



Volumen 18, número 1, 2022

Exámenes departamentales. Instrumento de evaluación para una asignatura de la enseñanza de lenguas

José Manuel Casillas DomínguezⁱLilia Martínez LobatosⁱⁱLaura Emilia Fierro Lópezⁱⁱⁱ**Resumen**

La evaluación educativa es parte importante en todo programa de estudios debido a que permite conocer el grado en que los estudiantes van adquiriendo las competencias establecidas en las asignaturas y en el perfil profesional. Sin embargo, la evaluación se debe llevar a cabo con responsabilidad y debe tener un seguimiento. En este estudio, se analizaron los resultados del examen departamental de la asignatura Morfosintaxis de la Segunda Lengua para conocer el índice de homogeneidad entre los estudiantes que llevaron la materia de Morfosintaxis de la Segunda Lengua con diferentes docentes. Para conocer el índice de homogeneidad se usó el método cuantitativo. Primero, se analizaron los resultados del examen con apoyo del programa Excel y se obtuvieron las medidas de tendencia central como la media y mediana. Por otra parte, se analizó el índice de dispersión de los resultados aplicando fórmulas para identificar la desviación estándar y la correlación de Pearson. No obstante, al revisar los resultados del examen surgieron dudas sobre la calidad técnica del examen al encontrar que todos los grupos evaluados obtuvieron una media general baja. Por tal motivo se decidió usar el *Test Analysis Program* (TAP) para identificar el índice de dificultad y la capacidad de discriminar de cada uno de los reactivos. Como resultado de estos análisis se encontró que existe cierto índice de homogeneidad. Sin embargo, el examen es considerado difícil debido a que cuenta con una media general baja y por el número de ítems en el examen considerados como difíciles.

Palabras clave: *Exámenes estandarizados, evaluación, medición, análisis, currículum.*

*Department exam. Assessment instrument for a language teaching subject***Abstract**

Assessment is an important part of any educational program. It lets us know the degree to which students are acquiring the competencies established in each class and in an educational program. None the less, assessment must be done with responsibility and it must have a follow up. In this study, we analyzed the results of the department test for students who take the class named, "Morfosintaxis de la Segunda Lengua" to know the degree of homogeneity amongst students taking the same class with different teachers. To know the degree of homogeneity we used the quantitative method. We first analyzed the results of the test with Excel in order to know the mean, standard deviation and the Pearson correlation. None the less, as we analyzed the results from the exam, we started having doubts on the technical quality of the exam because all the groups that were evaluated obtained a low general mean. Because of this, we decided to use the program called, "Test Analysis Program (TAP)" to identify the item difficulty and the item discrimination in every item. As a result, of this analysis we found evidence of homogeneity, but the test was considered difficult because of its mean which is low and because of the number of items that were considered to be difficult.

Key words: *Standardized exams, testing, measurement, analysis, curriculum*

Introducción

Los procesos de evaluación de aprendizajes desde perspectivas estandarizadas han mostrado ventajas y desventajas, así como muy diversas discusiones en las instituciones. Se presenta como una alternativa en la educación superior para verificar el logro de aprendizajes, principalmente en aquellos planes de estudio diseñados bajo competencias centrados en alcanzar aprendizajes esperados de manera estandarizada. Se podría considerar como una opción para los momentos o etapas iniciales de la formación profesional que se caracterizan por poseer troncos o asignaturas comunes a carreras diferentes o bien a asignaturas comunes entre unidades académicas. Desde la gestión curricular facilita la organización del trabajo académico de los profesores, el logro de aprendizajes esperados al uniformar evidencias en las competencias.

Entre las principales limitaciones que muestra la evaluación departamental se podría mencionar a la imposibilidad de generar trabajo académico desde perspectivas de construcción divergente, en las cuales, podría verse obstaculizada la implementación de didácticas que favorezcan el uso de situaciones auténticas o en sitio. Otro aspecto importante se encuentra en que podría perderse la individualidad del estudiante ya que se parte de suponer que todos muestran las mismas características y que, por lo tanto, todos pueden ser objeto del mismo examen.

La evaluación estandarizada o departamental de acuerdo con Linn (2000) tiene antecedentes desde la década de los cincuenta mediante diversos desarrollos donde impactó a las políticas de evaluación y las instituciones. No obstante, se reconoce que a partir de la década de los noventa se han manifestado algunas experiencias donde los exámenes departamentales han cobrado importancia. Hernández, Ramírez y Gamboa, (2018) señalan que la evaluación estandarizada permite una primera aproximación hacia una homologación de una variedad de experiencias de aprendizaje de los estudiantes en una misma disciplina.

