27-3.17218
https://www.revistas.una.ac.cr/index.php/educare
educare@una.ac.cr

En el caso de los atributos que permiten demostrar la confiabilidad de una prueba se emplean cálculos de consistencia interna (Reidl-Martínez, 2013; Viladrich et al. 2017) como el índice de confiabilidad de Kuder-Richardson (KR20), también se encuentra el índice del alfa de Cronbach, aunque este es mayormente utilizado para pruebas con ítems que evalúan un único dominio o unidimensionales (para ítems que examinan más de una dimensión se emplea el cálculo del alfa de Cronbach estratificado). Cuantificablemente, estos índices deben encontrarse en un rango de .8 para el alfa de Cronbach, mientras que para el índice KR20 este debe ser mayor a .7. Otro procedimiento para evaluar "la confiabilidad de una prueba es mediante [la aplicación del] método test retest, el cual consiste en aplicar la prueba en diferentes momentos a la misma muestra" (Gutiérrez Benítez & Acuña Gamboa, 2022, p. 335) de participantes, con el fin de evaluar las fluctuaciones o variaciones presentes en los resultados obtenidos. Para identificar o evaluar este grado de confiabilidad se efectúa el cálculo del coeficiente de correlación intraclase (CII), o también llamado como el índice de concordancia (Correa-Rojas, 2021; Mandeville, 2005), dicho índice debe encontrarse lo más cercano a uno para ser considerado de calidad.

Metodología

Unidad de análisis

La Facultad de Idiomas de la Universidad Autónoma de Baja California tiene presencia en cuatro municipios del Estado, por lo que para esta prueba se trabajó con distintas poblaciones a lo largo de un lapso de 2 años (2018-2019), con lo cual el instrumento se ha ido calibrando hasta lograr lo mostrado en este artículo. Se contó con la participación de 260 estudiantes provenientes de las cuatro sedes (Mexicali, Ensenada, Tijuana y Tecate), cursan el primer semestre del tronco común de la Licenciatura en Idiomas. El sexo de la población fue de un 50% que se identificó como masculino y un 50% que se identificó como femenino. Con una edad promedio de 18 años, de los cuales se eligieron a su vez un 33% de alumnado con bajo rendimiento (menores o iguales a 6,9), 33% del alumnado con rendimiento regular (entre 7 y 8,9) y 34% de alumando con alto rendimiento (mayores o iguales a 9), esto con el fin de contar con una muestra más representativa del universo de la Facultad de Idiomas de la UABC, a la luz de analizar los resultados de la evaluación en diferentes cohortes de desempeño académico. Además, esta distribución responde al método empleado para la realización del análisis psicométrico, que como ya se mencionó anteriormente está basado en los procedimientos convencionales de la teoría de respuesta al ítem. La muestra compuesta por los 260 escolares corresponde a un cálculo estadístico considerando un 95% de nivel de confianza y un 5% de margen de error.

Instrumentos

En relación con el instrumento que se empleó, se precisa que el examen departamental de morfología (entendiendo a la morfología como la rama de la lingüística que se encarga del estudio de la estructura interna de las palabras y el proceso de formación de estas mismas)

https://doi.org/10.15359/ree.27-3.17218
https://www.revistas.una.ac.cr/index.php/educare
educare@una.ac.cr

está diseñado para evaluar el nivel de dominio que el estudiantado posee en relación con todo el universo de conocimientos curriculares de dicha asignatura. Al ser una asignatura seriada, implica la adquisición de conocimientos clave que permiten concretar otros conocimientos, por ello la importancia de determinar el nivel de aprendizaje alcanzado por el alumnado y con ello hacer predicciones sobre el desempeño en futuras asignaturas.

Un producto generado en una de las etapas de desarrollo de la prueba es la tabla de especificaciones, donde se detallan la cantidad de ítems por contenido temático, así como el foco y nivel taxonómico de cada uno. En este sentido, para esta prueba se elaboraron 63 ítems de opción múltiple, el 3% de los ítems son de nivel de conocimiento, 22% de comprensión, 28% de aplicación, 38% análisis y 8% de evaluar esto acorde con la taxonomía de Bloom (Cuenca et al., 2021; Parra Giménez, 2017). La distribución de los reactivos está dada por la ponderación en cuanto a relevancia curricular para la cual se elabora un ítem, así existirán más ítems representativos en el examen de aquellos temas, subtemas, etc., que tengan mayor relevancia curricular para el logro de la competencia general del curso, así como aquellos temas que son críticos para la concreción de conocimientos futuros dentro de la misma asignatura; o bien, para asignaturas futuras comprendidas en el programa.

Lo anterior no se realiza de forma arbitraria, por el contrario, dentro de la metodología de desarrollo de la prueba hay una etapa dedicada a la obtención de dicho índice de relevancia curricular (IRC) así como el índice de concordancia (índice *Kappa*) o validez de contenido. Para lograr lo anterior se implementa un jueceo por personal experto, se determina jerárquicamente qué contenidos son más importantes y por qué (se definen criterios de evaluación concretos), y mediante una serie de análisis matemáticos se obtienen los índices de IRC y *Kappa*, que en el caso de este último se emplea también la escala de valoración propuesta por Landis & Koch (1977).

En relación con los análisis psicométricos de la dificultad y discriminación de la prueba, estos fueron realizados mediante el programa de cómputo especializado ITEMAN, así como mediante el software TAP publicado por Brooks & Johanson (2003). Además de efectuar el análisis con estos softwares se utilizó el mismo software en el que se aplicó el examen, el cual integra un módulo de evaluación psicométrico. Este software almacenó en su base de datos las respuestas de cada sujeto de los 260 sustentantes y a su vez generó el archivo fuente para efectuar este análisis con los softwares de terceros antes mencionados. Dicho software es una tecnología innovadora y lleva por nombre Sistema de exámenes estandarizados (SIEXAES). Se analizaron los resultados de los tres softwares para observar variaciones entre las mediciones y con ello calibrar los cálculos de forma que se obtengan números exactos y reducir errores por operaciones con menor cantidad de decimales o bien redondeo.

