Evaluación estandarizada del aprendizaje en la educación superior: un estudio de caso en México

Jorge Gutierrez, Luis Alan Acuña Gamboa

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

Envíado en: 2022-11-22
Postado en: 2022-11-25 (versión 1)
(AAAA-MM-DD)

Evaluación estandarizada del aprendizaje en la educación superior: un estudio de caso en México

Standardized Learning Assessment in higher education: a México case study

Avaliação estandardizada da aprendizagem na educação superior: estudo de caso no México

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

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

Financiamiento: autofinanciado.

Conflictos de interés: los autores declaran no presentar conflictos de interés.

Contribuciones de autoría

Jorge Gustavo Gutiérrez Benítez:

Visualización, revisión-edición, primer borrador, validación, supervisión, software, conceptualización, metodología, análisis.

Luis Alan Acuña Gamboa:

Revisión-edición, supervisión, conceptualización, análisis.

Resumen

La Universidad Autónoma de Baja California ha impulsado en los últimos años el desarrollo de pruebas departamentales estandarizadas con las cuales se pretende contribuir al cumplimiento de la meta de aplicar exámenes departamentales para mejorar continuamente los niveles de aprendizaje de los alumnos, prioritaria en el Plan de Desarrollo Institucional de esta universidad. Este trabajo de investigación presenta la estructura y composición de una prueba departamental estandarizada de una de las asignaturas de mayor relevancia curricular para el tronco común de Lenguas perteneciente a la Facultad de Idiomas, así como los resultados de la evaluación psicométrica realizada a la misma. Se empleó el método de análisis psicométrico basado en los procedimientos convencionales de la Teoría de Respuesta al Ítem, midiendo atributos tales como el índice de discriminación, coeficiente de discriminación y el índice de dificultad. Los resultados muestran que la prueba alcanzó un buen poder discriminatorio, así como una dificultad media, además destacaron los índices superiores al estándar obtenidos en la unidad temática de mayor relevancia curricular para la asignatura en cuestión. Los valores obtenidos en los atributos de confiabilidad y validez permitieron evidenciar la calidad del instrumento, contribuyendo a la certeza del nivel de dominio que los estudiantes reflejan en relación al universo de conocimientos representados en la prueba.

Palabras clave: Aprendizaje, estandarización, evaluación, exámenes, psicometría.

Abstract

In recent years the Autonomous University of Baja California has promoted standardized departmental tests development which are intended to contribute to the institution´s goal of applying departmental tests´ achievement to continuously improve student learning levels, a priority in the Institutional Development Plan of this University. This research paper presents the results of the psychometric evaluation carried out on a standardized departmental test of one of the subjects of outmost curricular relevance for the core curricula belonging to the Faculty of Languages, as well as the analysis of said test psychometric evaluation results. The psychometric analysis is based on the conventional procedures of the Item Response Theory measuring features such as discrimination index,

discrimination coefficient and difficulty index. The results show that the test achieved a good discriminatory range, as well as medium difficulty, and the above average indexes obtained in the thematic unit of outmost curricular relevance for the studied subject. The obtained values on reliability and validity features allowed to evidence the instrument´s quality, directly contributing to the domain level certainty that students show in relation to the acquired knowledge universe which the test presents.
Keywords: Evaluation, learning, psychometric, standardization, test.

Resumo

A Universidade Autônoma de Baixa California tem impulsado nos últimos anos o desenvolvimento das provas departamentais estandardizadas nas quais procura-se contribuir ao cumprimento da meta de aplicar exames departamentais para melhorar continuamente os níveis de aprendizagem dos alunos, prioritária no Plano de Desenvolvimento Institucional desta Universidade. Este trabalho de pesquisa apresenta a estrutura e composição duma prova departamental estandardizada de uma das disciplinas de maior relevância curricular para o tronco comum de Línguas que pertencem à Faculdade de Idiomas, assim como os resultados da avaliação psicométrica feita à mesma. Foi utilizado o método de análise psicométrico baseado nos procedimentos convencionais da Teoria de Resposta ao Item, medindo atributos tais como o índice de discriminação, coeficiente de discriminação e o índice de dificuldade. Os resultados mostram que a prova atingiu um bom grau de poder discriminatório, e também uma dificuldade media, alias, destacaram os índices superiores ao estândar obtidos na unidade temática de maior relevância curricular para as matérias envolvidas. 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: Aprendizagem, Estandardização, Avalição, Exames, Psicometria.

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 estudiante universitario. Varios autores (Fernández, 2013; Fernández, Alcaraz y Sola, 2017; Márquez, 2014) expresan que se optan por este tipo de pruebas con la intención de contar con instrumentos de evaluación válidos y confiables que permitan formar a los estudiantes con los atributos y características necesarias para responder a las demandas sociales y empresariales sobre el tipo de egresado universitario que se desea.

Este tipo de exámenes estandarizados han sido implementados con fines distintos, por ejemplo, siendo de carácter normativo en los filtros de selección de los aspirantes con mejor rendimiento, para ser admitidos a la universidad o promocionarse de grado, así como para pronosticar el desempeño académico futuro de los estudiantes (Hernández, Ramírez y Gamboa, 2018). Así mismo 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.