La Universidad Autónoma de Baja California, UABC, considera entre las evaluaciones de carácter institucional aquellas que podrá aplicar periódicamente y que revelen el grado de aprendizaje de los alumnos inscritos en un programa educativo para valorar los resultados del proceso educativo. Establece en el Estatuto Escolar (2018), Artículo 78, para los programas de licenciatura a los exámenes departamentales. Las unidades académicas tendrán la atribución de diseñar los exámenes parciales, ordinarios y extraordinarios para que sean aplicados de manera

colegiada bajo la modalidad de un examen departamental. Los cuales de acuerdo al artículo 80, tendrán los siguientes objetivos específicos: I. Conocer el grado de dominio que el alumno ha obtenido sobre la unidad de aprendizaje que cursa; II. Verificar el grado de avance del programa de la unidad de aprendizaje de conformidad con lo establecido en el presente estatuto, y III. Conocer el grado de homogeneidad de los aprendizajes logrados por los alumnos de la misma unidad de aprendizaje que recibieron el curso con distintos profesores.

Serán los profesores de las unidades académicas que compartan asignaturas en común los encargados del diseño de los exámenes departamentales, los cuales se harán de manera simultánea a todos los alumnos inscritos en la misma unidad de aprendizaje, bajo la supervisión del órgano colegiado de profesores del área. El resultado obtenido en la evaluación incidirá en un porcentaje de la calificación que corresponda al alumno. El mencionado porcentaje será determinado por la unidad académica, sin que pueda exceder de cincuenta por ciento. El profesor podrá extenderlo hasta cien por ciento.

Marco teórico

Se inicia este apartado con las definiciones que se consideran esenciales en este trabajo que son:

La evaluación

La evaluación, en términos generales, es formar un juicio sobre el desempeño de una actividad basado en lo que percibimos. Usualmente nuestros juicios son acompañados de decisiones que se basan en dichos juicios y es en estos momentos que debemos asegurar que nuestros juicios tengan un buen fundamento. De acuerdo con Pérez (2014),

La evaluación, por tanto, se caracteriza por ser un proceso que implica la recopilación de información. Posteriormente se interpreta y contrasta con determinadas instancias de referencia o patrones de deseabilidad que permita emitir un juicio de valor para orientar acciones y tomar decisiones (p. 16).

La descripción sobre evaluación que nos brinda Pérez incluye procesos, recopilación de información, así como interpretación, entre otros elementos que nos ayudan a tomar decisiones en nuestro día a día.

La evaluación educativa

La evaluación dentro del ámbito de la educación es conocido como la evaluación educativa. De acuerdo con Airasian (1996), la evaluación educativa es un proceso en el cual se recolecta información para analizar e interpretar para finalmente tomar decisiones. Como se puede observar, esta definición comparte varios elementos importantes con la definición que presenta Pérez sobre evaluación lo cual hace difícil distinguir entre evaluación y evaluación educativa.

Por otra parte, Sandoval, Maldonado-Fuentes y Tapia-Landino (2022) nos dan una definición más orientada hacia la educación:

...la evaluación puede ser entendida como un juicio que otorga valor a una cosa, hecho o fenómeno a partir de información obtenida rigurosamente, cuya finalidad es tomar decisiones pedagógicas (mejorar el proceso enseñanza y aprendizaje) o sociales (la institución o el sistema). De esta forma, se desprende que evaluar en Educación tiene una razón de ser particular: ser una tarea inherente al proceso de enseñanza y de aprendizaje para el mejoramiento permanente de los procesos formativos en sus distintos niveles (pp. 56-57).

Esta definición también comparte varios elementos encontrados en las definiciones de Pérez y Airasian, sin embargo, enfocan todos los esfuerzos (análisis, interpretación y toma de decisiones) hacia los aspectos pedagógicos y los procesos de enseñanza aprendizaje, motivo por el cual en este trabajo se adoptó esta definición de evaluación educativa.

El examen

La evaluación se apoya en diferentes instrumentos y herramientas que brindan una mayor objetividad a las decisiones que surgen de dicho proceso. Entre dichos instrumentos, el examen suele ser uno de los instrumentos más usados en la evaluación. Según Saavedra (2001), “históricamente el examen aparece en la institución escolar con la finalidad de determinar el rendimiento de los alumnos y de seleccionar los (rendimientos-selección) para ocupar las distintas funciones de la división del trabajo” (p. 2).

Por otra parte, Haladyna, (2004) define el examen como un instrumento de evaluación que se aplica siguiendo un protocolo de aplicación y tiene como objetivo describir los aprendizajes logrados por los evaluados.

Exámenes referidos al criterio o a la norma

Una vez que se presentan los exámenes los resultados de acuerdo con Fein (2012) pueden ser interpretados dependiendo si son exámenes referidos al criterio o a la norma. Cuando los resultados son referidos a la norma, se comparan los resultados de un evaluado con los resultados del resto de los evaluados en el mismo periodo y con el mismo examen. Cuando los exámenes son referidos al criterio los resultados de los evaluados son comparados con un estándar previamente establecido. La mayoría de los exámenes que se aplican en la escuela por ejemplo son referidos al criterio debido a que las calificaciones van del 0 al 10 o del 0 al 100. En estos exámenes, si un evaluado obtiene un 4 en la escala del 0 al 10, el alumno habrá obtenido una calificación reprobatoria. Si fuera un examen referido a la norma y el cuatro es la puntuación más alta en el examen es probable que el cuatro no necesariamente sea una calificación reprobatoria.