Procedimiento

La metodología utilizada en la elaboración de este examen fue formulada en el IIDE (Contreras Niño, 2000; Contreras Niño & Backhoff Escudero, 2004) empleando como fundamento el paradigma psicométrico expuesto por Nitko (1994) para la confección de pruebas de gran

https://doi.org/10.15359/ree.27-3.17218
https://www.revistas.una.ac.cr/index.php/educare
educare@una.ac.cr

escala desde el enfoque de criterios. Esto tipo de exámenes proporcionan resultados que describen el número de competencias que el estudiantado domina, en relación con el total de competencias evaluadas (Backoff Escudero, 2018). Estos exámenes están orientados por el currículo, que implica que todas aquellas decisiones sobre lo que se va a evaluar y la forma de evaluarlos están directamente orientados por lo que se establece en el currículo.

Dicha metodología contempla seis etapas en el desarrollo del examen: 1) definición del dominio de resultados que pretende el currículo; 2) análisis del currículo; 3) desarrollo de un plan de evaluación; 4) producción y validación de ítems; 5) análisis primario de resultados y; 6) análisis secundario de resultados. En relación con la etapa tres y cuatro, entre los principales productos derivados de estas, fueron la tabla de especificaciones de ítems (consultar Tabla 1) y el desarrollo de especificaciones de ítems (consultar Figura 1). Con referencia a la Tabla 1 de especificaciones de ítems este proceso presenta las decisiones de estrategia evaluativa del examen.

Tabla 1: Ejemplo de especificación de un contenido temático

Contenido	IRC	Especificaciones	Ítems	Foco del ítem	Tipo de ítem	Nivel taxonómico
1.1.1 Definition of the discipline	0.600	1	2	Probar el dominio del concepto morfología mediante su definición.	OM	Comprender
				Probar el dominio del concepto morfología mediante la identificación de sus características.	OM	Comprender

Nota: Elaboración propia.

En la Tabla 1 se ejemplifica para solo un contenido temático de todo el universo de conocimientos que comprende la asignatura una serie de atributos y características que permiten valorar e integrar el mismo en el diseño de la prueba. Por ejemplo, se detalla el índice de relevancia curricular (IRC), cuantos ítems para ese contenido deben producirse, cuál es el foco u objetivo que se busca alcanzar con ese ítem, qué tipo de ítem es (para este ejemplo OM se refiere a opción múltiple) y el nivel taxonómico que el ítem comprende, es decir, la dimensión del nivel cognitivo que el ítem implica.

Por otra parte, el desarrollo de especificaciones de ítems es un proceso formal que describe al responsable quien finalmente elaborará el ítem, las características que debe tener la tarea evaluativa. En otras palabras, se habla de un retrato por escrito de un ítem, este retrato detalla las características que deben tener los reactivos y las respuestas a estos, de manera que pueda considerarse válida la tarea evaluativa. Un ejemplo de una especificación de ítem es la mostrada en la Figura 1.

https://doi.org/10.15359/ree.27-3.17218
https://www.revistas.una.ac.cr/index.php/educare
educare@una.ac.cr

Figura 1: Ejemplo de una especificación de ítems

Información contextual
2.1 Derivation
In linguistics refers to the formation of a word based on another word; It is also called derivation to the relation that has a word with its base.
Unidad 2. Rules of Word formation **Competencia:** Identificar los mecanismos de formación de palabras propios de la lengua inglesa en ejemplos reales de uso, para comprender los principios de la formación del léxico mostrando iniciativa en el desarrollo de las clases.
IRC: 0.725 **Criterios**: • Contribución al logro de la competencia de la unidad: .20 • Dosificación: .10 • Carga horaria: .10 • Relevancia disciplinaria: .20 • Nº servicios que recibe: 1 • Nº servicios que proporciona: 3
Recibe servicios de: 1.3.1 Morphemes and allomorphs **Proporciona servicios a:** 3.1 Nouns 3.2 Adjectives **Práctica 2** P2. Elaborar un resumen en el que plasmen las características de los diferentes mecanismos de formación de palabras, proveyendo ejemplos reales de uso que faciliten el uso posterior como texto de referencia, de acuerdo a la normatividad de la APA.

El contenido Derivation es el primer tema conceptual que el alumno aborda en la Unidad 2. Rules of word formation, en el cual se enseña sobre la estructura y características de la derivación de palabras.

Para el aprendizaje del contenido de Derivation es importante que el estudiante haya comprendido previamente los conceptos de Morphemes and allomorphs, vistos en la Unidad 1. Basic concepts. Este contenido da servicio a la práctica de la Unidad 2 para el logro de la competencia en lo que se refiere a identificar los mecanismos de formación de palabras propios de la lengua inglesa, para comprender mejor los principios de la formación del léxico. El dominio de este contenido es fundamental para el aprendizaje de contenidos posteriores, como Nouns y de Adjectives (que forman parte de la Unidad 3).

Por estos motivos, Derivation fue uno de los que obtuvo las puntuaciones más altas en los criterios de contribución al logro de la competencia del curso, y más altos en relación al índice de relevancia curricular. Por ello se elaborarán 2 ítems: uno que ponga a prueba que el alumno identifica la función de una palabra dada; y otro ítem que ponga a prueba que el alumno identifica la raíz de una palabra dada.

Información contextual o indicaciones para responder este ítem:
Ninguna.

Información tabular, gráfica o textual a emplear en el ítem:
Ninguna.