Con este tipo de iniciativas y procesos de evaluación estandarizada, las instituciones educativas buscan aumentar las probabilidades de éxito en los estudiantes, tanto durante todo su trayecto escolar, así como en cada una de las unidades que componen una asignatura en particular.

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, Ramírez y Gamboa, 2018). 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 y Pedraza, 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, logrando 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, 2017).

Este tipo de iniciativas a nivel nacional han motivado y propiciado que instituciones como la Universidad Autónoma de Baja California (UABC) desarrolle 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 Investigación y Desarrollo Educativo (IIDE), donde se invitó a participar a la mayor cantidad posible de Facultades de la UABC 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 de los alumnos.

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 cuáles son 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.

2 Referentes teóricos

2.1 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 y Pedraza, 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 (Jornet, 2017). La principal característica de las evaluaciones estandarizadas radica en los marcos de referencia teóricos y metodológicos rigurosos (Backhoff, 2018; Fernández, Alcaraz y Sola, 2017). Estos marcos de referencia permiten efectuar mediciones que dan como resultado valoraciones cuantificables de atributos asociados a la calidad de la prueba, como son la validez y la confiabilidad.

Una de los aportes de este tipo de evaluaciones es la posibilidad de tener un mayor acercamiento a la realidad, permitiendo 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 el proceso de construcción del mismo (Jornet, 2017). Este tipo de pruebas, cuando se alinean con objetivos de aprendizaje, suministran la base para procesos de retroalimentación que el docente puede utilizar para precisar fortalezas y debilidades curriculares, así como comprobar el logro alcanzado de manera individual en cada estudiante (Ravela, 2010).

2.2 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. Así mismo se le identifica como un campo metodológico que abarca aspectos como las teorías, métodos y usos de la medición psicológica. En el caso particular de la perspectiva teórica incluye las teorías que tratan de las medidas en psicología, proporcionado una descripción, categorías, evaluación de su utilidad y precisión, así como la investigación de nuevos métodos, teorías y modelos matemáticos

que provean de mejores instrumentos de medida. Una característica sobresaliente de la psicometría es el uso del lenguaje formal y estructurado de las matemáticas (Aliaga, 2007).

Dos atributos psicométricos de especial interés para esta investigación son la confiabilidad y la validez de la prueba. Por un lado, la confiabilidad tiene que ver con los errores generados producto de la medición, y busca responder al problema de con cuanta certeza las cantidades observadas reflejan la puntuación verdadera del sustentante (Martínez, Hernández y Hernández, 2014). En otras palabras, la confiabilidad puede explicarse si a falta de cualquier cambio permanente en un individuo o sustentante las calificaciones de una prueba varían considerablemente con el tiempo o en diferentes situaciones, lo anterior significaría que la prueba no es confiable, motivo por el cual no podría emplearse para explicar o predecir el comportamiento de quienes realizan la prueba (Árraga y Sánchez, 2012).

Junto a este atributo de confiabilidad se encuentra ligada la validez, ya que si bien, una prueba puede ser confiable, pero a la misma vez no ser válida. Lo anterior quiere decir que carecería de utilidad contar con un instrumento confiable, si el mismo no es el más apropiado para la medición o evaluación que se desea efectuar.

Formalmente la validez es un juicio evaluativo que implica una evidencia empírica y sustento teórico que permiten avalar la suficiencia y lo apropiado de las interpretaciones con base en los puntajes obtenidos en las pruebas, yendo más allá de solo limitarse a los ítems de la prueba, sino que también contempla el contexto en el que se desarrolla la prueba (Aliaga, 2007; Árraga y Sánchez, 2012).

2.3 Atributos de confiabilidad y validez en un instrumento

Este trabajo de investigación emplea la Teoría de Respuesta el Ítem 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. El desarrollo de la TRI surge como una necesidad de mejorar la teoría clásica de los test, buscando obtener una mejor medición de los atributos de interés.

La Teoría de Respuesta al Ítem pretende suministrar el fundamento probabilístico al problema de medir constructos latentes en especial aquellos que no son observables, considerando al ítem como la unidad básica de medición, permitiendo así observar el posible comportamiento de sustentante a un ítem en particular (Cortada, 2004). 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.

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 (Medina, Ramírez y Miranda, 2019). Croker y Algina (en Backhoff, Larrazolo y Rosas, 2000) mencionan que usualmente, a esta proporción se le denota con una *p*, la cual indica la dificultad del ítem. El cálculo de este atributo se realiza mediante la división del número de personas que contesto acertadamente el ítem entre el número de personas que en total contestaron el ítem.

Respecto al índice de discriminación de la prueba, este se entiende como la propiedad que tiene un ítem para poder separar a los sustentantes que tienen una mejor puntuación final en la prueba de aquellos que tienen una menor puntuación (Medina et al.,2019).

En relación al coeficiente de discriminación (Pérez, Acuna y Arratia, 2008), también conocido como el punto de correlación biserial (rpbis), es aquel atributo que permite medir con mayor certeza la discriminación de un ítem. Este atributo permite observar la correlación existente entre los puntajes obtenidos por los sujetos en un ítem especifico y el puntaje obtenido en su totalidad en la prueba. Así mismo da razón de la probabilidad existente de que un ítem sea acertado por aquellos sustentantes con mayor puntuación en la prueba.