En el caso del examen departamental de Morfosintaxis de la Segunda Lengua, es un examen referido a la norma debido a que las puntuaciones obtenidas por los evaluados son comparadas con las evaluaciones del resto de los evaluados en ese mismo periodo.

Exámenes de alto impacto

Los exámenes de alto impacto se caracterizan por tener un impacto directo al alumno ya sea positivo o negativo. De acuerdo con Mendoza (2015), “un examen de alto impacto es aquel cuyo resultado es empleado para tomar decisiones importantes que conllevan consecuencias positivas o negativas en quienes los sustentan” (p. 170). Algunas de las consecuencias negativas de este tipo de exámenes pueden ser reprobación de una asignatura, no obtener un certificado, no entrar a la universidad deseada, etc.

Todo examen de alto impacto debe ser sometido a una serie de análisis antes de asignarle un valor. De acuerdo con Mendoza (2015), “debido a que los exámenes de alto impacto juegan un papel primordial en la vida de los examinados, es imprescindible que tales exámenes evalúen de manera justa y equitativa a los candidatos que cada año los realizan” (p. 170).

Ítem

El ítem es una unidad de medida dentro del examen en forma de pregunta o enunciado imperativo que solicite al evaluado realizar una tarea con la cual se demostrará el dominio del contenido o habilidad del evaluado. De acuerdo con Osterlind (1998) algunos de los formatos más comunes que puede tener un ítem son preguntas de opción múltiple, relación de columnas y completar enunciados.

Calidad técnica de los exámenes

Al diseñar un examen es importante realizar estudios para determinar la calidad técnica del mismo. Esto se lleva a través de estudios realizados a las respuestas del examen en general y a las respuestas de cada uno de los reactivos. Evers et al. (2013) mencionan que, así como en cualquier otra área científica, la calidad métrica de un examen es necesaria para lograr resultados eficientes.

Por otra parte, Adom, Adu y Atsu (2020) dicen que el enfoque estadístico involucra el uso de análisis estadístico para determinar qué tan bueno es un ítem en términos de dificultad, capacidad para discriminar entre los evaluados y el estado de sus distractores (p. 115).

Aunado a esto, la universidad de Washington (2022) menciona que el análisis de los ítems es importante para mejorar los ítems que serán reutilizados en exámenes futuros, pero también puede ser usado para eliminar ítems ambiguos o que confunden a los evaluados.

Finalmente, tomando en cuenta la información presentada en este apartado sobre evaluación, exámenes, cómo se interpretan los resultados y la calidad técnica de los mismos, se realizaron diferentes estudios al examen departamental, Morfosintaxis de la Segunda Lengua, un examen referenciado en la norma y de alto impacto.

Aspectos teóricos

Para determinar el grado de homogeneidad entre docentes que imparten la misma asignatura, el análisis se centró en medidas de tendencia central como lo son la media y mediana. También se analizó el índice de dispersión de los resultados aplicando fórmulas para identificar la desviación estándar. Finalmente, se apoyó en la correlación de Pearson para identificar si existía correlación entre los resultados de los exámenes y las calificaciones que obtuvieron los estudiantes al finalizar la asignatura.

Por otra parte, con el apoyo de un programa diseñado para evaluar los ítems de un examen, se analizó el índice de dificultad y la capacidad para discriminar de cada uno de los ítems. El índice de dificultad de acuerdo con Fein (2012) es la proporción de los examinados que contestan el ítem de manera correcta.

Finalmente, se analizó el índice de discriminación. De acuerdo con Haladyna (2004) los evaluados con mayores aciertos en un examen tienden a elegir la respuesta correcta de un ítem y los evaluados con el menor número de aciertos en un examen tienden a seleccionar una respuesta errónea.

La media, según Hernández, Fernández y Baptista (2010), es “la medida de tendencia central más utilizada” (p.293). Por otra parte, Guàrdia, Freixa, Pero y Turbany (2007) mencionan que esta medida se obtiene al sumar todos los resultados de los examinados y dividirlos por el número de evaluados.

La mediana es descrita por Guàrdia, Freixa, Pero y Turbany (2007), como “el punto de distribución que deja el mismo número de casos a cada lado de la distribución” (p.51). Esto se logra organizando los resultados de un examen de manera descendiente o ascendiente. Después, se identifica el resultado que se encuentra justo en el centro.

Es importante mencionar que cuando el número de evaluados es par, se debe de sumar los dos resultados que se encuentran en el centro y dividirlos entre dos para obtener la mediana. Por otra parte, se calculó la desviación estándar. Esta medida es definida por Olmos, Blanchart, Cebollero y Oset, (2012) como “la raíz cuadrada positiva de la varianza (como parámetro se representa por σ) y pretende resolverla dificultad que la varianza tiene al ofrecer valores en unidades cuadráticas” (p. 54).