Dimensión proceso cognitivo

Dimensión conocimiento	Recordar	Comprender	Aplicar	Analizar	Evaluar	Crear
Conocimiento factual						
Conocimiento conceptual			x			
Conocimiento procedimental						
Conocimiento metacognitivo						

Especificación de la base del ítem:
1) Podrá presentar una palabra y solicitar que se identifique la función de la palabra.
   Por ejemplo: *Identify the function of derivated word "intelligently"*
2) Podrá presentar una palabra y solicitar que se identifique la raíz de la palabra.
   Por ejemplo: *Identify the roots of derivated word "performing"*

Especificación de la respuesta correcta a.
1) Será la función gramatical
   Por ejemplo: *adverb*
2) Será la raíz en cuestión.
   Por ejemplo: *perform*

Especificación del distractor b.
1) Se podrá presentar una función gramatical incorrecta.
   Por ejemplo: *adjective*
2) Se podrá presentar una raíz incorrecta.
   Por ejemplo: *performs*

Especificación del distractor c.
1) Se podrá presentar una función gramatical incorrecta.
   Por ejemplo: *verb*
2) Se podrá presentar una raíz incorrecta.
   Por ejemplo: *forming*

En todos los casos, el distractor C no repetirá una característica ya incluida en alguna de las modalidades de la opción B.

Ítem muestra	
Identify the function of the derivated word "intelligently".	
a.	adverb
b.	adjective
c.	verb

Ítem muestra	
Identify the roots of derivated word "performing".	
a.	perform
b.	performs
c.	forming

Nota: Elaboración propia.

https://doi.org/10.15359/ree.27-3.17218
https://www.revistas.una.ac.cr/index.php/educare
educare@una.ac.cr

La etapa cuatro, referente a producir y validar ítems, es un trabajo técnico delicado y requiere mucha atención, puesto que se busca mantener estricta congruencia con la especificación que lo produce. Esto es altamente importante, porque de esta estricta congruencia se obtienen evidencias relacionadas con la validez del contenido del examen; caso contrario, se compromete de manera directa su calidad, confiabilidad y validez. Las etapas cinco y seis de la metodología son las de especial interés para este trabajo de investigación, y contemplan el análisis técnico de la calidad de los ítems. El análisis del comportamiento de los ítems del examen de Morfología de la Segunda Lengua ante los grupos examinados de la muestra se llevó a cabo mediante los procedimientos convencionales de la TRI (Aune & Attorresi, 2019; Gómez Rada, 2004; Hidalgo-Montesinos & French, 2016; Muñiz, 2010). Como se mencionó anteriormente, en términos psicométricos, la calidad se observa con la medición del índice de dificultad, el índice de discriminación y el coeficiente de discriminación de un ítem, que requieren de aplicar cálculos matemáticos específicos para la obtención de sus valores.

En el caso particular del coeficiente de discriminación, para señalar la calidad de este atributo algunos estudios (Carlos Martínez et al., 2011) concuerdan con la clasificación del coeficiente de discriminación mostrados en la Tabla 2.

Tabla 2: Clasificación de valores del coeficiente de discriminación

Calidad	Valor del coeficiente
Excelente	Mayor a .35
Buena	Mayor o igual a .26 y menor a .35
Regular	Mayor o igual a .18 y menor a .26
Pobre	Mayor a 0 y menor a .18
Descartar	Menor a 0

Nota: Elaboración propia a partir de los datos de Carlos Martínez et al. (2011).

Según la Tabla 2, idealmente se deben poseer valores de rpbis mayores o iguales a .26 para considerarse de calidad, cuanto más cercano a uno sea este valor mayor será la calidad. Aunque los estudios antes mencionados consideran como regular un valor mínimo de .15, para esta investigación se tomó de referencia un valor mínimo de .18 para ser considerado como regular.

Resultados

Los primeros resultados arrojados por los sistemas de análisis psicométricos fueron los promedios de valores del índice y coeficiente de discriminación, así como del índice de dificultad. Otros estadísticos descriptivos además de los antes mencionados son mostrados en la Tabla 3.

https://doi.org/10.15359/ree.27-3.17218
https://www.revistas.una.ac.cr/index.php/educare
educare@una.ac.cr

Tabla 3: Medias de los valores del índice de dificultad (p), índice de discriminación (d), coeficiente de discriminación (rpbis) y aciertos

Muestra (N)	Media de aciertos	Puntaje mínimo	Puntaje máximo	Media del valor P	Media del valor D	Media de Rpbis
260	43.263	34	60	0.647	.331	0.322

Nota: Elaboración propia.

Como se observa en la Tabla 3, el examen presenta una dificultad media con una leve tendencia a ser más fácil que difícil, obteniendo un valor p de .647. En cuanto a la discriminación se obtuvo un índice promedio de .330 para el valor D, asimismo fue interesante el promedio del coeficiente de correlación punto biserial o coeficiente de discriminación, el cual se ubicó en .322, que significa que el poder discriminatorio es bueno. En relación con la media de aciertos esta se ubicó en 43.263, significando que en promedio el estudiantado obtiene una calificación de 70 puntos en una escala de 0 a 100. La distribución porcentual de los 63 ítems en las cinco unidades quedó como se ilustra en la Tabla 4.

Tabla 4: Distribución porcentual de la representación de ítems en la prueba

Unidad Temática	Representación en la prueba
Unidad 1. Basic concepts.	20%
Unidad 2. Rules of word formation.	35%
Unidad 3. Open class words.	20%
Unidad 4. Closed class words.	15%
Unidad 5. Compound words, blends and phrasal words.	10%
Totales	**100**

Nota: Elaboración propia.