En el caso de los atributos que permiten demostrar la confiabilidad de una prueba se encuentra el Alfa de Cronbach y el índice de confiabilidad de Kuder-Richardson (KR20), estos miden la consistencia interna de una prueba (Reidl, 2013). 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, este consiste en aplicar la prueba en diferentes momentos a la misma muestra de participantes, con el fin de detectar las fluctuaciones o variaciones presentes en los resultados obtenidos. Este método es efectivo porque permite detectar cuan estable es la medición con el paso del tiempo a pesar de los cambios que pudieran presentarse (Serra y Peña, 2006). 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 (Mandeville, 2005), dicho índice debe encontrarse lo más cercano a uno posible para ser considerado como de calidad.

3 Metodología

3.1 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 a 2019) con la cual el instrumento se ha ido calibrando hasta lograr lo mostrado en este artículo. Se contó con la participación de 260 alumnos provenientes de las cuatro sedes (Mexicali, Ensenada, Tijuana y Tecate). Todos estos cursan el primer semestre del tronco común de la Licenciatura en Idiomas; de esta manera, el examen se desarrolló para la asignatura Morfología de la segunda lengua, la cual está seriada con Morfosintaxis, ambas ponderadas con la más alta relevancia curricular en el nivel académico de los estudiantes. El género de la población fue homogéneo (50% para cada cual), de los cuales se eligieron a su vez un 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 alumno con alto rendimiento (mayores o iguales a 9). Lo anterior con el fin de contar con una muestra más representativa del universo de estudiantes que representa 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.

3.2 Instrumentos

Con relación al instrumento que se empleó en esta investigación y sobre el cual se efectuó la recopilación de información, se precisa que el examen departamental de Morfología está diseñado para evaluar el nivel de dominio que los estudiantes poseen en relación a todo el universo de conocimientos que abarca el currículo de dicha asignatura. Al ser una asignatura seriada, implica la adquisición de conocimientos clave que permiten concretar otros conocimientos, por ello la importancia de poder determinar el nivel de aprendizaje alcanzado por los alumnos y con ello hacer predicciones sobre el desempeño en futuras asignaturas.

El examen está compuesto por un total de 63 reactivos de opción múltiple. 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, picotemas etc. que tengan una mayor relevancia curricular para el logro de la competencia general del curso, así como aquellos temas que son críticos para la concreción de conocimientos en un futuro, dentro de la misma asignatura o bien para asignaturas futuras comprendidas en el programa de estudios.

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, etapa que entre otros procedimientos aplica un jueceo por parte de expertos entre los cuales se determina jerárquicamente cuáles contenidos son más importantes y por qué (se definen criterios de evaluación concretos), y posterior, mediante una serie de análisis matemáticos se obtiene el índice de relevancia curricular (IRC) de cada uno de los temas que componen toda la asignatura y la consistencia interna del jueceo para asegurar dicho proceso. En la siguiente tabla se muestran la distribución de ítems por cada contenido temático, agrupados por unidad los cuales cubren la totalidad de temas que el currículo de la asignatura contempla.

Tabla 1

Distribución de temas y reactivos del examen departamental de morfología de la segunda lengua.

Unidad Temática	Cantidad de temas que abarca	Número de reactivos
Unidad 1. Basic concepts.	7	13
Unidad 2. Rules of word formation.	9	22
Unidad 3. Open class words.	4	12
Unidad 4. Closed class words.	4	10
Unidad 5. Compound words, blends and phrasal words.	7	6
Totales	31	63

Fuente: elaboración propia

La distribución de reactivos corresponde al IRC de cada tema, por lo que los que posean mayor IRC tendrán mayor representatividad en el examen, es decir existirá un mayor número de ítems de esos contenidos temáticos.

Con relación a los análisis psicométricos de la dificultad y discriminación de la prueba fueron realizados mediante un programa de cómputo especializado llamado ITEMAN de la compañía Assessment Systems Corporation, así como mediante el software TAP publicado por Brooks, G. & Johanson, G. (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 uno de los 260 sustentantes y a su vez generó el archivo fuente para poder efectuar el análisis psicométrico 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 de los mismos. Un ejemplo del instrumento de recopilación de información obtenido de

la ejecución de la prueba y procesado por los softwares antes mencionados se puede observar en la siguiente figura:

Quick Item Analysis

Item	Key	Number Correct	Item Diff	Disc. Index	# Correct in High Grp	# Correct in Low Grp	Point Biser	Adj PtBis
Item 08	(2)	29	0.76	0.35	9 (0.90)	6 (0.55)	0.38	0.33
Item 09	(3)	19	0.50	0.14	5 (0.50)	4 (0.36)	0.22	0.15
Item 10	(3)	9	0.24	0.31	4 (0.40)	1 (0.09)	0.40	0.35
Item 11	(2)	24	0.63	0.55	10 (1.00)	5 (0.45)	0.34	0.28
Item 12	(1)	25	0.66	0.44	8 (0.80)	4 (0.36)	0.30	0.24
Item 13	(2)	36	0.95	0.18	10 (1.00)	9 (0.82)	0.16	0.12
Item 14	(2)	21	0.55	0.35	8 (0.80)	5 (0.45)	0.24	0.17
Item 15	(3)	36	0.95	0.09	10 (1.00)	10 (0.91)	0.19	0.16
Item 16	(3)	34	0.89	0.27	10 (1.00)	8 (0.73)	0.26	0.22
Item 17	(1)	17	0.45	0.52	7 (0.70)	2 (0.18)	0.43	0.37

Figura 1. Información psicométrica recopilada con el software TAP. Fuente: elaboración propia usando TAP versión 16.1.18.