Asimismo, se realizó un estudio con el apoyo de la correlación de Pearson para conocer si existe relación entre la calificación obtenida en el examen y la calificación final obtenida en la asignatura. De acuerdo con Restrepo y González, (2007) la Correlación de Pearson “Tiene como objetivo medir la fuerza o grado de asociación entre dos variables aleatorias cuantitativas que poseen una distribución normal bivariada conjunta (p.185)”

Por otra parte, para conocer la calidad de los ítems se analizó el índice de dificultad y la capacidad para discriminar de cada uno de los ítems. El índice de dificultad representado por la

letra “P” se define en términos del número de personas que contestan correctamente un reactivo. Entre más sencillo sea el reactivo más elevado será el porcentaje (Anastasia y Urbina, 1998, p.173).

En cuanto al índice de discriminación, Aiken (2003) comenta que ayuda a identificar el número de alumnos con mayores promedios que aciertan un ítem y al mismo tiempo; el número de alumnos con los promedios más bajos que acertaron el mismo ítem. Esta medida nos ayuda a conocer el comportamiento de un ítem y sus distractores, su definición es la siguiente:

El índice de discriminación del reactivo (D) es una medida de la eficacia de un reactivo para discriminar entre quienes obtienen altas y bajas calificaciones en una prueba. Mientras más elevado sea el valor de D, resulta más eficaz para establecer dicha distinción. Cuando (D) es igual a 1.00, todos los examinados del grupo superior y ninguno del grupo inferior en las calificaciones totales de la prueba respondieron el reactivo en forma adecuada. (Aiken, 2003, p. 66).

Los exámenes estandarizados deben ser elaborados con gran responsabilidad, así como ser sometidos a diferentes estudios como los mencionados en esta sección para asegurar que las conclusiones derivadas de los resultados obtenidos por los estudiantes sean confiables.

Estrategia metodológica

La Facultad de Idiomas tiene como misión formar a profesionistas competentes en la enseñanza de lenguas y de traducción. Con el objetivo de dar seguimiento a las asignaturas con mayor índice de reprobación y asegurar la calidad de los programas de licenciatura que ofrece la Facultad de Idiomas, se identifica a la asignatura de Morfosintaxis de la Segunda Lengua como una de las asignaturas con el mayor índice de reprobación. Esta asignatura es de carácter obligatorio y se ofrece en el segundo semestre de Tronco Común, de las carreras ya mencionadas.

Para mejorar el índice de aprobación de la asignatura antes mencionada, en el año 2017, se aplica por primera vez un examen departamental como parte de su sistema de evaluación interna. En este apartado se detalla el procedimiento que se siguió para analizar los resultados del examen:

Población

El examen departamental de la asignatura de Morfosintaxis de la Segunda Lengua fue aplicado a todos los estudiantes de segundo semestre del Tronco Común de la Facultad de Idiomas, campus Mexicali. En esta evaluación participaron 74 estudiantes pertenecientes a cuatro

grupos que recién habían concluido la asignatura de Morfosintaxis de la Segunda Lengua. Después, se analizaron los resultados de los exámenes con los estudios que se describen en el apartado de diseño de la investigación.

Diseño de la investigación

En este trabajo se usó el método cuantitativo para analizar el examen y los resultados obtenidos por los evaluados. Al analizar el examen se encontró que cuenta con 61 ítems de opción múltiple. Los ítems cubren las cuatro unidades de la asignatura y los 26 temas que la conforman. Cada uno de los ítems cuenta con tres posibles respuestas.

Por otra parte, con el apoyo del programa Excel se realizaron las medidas de tendencia central conocidas como la mediana y media. La mediana se identificó organizando las calificaciones de los evaluados y organizándolas de manera descendente. Después se identificó la calificación que se encontraba justo en el centro del listado con las calificaciones, esta calificación es conocida como la mediana. Esta medida ayudó a identificar, más adelante, la dispersión de las calificaciones.

El programa Excel también ayudó a obtener la medida de tendencia central conocida como la media. Esta medida se obtuvo sumando las 74 calificaciones obtenidas por los evaluados y dividiéndolas entre 74. Este número permitió conocer el promedio general obtenido por los estudiantes en cada uno de los grupos e identificar el índice de homogeneidad. Más tarde se usó esta medida para obtener la desviación estándar de cada uno de los grupos evaluados.

Después, se realizaron análisis con las medidas de dispersión conocidas como la desviación estándar y la correlación de Pearson. La desviación estándar es otra medida aritmética que se considera importante al momento de analizar un examen. Esta medida ayudó a conocer el rango aceptable de las calificaciones obtenidas del examen departamental.

La desviación estándar de cada uno de los grupos se obtuvo de la siguiente forma. Primero se restó la media a cada una de las calificaciones obtenidas por los evaluados de un grupo. El siguiente paso fue obtener la raíz cuadrada de cada una de las respuestas obtenidas. Después, se sumó cada una de las raíces cuadradas y se dividió el resultado de estas sumas entre el número de evaluados. Finalmente, se encontró la raíz cuadrada del resultado obtenido en el paso anterior y se obtuvo la desviación estándar. Este procedimiento se repitió para cada uno de los grupos.