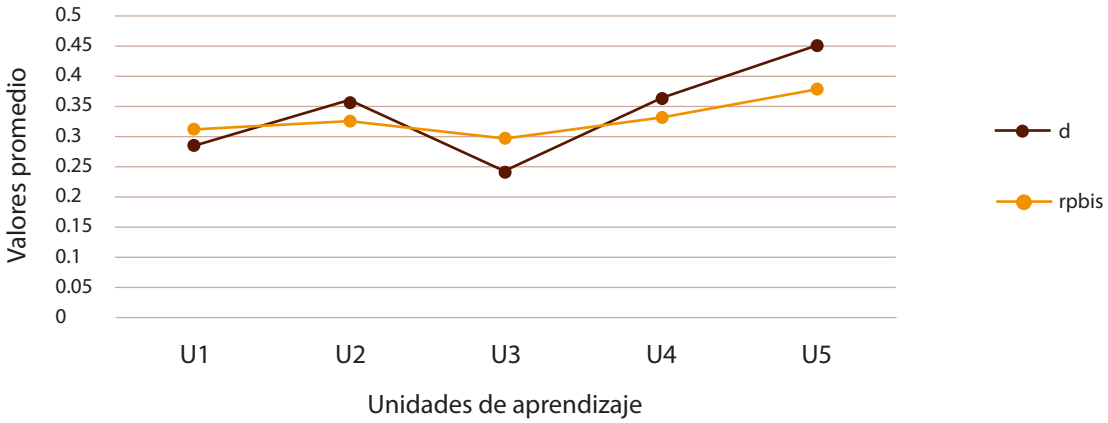
La Tabla 4 muestra que la unidad dos tiene una representación en ítems equivalente a un tercio de toda la prueba. El dato anterior es de especial interés, ya que la unidad dos al ser analizada según el índice de relevancia curricular obtuvo las puntuaciones más altas en este indicador. Los 31 temas que comprende el currículo de la asignatura fueron clasificados según su IRC, obteniendo así un total de 11 temas con el mayor IRC y por ende el más alto impacto en la concreción de conocimientos para toda la asignatura. De estos 11 temas, seis

https://doi.org/10.15359/ree.27-3.17218
https://www.revistas.una.ac.cr/index.php/educare
educare@una.ac.cr

corresponden específicamente a la unidad dos, esto significa que más del 50% de todos los temas de mayor importancia e impacto en el aprendizaje del estudiantado son abordados en la unidad dos. Lo anterior destaca la importancia de que los ítems aplicados en esta unidad posean las características más ideales al momento de evaluar.

En relación con esto fue relevante observar los valores obtenidos en los nueve temas que abarca la unidad dos, con especial atención en los seis de mayor IRC. La distribución de los valores del índice de discriminación y coeficiente de discriminación, para cada una de las cinco unidades que abarca la asignatura, se muestran en la Figura 2.

Figura 2: Distribución de los valores promedio del índice y coeficiente de discriminación por unidad



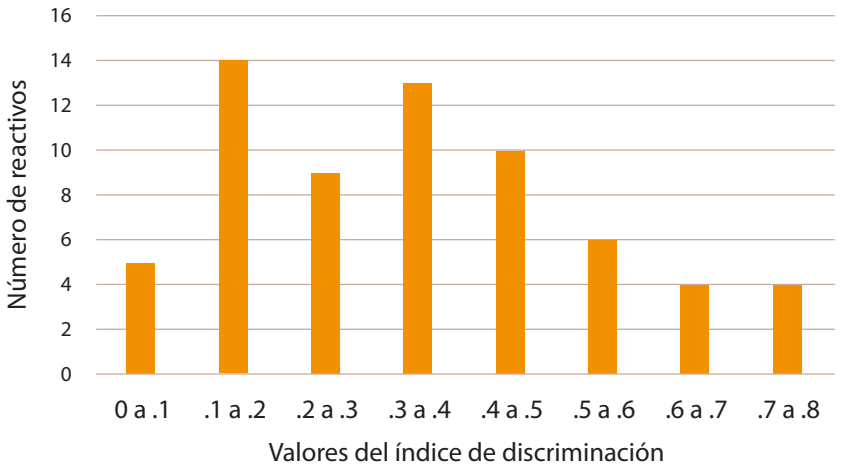
Nota: Elaboración propia.

En la Figura 2 anterior se observa que todas las unidades tuvieron valores buenos en promedio, tanto para el índice de discriminación como para el coeficiente de discriminación. En el caso específico de la unidad dos, el rango de valores en los que se ubicaron los índices de discriminación fueron de .17 y .72, por su parte el rango obtenido para el coeficiente de discriminación (Rpbis) fueron de .14 y .5. En general la unidad 2, que es la más importante para la asignatura, obtuvo un coeficiente de discriminación promedio de .32 y un índice de discriminación promedio de .35. Lo anterior significa que en cuanto a poder discriminatorio los ítems desarrollados para esta unidad cumplen satisfactoriamente con los estándares de calidad.

En cuanto a los índices de discriminación se puede observar la distribución de los 63 ítems de la prueba en la Figura 3, que si bien se identifica un número alto de valores en el rango de .1 a .2 (14), solo seis de ellos se encuentran por debajo de .19, lo que significa que tienen un margen de mejora positivo.

https://doi.org/10.15359/ree.27-3.17218
https://www.revistas.una.ac.cr/index.php/educare
educare@una.ac.cr

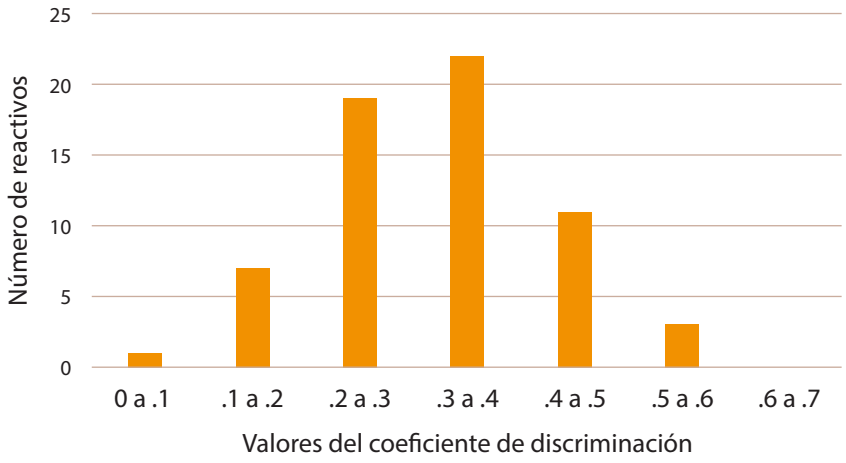
Figura 3: Distribución de los valores obtenidos para el índice de discriminación



Nota: Elaboración propia.