4 Procedimiento

La metodología utilizada en la elaboración de este examen fue formulada en el IIDE (Contreras, 2000; Contreras y Backhoff, 2004) con base en el modelo psicométrico propuesto originalmente por Nitko (1994) para la elaboración de exámenes de gran escala de referencia criterial. Esto tipo de exámenes proporcionan resultados que describen el número de competencias que un estudiante domina, en relación al total de competencias evaluadas (Backoff, 2018). Estos exámenes están orientados por el currículo, esto 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 currículum
3) desarrollo de un plan de evaluación
4) producción y validación de ítems
5) análisis primario de resultados
6) análisis secundario de resultados

En relación a la etapa tres y cuatro, entre los principales productos derivados de estas, fue la tabla de especificaciones de ítems (consultar tabla 2) y el desarrollo de especificaciones de ítems (consultar figura 2). Con referencia a la tabla de especificaciones de ítems este proceso presenta de una manera sintética las decisiones de estrategia evaluativa del examen.

Tabla 2

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 conceptos
				Probar el dominio del concepto Morfología mediante la identificación de sus características.	OM	Comprender conceptos

Fuente: elaboración propia.

En la tabla 2 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, cual es el foco u objetivo que se busca alcanzar con ese ítem, que 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 2.

Formato para la especificación de ítems de la prueba de Morfología del segundo idioma

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 conocimiento	Dimensión proceso cognitivo					
	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*

Figura 2. Ejemplo de una especificación de ítems. Fuente: elaboración propia.

La etapa cuatro referente a producir y validar ítems es un trabajo técnico muy delicado y que requiere mucha atención, esto puesto que se busca mantener una 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, en cuyo caso contrario de no cumplirse se compromete de manera directa la calidad, confiabilidad y validez del examen. 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 examinados de la muestra se llevó a cabo mediante los procedimientos convencionales de la teoría de la respuesta al ítem (Hidalgo y French, 2016; Muñiz, 2010; Gómez, 2015).

En términos psicométricos la calidad se observa entre otras cosas con la medición del índice de dificultad, el índice de discriminación y el coeficiente de discriminación de un ítem, para obtener estos valores se requieren aplicar cálculos matemáticos específicos. Además de explicar brevemente cual es el concepto de estos atributos se describen también cuales son las operaciones necesarias que deben realizarse.

Índice de dificultad

La dificultad de un ítem representado normalmente con *p* hace referencia a la expresión numérica del grado en el que una pregunta resulta difícil de responder de manera correcta para el grupo al cual se aplica (Hurtado, 2018) y se da en el intervalo específico de 0 a 1, entre más cercano a cero sea este valor significa que la pregunta es más difícil de acertar. El cálculo de este atributo se realiza mediante la división del número de personas que contesto acertadamente el ítem (A) entre el número de personas que en total contestaron el ítem (N). La fórmula que expresa la operación anterior se representa de la siguiente manera:

$$p = \frac{A}{N}$$

Índice de discriminación

La discriminación de una pregunta expresa el grado en que una pregunta ayuda a ampliar las diferencias estimadas entre los que obtienen un puntaje total de la prueba relativamente alto en comparación con los que obtuvieron un puntaje relativamente bajo (Bazan, citado en Hurtado, 2018). Los valores posibles de este índice están dados en el intervalo de -1 a 1. Un ítem discrimina correctamente si permite diferenciar o separar entre los sujetos con puntajes altos y los sujetos con puntajes bajo en la prueba. Si lo anterior sucede, el valor obtenido tendrá que ser positivo. El cálculo necesario para la obtención de este atributo está dado por la siguiente formula:

$$D = \frac{Ga - Gb}{\frac{N}{2}}$$

Donde Ga es el número de aciertos en el ítem del 27% de los sustentantes con las puntuaciones más altas en el examen, Gb es el número de aciertos en el ítem del 27% de los sustentantes con las puntuaciones más bajas en el examen, y por último N es el número de sustentantes.

El coeficiente de discriminación

Este atributo es empleado para saber si los sustentantes idóneos son los que obtienen las respuestas correctas, conocer el poder predictivo que tiene el ítem de la relación acierto/error con la calificación total obtenida en una prueba (Backhoff, Larrazolo y Rosas, 2000). Dicho atributo puede ser obtenido a partir del siguiente calculo:

$$r_{pbis=} \frac{\bar{X}_1 - \bar{X}_0}{S_x} * \sqrt{\frac{n_1 n_0}{n(n-1)}}$$

Donde Ortiz, Díaz, Llanos, Pérez y González (2015) explican que $\bar{X}_1$es la media de las puntuaciones totales de quienes respondieron correctamente el ítem, $\bar{X}_0$ es la media de las puntuaciones totales de quienes respondieron incorrectamente el ítem, S_x es la desviación estándar de las puntuaciones totales, n_1 se refiere al número de casos que acertaron correctamente al ítem, n_0 son el número de casos que acertaron incorrectamente al ítem y n es la suma del total de casos de acierto y el total de casos de fallo.