Por otra parte, con el objetivo de conocer mejor al examen de Morfosintaxis de la Segunda Lengua, se llevó a cabo un análisis de la desviación estándar donde se incluyeron los resultados de todos los evaluados de los cuatro grupos. Para realizar este ejercicio se siguió el mismo procedimiento explicado en el párrafo anterior.

Con el objetivo de conocer la relación que existe entre las calificaciones obtenidas en el examen departamental y las calificaciones que los evaluados obtuvieron en la asignatura de Morfosintaxis de la Segunda Lengua se realizó un estudio de la correlación de Pearson. Para obtener dicha correlación de un grupo, en una tabla Excel se colocaron las calificaciones que los evaluados obtuvieron en el examen en una columna y a la derecha de dicha columna se colocaron las calificaciones que los alumnos obtuvieron como calificación final en la asignatura antes mencionada. Después se ordenó al sistema de Excel que realizara la correlación de Pearson entre ambas columnas.

Por último, para analizar los ítems, se usó el programa *Test Analysis Program* (TAP), diseñado por Gordon P. Brooks. En este programa se vaciaron las respuestas y después se alimentó el programa con las respuestas de los evaluados. El siguiente paso fue correr el sistema con la información antes mencionada y se obtuvo una serie de análisis de los cuales destacan el índice de dificultad y la capacidad para discriminar de cada uno de los reactivos.

El índice de dificultad permitió identificar el número de evaluados que contestó de manera acertada cada reactivo. Esto ayudó a identificar reactivos que debían ser revisados y o modificados. De acuerdo con Ebel y Frisbie (1991), un ítem que sea contestado de manera correcta por todos los examinados, o un ítem que todos contesten de manera incorrecta, no contribuye al análisis del examen.

Por otra parte, se realizó un análisis de la distribución de los ítems de acuerdo a su índice de dificultad. Para este análisis se consideró la distribución propuesta por Backhoff, Larrazolo, y Rosas (2000). En esta propuesta sugieren que 5% de los ítems de un examen deben ser considerados fáciles, 20% medianamente fáciles, 50% con dificultad media, 20% medianamente difíciles y 5% de los ítems deben ser considerados difíciles. Para lograr esto, se elaboró una tabla con los porcentajes propuestos ya mencionados y una propuesta de distribución de índices que se adaptara con los porcentajes.

Capacidad para discriminar

El último análisis que se realizó al examen departamental fue para identificar la capacidad de cada ítem para discriminar entre los alumnos que obtuvieron las puntuaciones más altas y los que obtuvieron las calificaciones más bajas. De acuerdo a Fein (2012). En un ítem que funciona bien, proporcionalmente habrá más sujetos con las puntuaciones más altas que los sujetos con las puntuaciones más bajas (p. 151).

En este trabajo, se analizó la distribución de los ítems de acuerdo a su capacidad para discriminar. En este caso, se siguió la distribución recomendada por Ebel y Frisbie (1986) mismos que nos dan parámetros para poder clasificar a los ítems de acuerdo al grado de discriminación obtenido. Ellos comentan que un excelente ítem es aquel que cuenta con un grado de discriminación de .40 o más, un buen ítem cuenta con un grado de discriminación de entre .30 y .39, uno regular va entre .20 y .29; uno pobre entre .0 y .19; uno pésimo entre -.1 o menor. Ebel y Frisbie (1986) también comentan que aquellos ítems con un índice de discriminación de .19 o menor deben ser mejorados o rechazados.

Resultados

Grado de dominio

Para analizar el grado de dominio de los estudiantes sobre la asignatura, se identificó la calificación obtenida en el examen departamental de cada alumno, también se identificó la calificación final obtenida por los alumnos en la asignatura. Después, se compararon los resultados del examen y las calificaciones obtenidas por los estudiantes en la asignatura.

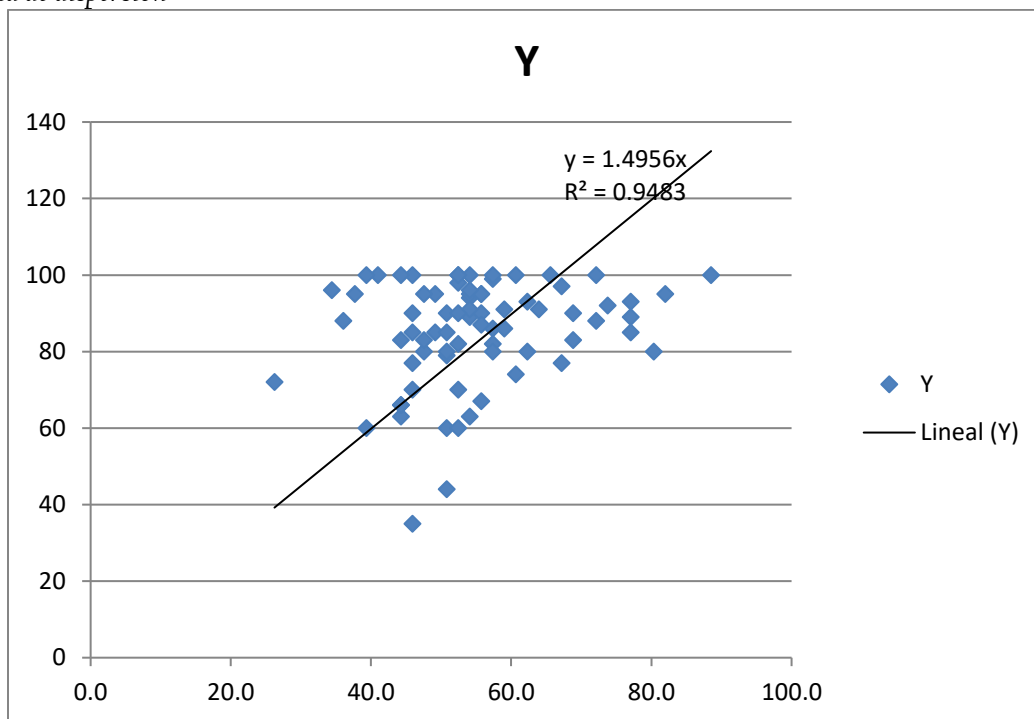
Al analizar los resultados del examen y los resultados de la asignatura se encontró que muchas de las calificaciones no coincidían. Por ejemplo, los evaluados obtuvieron en el examen un promedio de 55% y en la asignatura un promedio de 85%. También se encontró que muchos alumnos obtuvieron calificaciones altas en el examen y calificaciones bajas en la asignatura.