Por su parte, las distribuciones de los valores del coeficiente de discriminación en los 63 ítems se muestran en la Figura 4.

Figura 4: Distribución de los valores obtenidos para el coeficiente de discriminación (rpbis)



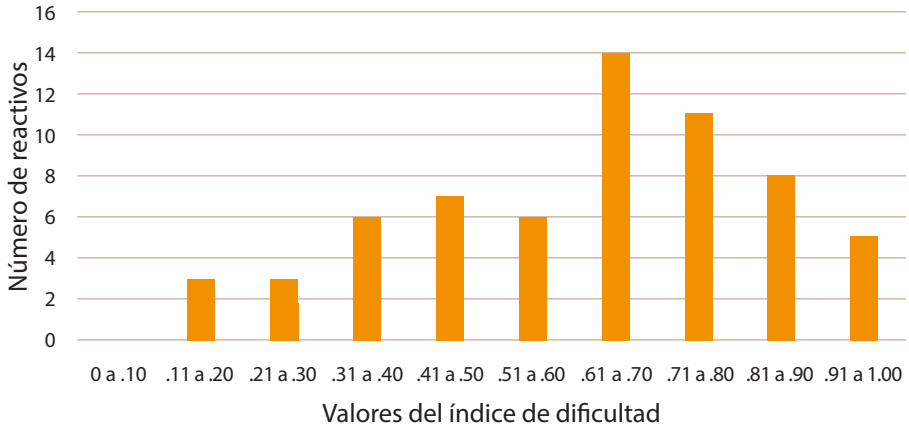
Nota: Elaboración propia.

Como se logra apreciar, de los 63 ítems que comprenden el examen, solo el 8% de ellos tiene un coeficiente de discriminación bajo (menor a .18), mientras que más del 70% de los ítems presenta un valor de rpbis dentro del rango de bueno a excelente. El dato anterior recalca la confiabilidad del instrumento al momento de discriminar a la población sustentante, dando así seguridad sobre los resultados obtenidos con la prueba.

https://doi.org/10.15359/ree.27-3.17218
https://www.revistas.una.ac.cr/index.php/educare
educare@una.ac.cr

En relación con el índice de dificultad, la distribución de la dificultad de cada uno de los ítems se muestra en la Figura 5, en la cual se puede notar como el mayor porcentaje de ítems se encuentra en el rango de dificultad de .61 a .70. Por esta razón el promedio de dificultad del examen se ubica en .647; como se mencionó anteriormente, esto significa que el examen tiene una leve tendencia a ser más fácil que difícil.

Figura 5: Distribución de los valores obtenidos para el índice de dificultad (*p*)



Nota: Elaboración propia.

Es importante señalar que la dificultad de los ítems se encuentra distribuida a lo largo de las cinco unidades de la asignatura, aunque como se puede apreciar en la Figura 5, hay una proporción considerable de ítems en la clasificación de fáciles a muy fáciles. En relación con esta misma distribución de valores del índice de dificultad, en la Tabla 5 se presenta la clasificación de los 63 ítems.

Tabla 5: Clasificación de los ítems de la prueba según su dificultad

Dificultad	Valor de *p*	Porcentaje de ítems en la prueba
Muy fáciles	Mayores o iguales a .90	8%
Fáciles	.76 a .89	17%
Dificultad media	.46 a .75	54%
Difíciles	.21 a .45	16%
Muy difíciles	Menores o iguales a .20	5%

Nota: Elaboración propia.

https://doi.org/10.15359/ree.27-3.17218
https://www.revistas.una.ac.cr/index.php/educare
educare@una.ac.cr

En la Tabla 5 se hace notar que en general la prueba tiene un 25% de ítems que están en el rango de fáciles a muy fáciles, mientras que, en el lado opuesto el 21% de ítems está ubicado en el rango de difíciles a muy difíciles. Esto hace que la prueba tenga una proporción positivamente balanceada de la distribución de dificultad de ítems en toda la prueba y cercano a lo que sugieren ciertos estudios (Backhoff Escudero et al., 2000; Ortiz Romero et al., 2015).

Una vez realizados los análisis anteriores, fue necesario determinar los criterios de calidad aceptables de los 63 ítems que compusieron la prueba, con el fin de mejorar aquellos ítems que estaban en un rango muy cercano a regular o bien dentro de este. Así mismo desechar aquellos que no cumplían ni con la medida regular. Para lograr lo anterior se consideraron los siguientes criterios: ítems con coeficiente de discriminación menor a .18, ítems con índice de discriminación menor a .2 y por ultimo ítems que tuvieran una dificultad menor a .2 y mayor a .9.

Dicho lo anterior se concluyó que de los 63 ítems que componen la prueba deben ser revisados para mejorar su poder discriminatorio un total de 8 ítems, mientras que hay 10 ítems que tienen que ser elaborados nuevamente, cinco de ellos por que poseen una discriminación nula o menor a .1, mientras que otros cinco tienen una dificultad superior a .9. Es interesante notar que solo un número reducido de la representatividad total de la prueba requiere de ser cambiada drásticamente, dando un margen de calidad aceptable en toda la prueba. Además, se retoma el hecho de que la unidad con mayor IRC para toda la asignatura cumplió con valores que se ubicaron en las clasificaciones de buenas a excelentes para el coeficiente de discriminación.

Conclusiones

Es claro que para hacer estimaciones y tomar decisiones acertadas con base en el resultado de una prueba, es necesario que esta sea válida y confiable, de otra manera se pudieran emitir juicios o conclusiones erróneas. En el caso particular de las pruebas departamentales se realizan con el fin de evaluar de manera general todo el universo de conocimiento implicado en una determinada asignatura, con cuyos resultados se pueda estimar el nivel de dominio que presenta cada estudiante y poder hacer apreciaciones sobre diferentes aspectos, tales como la cobertura del total de contenidos de la materia, comparaciones entre la evaluación departamental y la realizada por parte del personal docente, hacer predicciones sobre rendimiento futuro entre otros.