Algunos autores (Carlos, Galván y Ruiz, 2011) concuerdan con la clasificación de la calidad del coeficiente de discriminación mostrada en la siguiente tabla:

Tabla 3.

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

Fuente: elaboración propia a partir de los datos de Carlos, et.al. 2011.

Según la tabla anterior, idealmente se deben poseer valores de rpbis mayores o iguales a .26 para considerarse de calidad, entre más cercano a uno sea este valor mayor será la calidad. Aunque los autores 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.

5 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 4.

Tabla 4.

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

Fuente: elaboración propia

Como se observa en la tabla, en términos generales 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, así mismo fue interesante el promedio del coeficiente de correlación punto biserial o coeficiente de discriminación, el cual se ubicó en .322, esto significa que el poder discriminatorio es bueno. En relación a la media de aciertos esta se ubicó en 43.263, significando lo anterior que en promedio los estudiantes obtienen una calificación de 70 puntos en una escala de 0 a 100. La distribución porcentual de los 63 ítems en las 5 unidades quedó como se ilustra a continuación en la tabla número 5.

Tabla 5.

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

Fuente: elaboración propia

La tabla 5 muestra como 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 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 estudiante 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 a esto fue relevante observar los valores obtenidos en los nueve temas que abarca la unidad dos, con especial interés 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 3.

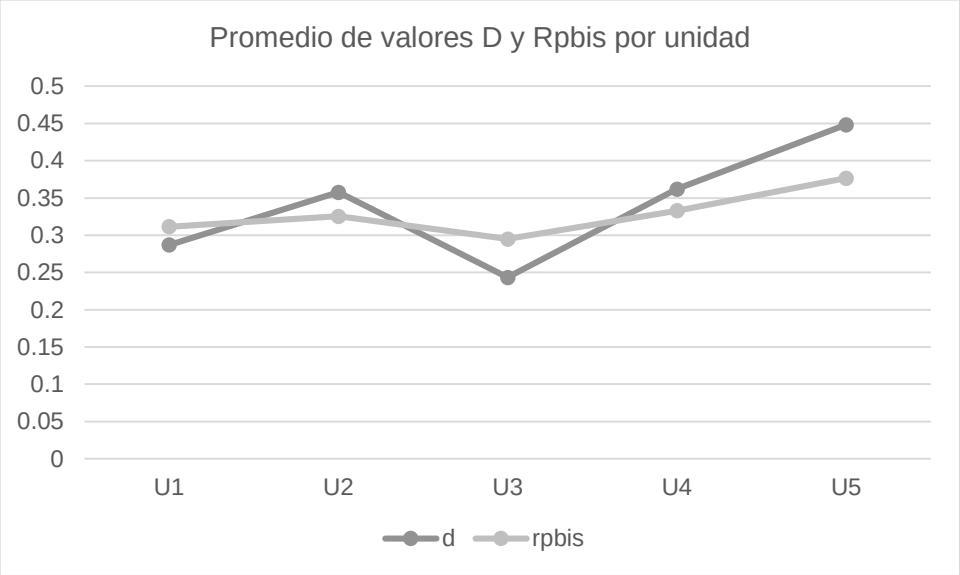

*Figura 3***.** Distribución de los valores promedio del índice y coeficiente de discriminación por unidad. Fuente: elaboración propia.

En la figura 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 la unidad dos 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 4, que si bien se identifica un número alto

de valores en el rango de .1 a .2 (14), solo 6 de ellos se encuentran por debajo de .19, lo que significa que tienen un margen de mejora positivo.

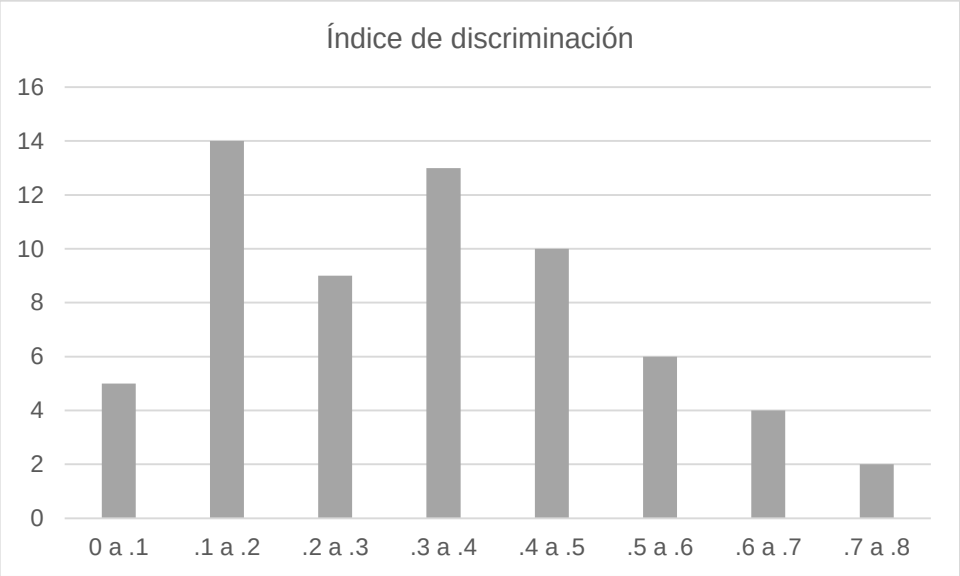

Figura 4. Distribución de los valores D. Fuente: elaboración propia.