Estos resultados se pueden deber a muchos factores, entre ellos; falta de motivación al presentar el examen departamental, que las evaluaciones de la clase hayan sido más holísticas, que los reactivos en el examen necesitan mejorar su calidad técnica, etc. Sin embargo, no se puede hacer a un lado el hecho de que los resultados del examen no coinciden con las calificaciones obtenidas en la asignatura.

El análisis presentado en los párrafos anteriores llevó a realizar un segundo estudio para asegurar que los resultados obtenidos fueran correctos. Por tal motivo, se llevó a cabo un estudio para saber si existe correlación entre la calificación obtenida en el examen departamental y la calificación obtenida en la asignatura. El estudio se hizo con la correlación lineal de Pearson.

Con este estudio se encontró que existe una correlación positiva, pero baja. Aun así, debido a lo disperso de la correlación se consideró que no es evidencia suficiente para demostrar el dominio de los alumnos sobre los contenidos de la asignatura (véase la tabla 1).

Tabla 1.
Diagrama de dispersión



Nota: Tabla elaborada con información obtenida de los resultados del examen de Morfosintaxis de la Segunda Lengua y las calificaciones obtenidas por los alumnos en la asignatura con el mismo nombre.

Grado de homogeneidad

Para identificar el grado de homogeneidad, se hizo un análisis entre los resultados logrados por los alumnos de la misma unidad de aprendizaje que recibieron el curso con distintos profesores. En este estudio, se identificaron a cuatro grupos; 102, 103, 104 y 105. El grupo 102, obtuvo una media de 32.8 aciertos en el examen de Morfosintaxis de la Segunda Lengua, mismo

que cuenta con 61 ítems; el grupo 103 obtuvo una media de 29.2; el grupo 104 con una media de 31. Y el grupo 105 logró una media de 37.9.

Al analizar estos resultados, se puede apreciar que las diferencias son mínimas, (29, 31, 32, 37). Esta información sugiere que existe cierto grado de homogeneidad, a pesar de que la media obtenida por cada grupo sea baja y poco positiva. Aun así, se sugiere que los docentes que atendieron a los grupos 102, 103, y 104 se entrevisten con el docente del grupo 105 para hablar sobre los métodos que emplea en su salón de clase.

Desviación estándar

Para comprender mejor el examen se llevó a cabo un análisis de la desviación estándar donde se incluyeron los resultados de todos los evaluados para identificar el índice de dificultad del examen en su totalidad. En este ejercicio se identificó que se cuenta con una media general de 33.5 y de una desviación estándar de 7.1. Este dato nos dice que cualquier evaluado que haya acertado entre 26.3 y 40.6 reactivos se encuentra en un rango aceptable para este examen. Esto a pesar de que 26 reactivos acertados de 61 reactivos, con los que cuentan el examen, nos da un porcentaje de 43% de aciertos sobre el 100%. Este resultado nos lleva a pensar que el examen es difícil.

Índice de dificultad

Los resultados derivados del análisis hecho a los reactivos sobre su índice de dificultad indican que la gran mayoría de los reactivos son clasificados como fáciles, dificultad media o difíciles, lo cual le da cierto sentido de simetría en cuanto a la distribución de los ítems. Sin embargo, en la tabla 2 se puede observar que la distribución de los ítems de acuerdo al índice de dificultad es poco asimétrica, de acuerdo a la propuesta de Backhoff, Larrazolo y Rosas (2000). Se cuenta con más ítems clasificados como fáciles o difíciles de los que debería tener y la cantidad de ítems con dificultad media está muy por debajo de lo que debería ser (véase la tabla 2).

Tabla 2.
Índice de dificultad.