Hablando en el particular del índice de dificultad, la prueba presentó una dificultad promedio de .647, lo que significa que la prueba tiende a ser más fácil que difícil, más del 50% de los ítems en la prueba tuvieron una dificultad ubicada dentro del rango de medio a muy fácil, considerando que este índice se mide de 0 a 1, podemos expresar que cualquier ítem con un valor de .65 en adelante tiende a ser más fácil que difícil. En general la prueba muestra una distribución de todo el rango de valores para el índice de dificultad, sin embargo, como se dijo anteriormente hay una ligera mayor representación de ítems clasificados como fáciles y muy fáciles (25%) en comparación con los ítems clasificados como difíciles y muy difíciles (21%).

https://doi.org/10.15359/ree.27-3.17218
https://www.revistas.una.ac.cr/index.php/educare
educare@una.ac.cr

Sin embargo, se identificó un 13% de ítems que no cumplen con las normas de calidad deseadas, ya que presentan un índice de dificultad menor a .2 o bien superior a .9, aunque cabe mencionar que de estos el 6% presentan una discriminación positiva, pero los valores obtenidos no fueron muy altos. En el caso de los muy fáciles, 5 ítems superaron la barrera de dificultad de .9, lo que los hace demasiado fáciles de contestar, por lo que para efectos de discriminación también fallan, puesto por su facilidad son acertados casi en igual proporción por el grupo con mejor rendimiento como por el grupo con menor rendimiento. En relación con el coeficiente de discriminación, la prueba mostro un promedio general de .322, dato muy importante porque a diferencia del índice de discriminación el coeficiente permite también medir el hecho de que los sujetos sustentantes que tienen un mejor dominio a nivel general en la prueba sean quienes acierten correctamente los ítems. Además, permite valorar la relación predictiva entre acertar correctamente un ítem y la calificación obtenida en la prueba. Es notorio también el hecho de que menos del 10% de todos los ítems que componen la prueba tuvieron valores por debajo del estándar de calidad regular (menores a .18).

En referencia al índice de discriminación de la prueba, se hicieron notar ciertas situaciones, una de ellas fue que hubo un número considerable de ítems que se encontraron en el rango de valores menor a .2, sin embargo, a pesar de que hay 14 ítems dentro del rango de valores de .1 a .2, solo 7 están por debajo del estándar mínimo de aceptación de .18. Dentro de los valores promedios por cada unidad en este índice el más bajo se ubicó en .243, y correspondió a la unidad 3. Para las restantes unidades los valores promedios se ubicaron en el rango de .286 y .448. Aunque hay que considerar que estos valores también se ven afectados por la cantidad de ítems de cada unidad, como por ejemplo la unidad 3 tiene el doble de ítems que la unidad 5.

Los análisis psicométricos realizados a la prueba departamental de morfología de la segunda lengua permitieron evidenciar la calidad del instrumento, lo que ofrece una seguridad al momento de tomar decisiones sobre el nivel de dominio que el estudiantado refleja en la prueba. Es un hecho conocido que la actividad evaluativa no se limita a un solo instrumento o a una sola acción, ya que la importancia de evaluar se ve completa al tomar acciones correctivas con base en los resultados obtenidos tanto en las actividades evaluativas diarias frente a clase, evaluaciones parciales y evaluaciones semestrales como en el caso de los exámenes departamentales. Dicho de otra manera, la evaluación debe complementarse entre las que son de tipo formativo y las que son de tipo sumativa, que en el caso de esta investigación se habla de una evaluación de carácter más sumativo. Sin embargo, no hay duda de que un instrumento de esta naturaleza permite apreciar de una manera valida y confiable el nivel de dominio del estudiantado al momento de ejecutar la prueba, claro está todo esto dentro del marco de lo establecido en el currículo de la asignatura.

Como se ha expresado anteriormente, por medio de este tipo de evaluación se trata de comprobar el nivel de dominio alcanzado en forma individual por el estudiantado; esto desde luego no debe significar un conocimiento totalmente nuevo para las personas implicadas hablando del personal docente y el alumnado, puesto que previo a la realización de la prueba departamental existe ya un sustento del nivel de dominio alcanzado por cada alumno y alumna como resultado de

https://doi.org/10.15359/ree.27-3.17218
https://www.revistas.una.ac.cr/index.php/educare
educare@una.ac.cr

todas las actividades evaluativas de carácter formativo que se han realizado durante el semestre. Lo anterior permitiría pronosticar un posible resultado de la evaluación sumativa en relación con lo logrado con las formativas, significando así que ambos resultados tendrían que mantener una similitud. Si lo antes mencionado no se presentara, se hablaría entonces de una posible deficiencia en la calidad de los procedimientos o instrumentos evaluativos empleados de manera formativa, situación que considerando el hecho de que este tipo de evaluaciones en la mayoría de los casos no emplean una metodología que de sustento a la calidad de las mismas.

La conclusión anterior tendrá mayor validez si la evaluación sumativa a diferencia de las formativas, si posee un fundamento sólido establecido por una metodología rigurosa que permita comprobar la calidad de la prueba en término psicométricos, es decir, que realmente evalúe aquello que se supone debe evaluar. Esto último nuevamente pone la atención en la importancia de que este tipo de pruebas cumpla con las normas establecidas de calidad para considerarlas válidas y confiables, permitiendo hacer comparaciones y valoraciones de la calidad de las evaluaciones formativas durante el semestre. Además, en el caso particular de la materia implicada en este examen departamental, al estar seriada con otra materia que también posee el mayor índice de relevancia curricular, el alcance de los resultados trasciende la asignatura, ya que la información proporcionada con esta prueba permitirá hacer estimaciones o pronosticar el desempeño en futuras asignaturas que requieran de conocimientos adquiridos en esta materia, y en especial cuando se trata de materias cuya relevancia curricular es sobresaliente.