La figura 4 también muestra como aproximadamente más del 80% de los ítems se encuentran distribuidos en el rango de .2 a .8 con relación al índice de discriminación. Lo anterior considerando que hay un total de 11 ítems que están por debajo de .18 en la calidad deseada, por lo que como se mostró anteriormente el promedio para el valor D en toda la prueba se ubicó en .331 muy cercano al criterio de excelencia.

Por su parte las distribuciones de los valores del coeficiente de discriminación en los 63 ítems se muestran en la figura 5.

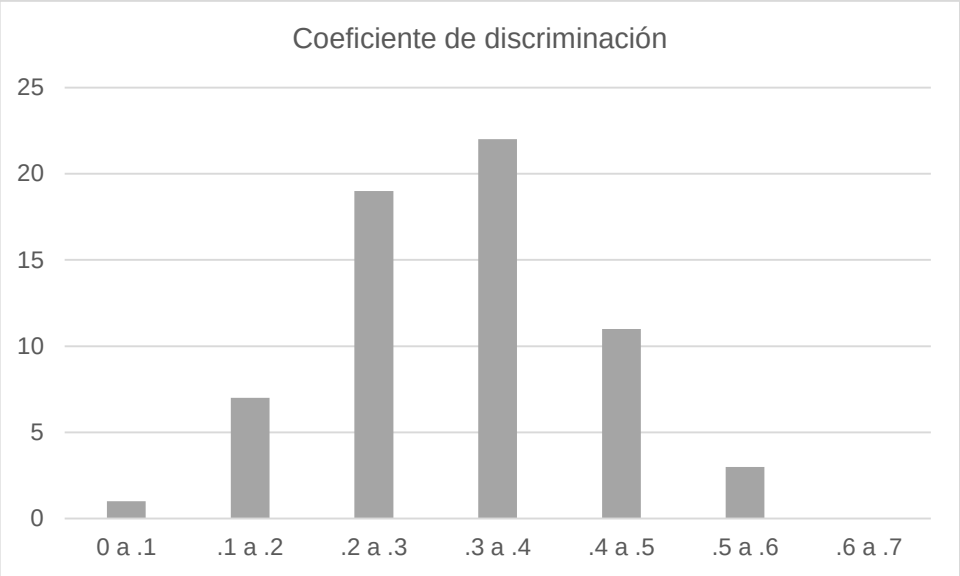

Figura 5. Distribución de los valores obtenidos para el coeficiente de discriminación (rpbis). Fuente: elaboración propia.

Como se logra apreciar en la figura anterior, de los 63 ítems que comprenden el examen, aproximadamente solo el 8% de ellos tiene un coeficiente de discriminación bajo (menor a .18), mientras que más del 70% de los ítems en la prueba 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.

Con relación al índice de dificultad, la distribución de la dificultad de cada uno de los ítems se muestra en la figura 6, en la cual se puede notar como el mayor porcentaje de ítems se encuentra en el rango de dificultad de .6 a .7. 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.

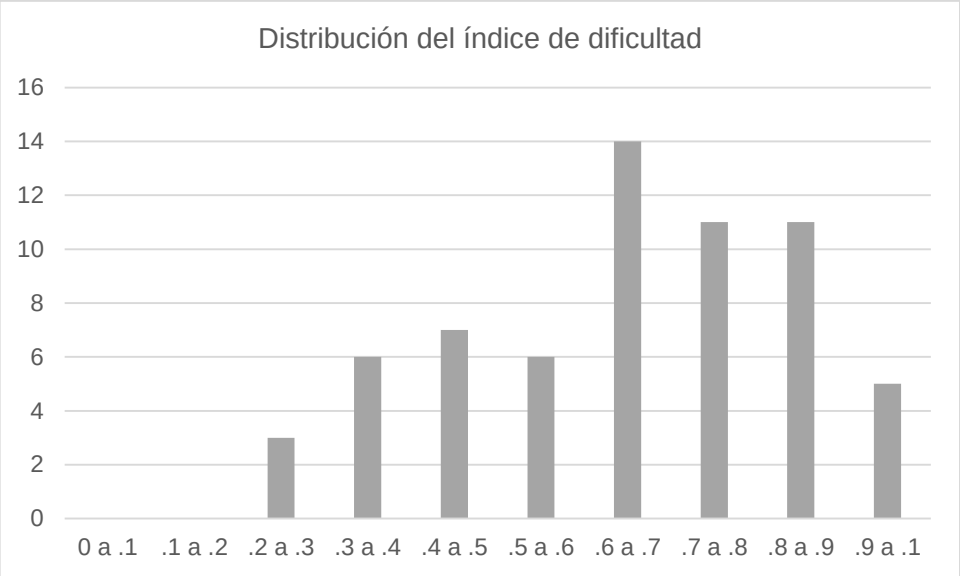

Figura 6. Distribución de los valores obtenidos para el índice de dificultad (p). Fuente: elaboración propia.