Recomendado	Escala	Ítems	Porcentaje
5% Fáciles	.7- más	13	21%
20% Medianamente fáciles	.6-.7	8	13%
50% Dificultad media	.5-.6	16	26%
20% Medianamente difíciles	.4-.5	10	16%
5% Difíciles	.39-menor	14	23%

Nota: Tabla adaptada de Backhoff, Larrazolo y Rosas (2000).

Grado de discriminación

Un ítem debe ser diseñado de tal forma que los ítems puedan confundir a los estudiantes que obtiene los promedios más bajos más no a los que cuentan con los promedios más altos.

Tabla 3.
Índice de discriminación.

Mexicali		
Escala	Evaluación del ítem	Ítem
.40-más	Excelente	17
.30-.39	Buena	3
.20-.29	Regular	17
.0-.19	Pobre	20
-.1-menor	Pésima	4

Nota: Tabla adaptada de Ebel y Frisbie (1986), con información obtenida de los resultados en el examen aplicado.

Al analizar la tabla 3 se puede observar que 20 ítems cuentan con un índice de discriminación de .19 o menor lo cual significa que deben ser mejorados o rechazados. También se puede observar que 4 ítems cuentan con un índice de discriminación de -.1 o inferior por lo que se recomienda eliminar dichos ítems. Estos 24 ítems representan el 39% de los ítems en el examen que tienen que ser mejorados o rechazados.

Conclusiones

Los exámenes estandarizados permiten ver de manera objetiva los conocimientos adquiridos por los estudiantes para después identificar a grupos o docentes sobresalientes mismos que pueden ayudar a mejorar las prácticas educativas buscando un mayor índice de homogeneidad entre diferentes grupos que llevan una misma asignatura.

Por tal motivo, es importante asegurar el buen funcionamiento de un examen de este tipo. Se deben realizar una serie de análisis, como los que se han planteado en este documento, para que las decisiones que deriven de él cuenten con la validez suficiente para poder tomar decisiones que permitan la mejora de la asignatura y de los conocimientos y habilidades adquiridos por los evaluados.

En cuanto a los estudios realizados para identificar la correlación entre las calificaciones obtenidas en el examen, mismas que cuentan con un promedio de 55% sobre 100%, y aquellas obtenidas en la asignatura, con un promedio de 85% sobre 100%, cuentan con una diferencia de 30 puntos. Debido a esta amplia diferencia, se motiva a realizar estudios adicionales que ayuden a mejorar el examen de Morfosintaxis de la Segunda Lengua para fortalecer la correlación obtenida en este estudio.

Por otro lado, se muestra que existe un índice de homogeneidad. Se observó una diferencia mínima entre las medias obtenidas por cada grupo. Esta información nos sugiere que, si bien todos los grupos obtuvieron una media baja, el examen tiene mucho que ver debido a que no es un hecho aislado. Si todos los grupos obtuvieron una media baja en el examen se puede deber a que el examen es difícil de contestar y se debe revisar.

El clasificar al examen como difícil también surge del análisis hecho para identificar la desviación estándar del examen. Esto se debe a que los resultados de este análisis colocan a 26 aciertos de 61 dentro de un rango aceptable para este examen. Cabe destacar que 26 ítems de 61 equivalen a 43% de aciertos sobre el 100%. Por tal motivo se debe trabajar para aumentar el 43% a un mínimo de 60%.

El análisis hecho a los ítems para conocer su índice de dificultad da como resultado una carga considerable de los ítems hacia aquellos que son considerados como difíciles. Esto se debe a que 23% de los reactivos son vistos como difíciles superando al 5% propuesto por Backhoff, Larrazolo y Rosas (2000) por 18 puntos.

Finalmente, el índice de discriminación de los ítems indica que por lo menos 20 ítems deben ser revisados y o eliminados. Este resultado podría explicar el por qué este examen es visto como difícil de acuerdo a los diferentes análisis realizados al examen. Por tal motivo se sugiere dar seguimiento a los resultados del examen y mejorar la calidad técnica del mismo para que en un futuro se pueda tener más información sobre el índice de homogeneidad y los conocimientos adquiridos por los evaluados.