Declaración de contribuciones

Las personas autoras declaran que han contribuido en los siguientes roles: **J. G. G. B.** contribuyó en la escritura del artículo, la validación y supervisión del proceso investigativo, la obtención de software, y la conceptualización, metodología y análisis de la investigación. **L. A. A. G.** contribuyó en la revisión-edición del artículo, la supervisión del proceso investigativo, y la conceptualización y análisis de la investigación.

Declaración de material complementario

Este artículo tiene disponible material complementario:

Preprint en https://doi.org/10.1590/SciELOPreprints.5126

Referencias

Aguilar-Salinas, W. E., & De las Fuentes Lara, M. (2023). Examen colegiado y predictores de éxito en los estudiantes de álgebra lineal. *Bolema: Boletim de Educação Matemática*, *37*(6), 797-822. http://dx.doi.org/10.1590/1980-4415v37n76a20

https://doi.org/10.15359/ree.27-3.17218
https://www.revistas.una.ac.cr/index.php/educare
educare@una.ac.cr

Aliaga Tovar, J. (2007). Psicometría: Tests psicométricos, confiabilidad y validez. *Psicología: Tópicos de Actualidad*, *8*, 85-108. https://www.academia.edu/download/33465691/CONFIABILIDAD_Y_VALIDEZ.pdf

Árraga Barrios, M. V. & Sánchez Villarroel, M. (2012). Validez y confiabilidad de la Escala de felicidad de Lima en adultos mayores venezolanos. *Universitas Psychologica, 11*(2), 381-393. https://doi.org/10.11144/Javeriana.upsy11-2.vcef

Aune, S. & Attorresi, H. F. (2019). Teoría de la respuesta al ítem: Su utilización en América Latina. Supuestos de unidimensionalidad e independencia local. XI *Congreso Internacional de Investigación y Práctica Profesional en Psicología. XXVI Jornadas de Investigación. XV Encuentro de Investigadores en Psicología del MERCOSUR. I Encuntro de Investigación de Terapia Ocupacional. I encuentro de Musicoterapia*. Universidad de Buenos Aires. https://www.aacademica.org/000-111/116

Backhoff Escudero, E. (2018). Evaluación estandarizada de logro educativo: Contribuciones y retos. *Revista Digital Universitaria*, *19*(6), 1-14. http://doi.org/10.22201/codeic.16076079e.2018.v19n6.a3

Backhoff Escudero, E., Larrazolo Reyna, N., & Rosas Morales, M. (2000). Nivel de dificultad y poder de discriminación del Examen de Habilidades y Conocimientos Básicos (EXHCOBA). *REDIE.Revista Electrónica de Investigación Educativa, 2*(1), 11-28. https://www.redalyc.org/articulo.oa?id=15502102

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

Carlos Martínez, E. A., Galván Parra, L. A., & Ruiz Moreno, R. (2011, 7-11 de noviembre). Análisis de las propiedades psicométricas de un examen de admisión para aspirantes a ingeniería. *XI Congreso Nacional de Investigación Educativa.* https://www.comie.org.mx/congreso/memoriaelectronica/v11/docs/area_01/1553.pdf

Centro Nacional de Evaluación para la Educación Superior (CENEVAL). (2017). *Origen y evolución del Ceneval*. https://docplayer.es/79887381-Origen-y-evolucion-del-ceneval.html

Contreras Niño, L. Á. (2000). *Desarrollo y pilotaje de un examen de español para la educación primaria en Baja California* [Tesis de Maestría, Universidad Autónoma de Baja California]. http://iide.ens.uabc.mx/documentos/divulgacion/tesis/MCE/1998/Luis_Angel_Contreras_Nino.pdf

Contreras Niño, L. Á. & Backhoff Escudero, E. (2004). Metodología para elaborar exámenes criteriales alineados al currículo. En S. Castañeda Figueiras (Ed.), *Educación, aprendizaje y cognición. Teoría en la práctica* (pp. 298-323). Manual Moderno.

https://doi.org/10.15359/ree.27-3.17218
https://www.revistas.una.ac.cr/index.php/educare
educare@una.ac.cr

Correa-Rojas, J. (2021). Coeficiente de correlación intraclase: Aplicaciones para estimar la estabilidad temporal de un instrumento de medida. *Ciencias Psicológicas*, *15*(2). https://doi.org/10.22235/cp.v15i2.2318

Cortada de Kohan, N. (2004). Teoría de respuesta al ítem: Supuestos básicos. *Revista Evaluar,4*(1), 95-110. https://doi.org/10.35670/1667-4545.v4.n1.600

Cuenca, A. A., Álvarez, M., Ontaneda, L. J., Ontaneda, E. A., & Ontaneda, S. E. (2021). La taxonomía de Bloom para la era digital: Actividades digitales docentes en octavo, noveno y décimo grado de Educación General Básica (EGB) en la habilidad de comprender. *Revista Espacios*, *42*(11), 11-25. 10.48082/espacios-a21v42n11p02

Fernández Martínez, M. A. (2013). Las pruebas estandarizadas y el diseño de la política educativa en México. *Este país,* (269), 34-36. https://biblat.unam.mx/es/revista/este-pais-mexico-d-f/articulo/las-pruebas-estandarizadas-y-el-diseno-de-la-politica-educativa-en-mexico

Fernández Navas, M., Alcaraz Salarirche, N., & Sola Fernández, M. (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

Gómez Rada, C. A. (2004). 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. https://www.redalyc.org/pdf/798/79801108.pdf

González Campos, J. A., & Aspeé Chacón, J. E. (2021). Propuesta de estimador de la fiabilidad mediante Alfa-Game: La significancia estadística del coeficiente de fiabilidad. *Revista Iberoamericana de Psicología, 14*(1), 1-10. https://doi.org/10.33881/2027-1786.rip.14101

Gutiérrez Benítez, J. G. & Acuña Gamboa, L. A. (2020). Evaluación estandarizada de los aprendizajes en la UABC: Innovación desde el análisis psicométrico. *Apertura*, *12*(1), 118-131. https://doi.org/10.32870/Ap.v12n1.1698