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 mayor proporción de ítems en la clasificación de fáciles a muy fáciles. Con relación a esta misma distribución de valores del índice de dificultad, se puede clasificar a los 63 ítems en la siguiente proporción:

Tabla 6

Clasificación de los ítems de la prueba según su dificultad

Dificultad	**Valor de *p***	**Porcentaje de ítems en la prueba**
Muy fáciles	Mayores a .9	8%
Fáciles	.75 a .9	17%
Dificultad Media	.45 a .75	54%
Difíciles	.2 a .45	16%
Muy difíciles	Menores a .2	5%

Fuente: elaboración propia

La tabla 6 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 autores (Backhoff et al, 2000, Ortiz 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.

6 Conclusiones

Es claro que para hacer estimaciones y una toma decisiones acertada 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 docente, hacer predicciones sobre rendimiento futuro entre otros.

Para poder determinar la calidad de una prueba se pueden realizar distintos procesos de valoración, entre ellos destacan los análisis psicométricos que se efectúan a cada uno de los ítems que componen la prueba. En razón de lo anterior se analizaron para la prueba departamental de Morfología en la segunda lengua el poder discriminatorio y el índice de dificultad de cada uno de los ítems que la integraron.

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%).

Sin embargo, se identificó un 15% 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 al coeficiente de discriminación la prueba mostro un promedio general de .322, este dato es muy importante porque a diferencia del índice de discriminación el coeficiente permite también medir el hecho de que los 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 los estudiantes reflejan 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 a 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 de un estudiante 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 estudiante; esto desde luego no debe significar un conocimiento totalmente nuevo para los implicados hablando del docente y el alumno, puesto que previo a la realización de la prueba departamental existe ya un sustento del nivel de dominio alcanzado por cada alumno como resultado de 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 a 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 desde luego 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 evalué 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 este materia, y en especial cuando se trata de materias cuya relevancia curricular es sobresaliente.

Referencias

Aliaga, J. (2007). Psicometría: tests psicométricos, confiabilidad y validez. *Psicología: Tópicos de actualidad*, *8*., pp. 85-108.

Árraga, M. y Sánchez, M. (2012). Validez y confiabilidad de la Escala de Felicidad de Lima en adultos mayores venezolanos. *Universitas Psychologica, 21*(2), 381-393.

Backhoff, E. (2018), “Evaluación estandarizada de logro educativo: contribuciones y retos”, *Revista Digital Universitaria*, *21*(6), pp. 1-14.

Backhoff, E., Larrazolo, N. y Rosas, M. (2000). Nivel de dificultad y poder de discriminación del Examen de Habilidades y Conocimientos Básicos

(EXHCOBA). *Revista Electrónica de Investigación Educativa, 2*(1), pp.11-28.

Brooks, G. y Johanson, G. (2003). TAP: Test Analysis Program. *Applied Psychological Measurement*, *27*(4), pp.303–304. https://doi.org/10.1177/0146621603027004007.

Carlos, E., Galvan, L. y Ruiz, R. (7-11 de noviembre de 2011). *Análisis de las propiedades psicométricas de un examen de admisión para aspirantes a ingeniería* [Presentación en papel]. XI Congreso Nacional de Investigación Educativa, México, D.F. Recuperado el 29 de abril de 2022, de https://www.comie.org.mx/congreso/memoriaelectronica/v11/docs/area_01/1553.pdf

Centro Nacional de Evaluación para la Educación Superior. (2017). Origen y evolución del Ceneval. Recuperado el 10 de octubre de 2020, de: http://www.ceneval.edu.mx/documents/20182/49855/OrigenEvolucionCeneval_2018/f8406659-7d28-4960-9ec1-964d90c76e4c

Contreras, L. (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 el 28 de septiembre de 2020, de: http://iide.ens.uabc.mx/documentos/divulgacion/tesis/MCE/1998/Luis_Angel_Contreras_Nino.pdf

Contreras, L. y Backhoff, E. (2004). Metodología para elaborar exámenes criteriales alineados al currículo. En: Castañeda, S. (Ed.), *Educación aprendizaje y cognición, teoría en la práctica*. México: Manual Moderno. ISBN 970 729 088 9.

Cortada, N. (2004). Teoría de respuesta al ítem: supuestos básicos. *Revista Evaluar, 4*(1), pp. 95-110.

Fernández, M. (2013). Las Pruebas Estandarizadas y el Diseño de la Política Educativa en México. *Este país,* (269), pp. 34-36.

Fernández, M., Alcaraz, N. y Sola, 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, C. (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), pp. 97-113.

Hernández, M., Ramírez, E. y Gamboa, S. (2018). La implementación de una evaluación estandarizada en una institución de educación superior. *Innovación educativa, 18*(76), pp.149-170.