Referencias

- Aiken, L. (2003). *Test psicológicos y evaluación*. Pearson Education.
- Airasian, P. (1996). *Assessment in the classroom*. McGraw-Hill, Inc.
- Adom, D. Adu, J. y Atsu, D. (2020). Test, measurement, and evaluation: Understanding and use of the concepts in education. *International Journal of Evaluation and Research in Education (IJERE)*, 9 (1), 109-119.
- Anastasia, A. y Urbina, S. (1998). *Tests psicológicos*, (7ª. Ed). Pearson Education.
- 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), 11-29. <http://redie.uabc.mx/vol2nol/contenido-backhoff.html>
- Boopathiraj, C. y Chellamani, K. (2013). Analysis of test items on difficulty level and discrimination index in the test for research in education. *International Journal of Social Science & Interdisciplinary Research*, 2 (2), 189-193. <https://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.1080.8467&rep=rep1&type=pdf>
- Ebel, R. y Frisbie, D. (1991). *Essentials of Educational Measurement* (5ta. ed.). Prentice-Hall of India
- Evers et al. (2013). Assessing the quality of tests: Revision of the EFPA review model. *Psicothema*, 25 (3), 283-291. DOI: <https://doi.org/10.7334/psicothema2013.97>
- Fein, M. (2012). *Test development: Fundamentals for certification and evaluation*. The American Society for Training and Development.
- Guardia, J., Freixa, M., Pero, M. y Turbany, J. (2012). *Análisis de datos en psicología* (2da. ed.). Delta Publicaciones.
- Hernández, R., Fernández, C. y Baptista, P. (2010). *Metodología de la investigación*. Mc.GrawHill.
- Hernández Madrigal, M., Ramírez Flores, E. y Gamboa Cerda, S. (2018). La implementación de una evaluación estandarizada en una institución de educación superior. *Innovación educativa*, 18 (76), 149-170. http://www.scielo.org.mx/scielo.php?script=sci_arttext&pid=S1665-26732018000100149&lng=es&tlng=es.
- Haladyna, T. (2004). *Developing and validating multiple-choice test items*. Lawrence Erlbaum Associates, Inc.
- Linn, R.L. (2000). Assessment and accountability. *Educational Researcher*, 29(2), 4-16. <https://journals.sagepub.com/toc/edr/29/2>
- Mendoza, R. (2015). La validez en los exámenes de alto impacto: Un enfoque desde la lógica argumentativa. *Perfiles Educativos*, 37 (149), 169-186. DOI: <https://doi.org/10.22201/iisue.24486167e.2015.149.53132>

- Olmos, J., Blanchart, M., Cebollero, M. y Oset, J. (2012). *Análisis de datos en psicología* (2da. ed.). Delta Publicaciones.
- Osterlind, S. (1998). *Constructing Test Items: Multiple-Choice, Constructed-Response, Performance, and Other Formats* (2da. ed.). Kluwer Academic Publishers.
- Pérez, B. (2014). *Metodología de la evaluación educativa*. Instituto Politécnico Nacional.
- Restrepo, L. y González, J. (2007). De Pearson a Spearman. *Revista Colombiana de Ciencias Pecuarias*, 20 (2), 183-192. <https://www.redalyc.org/pdf/2950/295023034010.pdf>
- Saavedra, M. (2001). *Evaluación del aprendizaje: conceptos y técnicas*. Editorial Pax.
- Sandoval, P., Maldonado-Fuentes, A. y Tapia-Landino, M. (2022). Evaluación educativa de los aprendizajes: conceptualizaciones básicas de un lenguaje profesional para su comprensión. *Páginas de Educación*, 15 (1), 49-75. DOI: <https://doi.org/10.22235/pe.v15i1.2638> .
- UABC. Gaceta UABC (2018). *Estatuto Escolar de la Universidad Autónoma de Baja California Estatuto Escolar*. 3 de diciembre. Anexo no. 408. http://sriagral.uabc.mx/Externos/AbogadoGeneral/Reglamentos/Estatutos/03_EstatutoEscolarUABC_ReformasDic032018.pdf
- University of Washington (2022). *Understanding Item Analyses*. [Report] <https://www.washington.edu/assessment/scanning-scoring/scoring/reports/item-analysis/>

ⁱ Doctor en Educación por la Universidad de Xochicalco, maestro en Docencia por parte de la UABC, licenciado en Docencia del Idioma Inglés. Profesor de tiempo completo en la Facultad de Idiomas de la UABC campus Mexicali donde imparte asignaturas y dirige tesis en los programas educativos de Licenciatura en Enseñanza de Lenguas y Maestría en Lenguas Modernas. Actualmente es Coordinador de Formación Profesional. Correo electrónico: jcasillas@uabc.edu.mx

ⁱⁱ Doctora en Ciencias Educativas por la UABC, maestra en Educación por la Universidad Autónoma de Ciudad Juárez, Especialidad en Docencia y Diplomados en Diseño Curricular y en Evaluación de Educación Superior, licenciatura en Ciencias de la Educación en la UABC. Es profesora investigadora de tiempo completo en la Facultad de Idiomas en la UABC en Mexicali, donde atiende docencia y dirige tesis en licenciatura, maestría y doctorado. Autora en los temas de Currículum, didáctica y vinculación en educación básica y superior. Miembro del Sistema Nacional de Investigadores. Correo electrónico: liliam@uabc.edu.mx

ⁱⁱⁱ Doctora en Ciencias Educativas, maestra en Docencia y licenciada en Docencia del Idioma Inglés (UABC). Profesora-investigadora de tiempo completo en la Facultad de Idiomas de la UABC, donde imparte asignaturas y dirige tesis en los programas educativos de licenciatura en Enseñanza de Lenguas, Maestría en Lenguas Modernas y Doctorado en Ciencias del Lenguaje. Actualmente es coordinadora de Investigación y Posgrado. Su área de investigación se desarrolla en los temas de currículum y formación profesional de profesores de lenguas. Candidato a SNI. Correo electrónico: laurafierro@uabc.edu.mx