Gutiérrez Benítez, J. G. & Acuña Gamboa, L. A. (2022). Evaluación estandarizada de los aprendizajes: Una revisión sistemática de la literatura. *CPU-e, Revista de Investigación Educativa*, (34), 321-351. https://doi.org/10.25009/cpue.v0i34.2800

Hernández Madrigal, M., Ramírez Flores, É., & Gamboa Cerda, S. (2018). La implementación de una evaluación estandarizada en una institución de educación superior. *Innovación Educativa, 18*(76), 149-170. https://dialnet.unirioja.es/servlet/articulo?codigo=6791099

Hidalgo-Montesinos, M. D. & French, B. F. (2016). Una introducción didáctica a la teoría de respuesta al ítem para comprender la construcción de escalas. *Revista de Psicología Clínica con Niños y Adolescentes*, *3*(2), 13-21. https://www.redalyc.org/pdf/4771/477152554002.pdf

https://doi.org/10.15359/ree.27-3.17218
https://www.revistas.una.ac.cr/index.php/educare
educare@una.ac.cr

Hurtado Mondoñedo, L. L. (2018). Relación entre los índices de dificultad y discriminación. *Revista Digital de Investigación en Docencia Universitaria*, *12*(1), 273-300. https://doi.org/10.19083/ridu.12.614

Jornet Meliá, J. M. (2017). Evaluación estandarizada. *Revista Iberoamericana de Evaluación Educativa, 10*(1), 5-8. https://revistas.uam.es/riee/article/view/7590

Landis J. R. & Koch G. G. (1977) The measurement of observer agreement for categorical data. *Biometrics***,** *33*(1), 159-174. https://doi.org/10.2307/2529310

Mandeville, P. (2005). El coeficiente de correlación intraclase (ICC). *Ciencia UANL, 8*(3), 414-416. https://www.redalyc.org/pdf/402/40280322.pdf

Márquez Jiménez, A. (2014). Las pruebas estandarizadas en entredicho. *Perfiles Educativos, 36*(144), 3-9. https://www.redalyc.org/pdf/132/13230751001.pdf

Martínez Arias, M. R., Hernández Lloreda, M. J., & Hernández Lloreda, M. V. (2014). *Psicometría*. Alianza Editorial.

Medina Paredes, J., Ramírez Díaz, M. H., & Miranda, I. (2019). Validez y confiablidad de un test en línea sobre los fenómenos de reflexión y refracción del sonido. *Apertura*, *11*(2), 104-121. http://dx.doi.org/10.32870/Ap.v11n2.1622

Medrano, L. A, & Pérez, E. (2019). *Manual de psicometría y evaluación psicológica* (2.ª ed.). Editorial Brujas. https://www.researchgate.net/publication/351094332_Manual_de_Psicometria_y_Evaluacion_Psicologica

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

Nitko, A. J. (1994). A model for curriculum-driven criterion-referenced and norm-referenced national examination for certification and selection of students. Documento presentado en la *Conference of Education, Evaluation and Assessment for the Association Studies of Educational Evaluation in Sudafrica (ASEESA)*. Sudáfrica. https://eric.ed.gov/?id=ED377200

Ortiz Romero, G. M., Díaz Rojas, P. A., Llanos Domínguez, O. R., Pérez Pérez, S. M., & González Sapsin, K. (2015). Dificultad y discriminación de los ítems del examen de metodología de la investigación y estadística. *Edumecentro*, *7*(2), 19-35. https://pesquisa.bvsalud.org/portal/resource/pt/lil-738427

Oyarzún Maldonado, C., & Soto González, R. (2021). La improcedencia de estandarizar el trabajo docente: Un análisis desde Chile. *ALTERIDAD Revista de Educación*, *16*(1), 105-116. https://doi.org/10.17163/alt.v16n1.2021.08

https://doi.org/10.15359/ree.27-3.17218
https://www.revistas.una.ac.cr/index.php/educare
educare@una.ac.cr

Parra Giménez, F. J. (2017). La taxonomía de Bloom en el modelo Flipped Classroom. *Publicaciones Didácticas*, (86), 176-179. https://core.ac.uk/download/pdf/235855538.pdf

Pérez Tapia, J. H., Acuña Aguilar, N., & Arratia Cuela, E. R. (2008). Nivel de dificultad y poder de discriminación del tercer y quinto examen parcial de la cátedra de cito-histología 2007 de la carrera de medicina de la UMSA. *Cuadernos Hospital de Clínicas*, *53*(2), 16-22. http://www.scielo.org.bo/pdf/chc/v53n2/v53n2a03.pdf

Ravela, P. (2010). ¿Qué pueden aportar las evaluaciones estandarizadas a la evaluación en el aula? Programa de Promoción de la Reforma Educativa en América Latina y el Caribe (Serie Documentos). *Preal,* (47), 3-25. https://hdl.handle.net/20.500.12820/392

Reidl-Martínez, L. M. (2013). Confiabilidad en la medición. *Investigación en educación médica, 2*(6), 107-111. https://www.redalyc.org/pdf/3497/349733227007.pdf

Robles Pastor, B. F. (2018). Índice de validez de contenido: Coeficiente V de Aiken. *Pueblo Continente*, *29*(1), 193-197. https://upao.edu.pe/descargas/categoria/index.php?link=revista-pueblo-continente

Tristán López, A. & Pedraza Corpus, N. Y. (2017). La objetividad en las pruebas estandarizadas. *Revista Iberoamericana de evaluación educativa*, *10*(1), 11-31. https://doi.org/10.15366/riee2017.10.1.001

Viladrich, C., Angulo-Brunet, A., & Doval, E. (2017). Un viaje alrededor de alfa y omega para estimar la fiabilidad de consistencia interna. *Anales de Psicología, 33*(3), 755-782. http://dx.doi.org/10.6018/analesps.33.3.268401