Hidalgo, M. y French, B. (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), pp. 13-21.

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

Jornet, J. (2017). Evaluación estandarizada. *Revista Iberoamericana de Evaluación Educativa, 10*(1), pp. 5-8.

Mandeville, P. (2005). El coeficiente de correlación intraclase (ICC). *Ciencia UANL,* 8(3), pp. 414-416.

Márquez, A. (2014). Las pruebas estandarizadas en entredicho. *Perfiles educativos, 36*(144), pp. 3-9.

Martínez, M., Hernández, M. y Hernández, M. (2014). *Psicometría*. Alianza Editorial.

Medina, J., Ramírez, M. y Miranda, I. (2019). Validez y confiablidad de un test en línea sobre los fenómenos de reflexión y refracción del sonido. *Apertura: Revista de Innovación Educativa*, 11(2), pp.104-121.

Muñiz, J. (2010). Las teorías de los tests: teoría clásica y teoría de respuesta a los ítems. *Papeles del Psicólogo, 31*(1), pp.57-66.

Nitko, A. (1994). *A Model for Development Curriculum-Driven Criterion-Referenced and Norm-Referenced 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. Recuperado el 5 de octubre de 2020, de: https://eric.ed.gov/?id=ED377200

Ortiz, G., Díaz, P., Llanos, O., Pérez Pérez, Silvia María y González Sapsin, Kariné. (2015). Dificultad y discriminación de los ítems del examen de

Metodología de la Investigación y Estadística. *Revista Educación Médica del Centro*, *7*(2), pp.19-35.

Pérez, J., Acuna, N. y Arratia, E. (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), pp.16-22.

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. Preal. *Serie Documentos*, (47), pp.3-25.

Reidl, L. (2013). Confiabilidad en la medición. *Investigación en educación médica, 2*(6), pp.107-111.

Serra, A. y Peña, J. (2006). Fiabilidad test-retest e interevaluador del Test Barcelona. *Neurología*, 21(6), pp. 277-281.

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

UNIVERSIDAD AUTÓNOMA DE BAJA CALIFORNIA

FACULTAD DE IDIOMAS

"2022, año de la erradicación de la violencia contra las mujeres en Baja California"

MEMORÁNDUM

A QUIÉN CORRESPONDA:

Por medio del presente y de la manera más atenta el comité de ética de la Facultad de Idiomas extensión Tecate de la Universidad Autónoma de Baja California autoriza que el manuscrito "Evaluación estandarizada del aprendizaje en la educación superior: un estudio de caso en México" pueda ser objeto de publicación preprint en la biblioteca electrónica SciELO.

Es importante mencionar que se han atendido adecuadamente los intereses, la identidad y privacidad de los seres humanos involucrados en dicho manuscrito.

Sin otro particular por el momento, quedamos a sus apreciables órdenes agradeciéndole la atención prestada a este memorándum.

ATENTAMENTE
"POR LA REALIZACIÓN PLENA DEL SER"
Tecate, Baja California, a 24 de noviembre de 2022.

Comité de Ética de la Facultad de Idiomas Tecate

Dra. Myriam Romero Monteverde

Dra. Karina Olguin Jiménez

UNIVERSIDAD AUTONOMA DE BAJA CALIFORNIA
FACULTAD DE IDIOMAS TECATE

C.c.p. Archivo

Este preprint fue presentado bajo las siguientes condiciones:

- Los autores declaran que son conscientes de que son los únicos responsables del contenido del preprint y que el depósito en SciELO Preprints no significa ningún compromiso por parte de SciELO, excepto su preservación y difusión.
- Los autores declaran que se obtuvieron los términos necesarios del consentimiento libre e informado de los participantes o pacientes en la investigación y se describen en el manuscrito, cuando corresponde.
- Los autores declaran que la preparación del manuscrito siguió las normas éticas de comunicación científica.
- Los autores declaran que los datos, las aplicaciones y otros contenidos subyacentes al manuscrito están referenciados.
- El manuscrito depositado está en formato PDF.
- Los autores declaran que la investigación que dio origen al manuscrito siguió buenas prácticas éticas y que las aprobaciones necesarias de los comités de ética de investigación, cuando corresponda, se describen en el manuscrito.
- Los autores declaran que una vez que un manuscrito es postado en el servidor SciELO Preprints, sólo puede ser retirado mediante solicitud a la Secretaría Editorial deSciELO Preprints, que publicará un aviso de retracción en su lugar.
- Los autores aceptan que el manuscrito aprobado esté disponible bajo licencia Creative Commons CC-BY.
- El autor que presenta el manuscrito declara que las contribuciones de todos los autores y la declaración de conflicto de intereses se incluyen explícitamente y en secciones específicas del manuscrito.
- Los autores declaran que el manuscrito no fue depositado y/o previamente puesto a disposición en otro servidor de preprints o publicado en una revista.
- Si el manuscrito está siendo evaluado o siendo preparando para su publicación pero aún no ha sido publicado por una revista, los autores declaran que han recibido autorización de la revista para hacer este depósito.
- El autor que envía el manuscrito declara que todos los autores del mismo están de acuerdo con el envío a SciELO Preprints.