ISSN: 2448 - 6574

Análisis psicométrico de exámenes de Matemáticas y Lenguaje y Comunicación en CONALEP Estado de México.

Psychometric analysis of Mathematics and Language and Communication exams in CONALEP State of Mexico.

Irvin Rodolfo Tapia Bernabé
irtb.tapia@gmail.com

Luis Enrique González Mejía
glezenkike@gmail.com
Colegio de Educación Profesional Técnica del Estado de México

Resumen

El presente trabajo demuestra en qué medida los reactivos utilizados en pruebas censales diagnóstica de Matemáticas y en Lenguaje y Comunicación, impactó el desempeño de estudiantes de segundo semestre de CONALEP Estado de México. Como objetivo específico, se tuvo medir el grado en el que los ítems son capaces de establecer diferencias en estudiantes con niveles altos o bajos en sus habilidades, capacidades y conocimientos, así como clasificar por nivel de dificultad cada uno de los ítems de los exámenes de diagnóstico, ambos bajo la perspectiva de la Teoría Clásica del Test. La investigación se encuentra centrada bajo un enfoque cuantitativo con un diseño descriptivo transeccional. Se llegó a la conclusión de que ambos exámenes presentan dificultad alta y una discriminación en sus reactivos baja.

Palabras clave: Análisis psicométrico, pruebas censales, dificultad y discriminación.

Abstract

The present work demonstrates the extent to which the reagents used in the census tests of Mathematics and Language and Communication, impacted the performance of second-year students of CONALEP Estado de México. As a specific objective, we had to measure the degree to which the items are able to establish differences in students with high or low levels in their skills, abilities and knowledge, as well as to classify by level of difficulty each one of the items of the exams. diagnosis, both from the perspective of the Classical Test Theory. The research is centered on a quantitative approach with a descriptive transectional design. It was concluded that both exams present high difficulty and low discrimination in their reagents.

Palabras clave: Psychometric analysis, census tests, difficulty and discrimination.

ISSN: 2448 - 6574

Planteamiento del problema

Objetivo General:

Describir estadísticamente los índices de dificultad y discriminación de ítems politómicos de una prueba diagnóstica de Matemáticas y de Lenguaje y Comunicación, a partir de los resultados del desempeño de estudiantes del ciclo escolar 1.17.18 en Conalep Estado de México.

Objetivos específicos:

- Establecer los factores ponderativos equivalentes a 0 y 1 obtenidos mediante la calificación de los ítems politómicos.
- Determinar los índices de dificultad y discriminación de los ítems que componen una prueba de conocimientos de matemáticas y lenguaje y comunicación.
- Determinar el grado de consistencia interna de la prueba a partir del nivel de correlación biserial de los ítems.

Pregunta de investigación

¿Cuál es el modelo psicométrico apropiado para determinar los índices de dificultad y discriminación en pruebas objetivas?

¿De qué manera impacta el índice de dificultad y discriminación del instrumento de medición en la evaluación de los estudiantes?

Justificación de la investigación

Elevar la calidad de la educación en México, es hoy en día uno de los principales ejes de atención de la política pública en nuestro país. A raíz de la creación de la Política Nacional de Evaluación de la Educación (INEE, 2015) se pretendió entre otros, desarrollar una cultura de evaluación y toma de decisiones basada en la evidencia, así como la construcción de los Programas Estatales de Evaluación y Mejora Educativa (PEEME).

Las áreas de evaluación en los estados se encuentran en proceso de fortalecimiento de sus capacidades institucionales y han desarrollado, de la mano del Instituto Nacional para la Evaluación de la Educación (INEE), sus propios Programas Estatales de Evaluación y Mejora Educativa (PEEME).

ISSN: 2448 - 6574

En el Estado de México, se ha integrado al PEEME de la EMS, el desarrollo de propuestas de evaluación para la mejora de los procesos a través del desarrollo de exámenes realizados por sus cuerpos académicos (CESPD, s.f.). Un ejemplo de lo anterior es el instrumento para evaluar las competencias Matemáticas y Lenguaje y comunicación en estudiantes de segundo semestre de Conalep Estado de México. Si bien este tipo hechos representa acciones plausibles ante este panorama, es importante cuestionarse:

¿Cumple este tipo de instrumentos de evaluación los estándares de calidad establecidos por las organizaciones internacionales?, ¿Cuál será la dificultad y el poder de discriminación de las pruebas empleadas para evaluar los aprendizajes en Matemáticas y Lenguaje y Comunicación de segundo semestre en el Conalep Estado de México?, ¿De qué manera impacta el índice de dificultad y discriminación de los instrumentos de medición en la evaluación de los estudiantes?

Cuando se utilizan instrumentos de gran escala y alto impacto, como en el caso de las pruebas censales, es necesario conocer los indicadores técnicos que definen la calidad del instrumento educativo que se emplea los exámenes de gran escala los cuales son diseñados para aplicarse en más de un plantel escolar. Por su dimensión y por el poderoso impacto social que tienen, su elaboración debe ajustarse a rigurosos estándares de calidad (Aiken, 1996).

El propósito de una prueba educativa es hacer inferencias acerca del conocimiento que tiene un estudiante respecto al dominio evaluativo que se pretende medir; información que es útil a los educadores para tomar decisiones tendientes a mejorar el proceso educativo (INEE, 2005). Sin embargo, sin importar qué tan cuidadosamente se elabore una prueba, los resultados no tienen ningún valor si esta no se administra y califica en forma adecuada. (Lewis R., 2013).

Una fuente importante de estos recursos son los Estándares para la Evaluación Educativa y Psicológica, una serie de 264 normas para construir, evaluar, administrar, calificar, interpretar y usar los resultados. Enfatizan la importancia de tomar en cuenta el bienestar de las personas que hacen una prueba y evitar el mal uso de los instrumentos de evaluación (Backoff, Larrazolo, & Rosas, 2000). Sin embargo, mientras que, en países desarrollados, es obligatorio que estos criterios de calidad se satisfagan, en México es inexistente esta normatividad.

El presente trabajo de investigación educativa busca examinar los resultados psicométricos relacionados con su nivel de dificultad y poder de discriminación de los reactivos de la evaluación diagnóstica dirigida a estudiantes de segundo semestre de la generación 2017-2020 en el Conalep Estado de México, guiado mediante la Teoría Clásica de los Test, vigente de acuerdo con (Backoff, Larrazolo, & Rosas, 2000).

ISSN: 2448 - 6574

El estudio generará precedentes en la valoración de la calidad de los instrumentos de medición empleados para evaluar el desarrollo de las competencias en estudiantes, a fin de proporcionar información correcta para la toma de decisiones. Además, aporta una guía a la labor de evaluación educativa de las instituciones de Educación Media Superior a través de la experiencia de Conalep Estado de México.

Fundamentación Teórica

Teoría clásica de los test

La TCT ha sido el modelo dominante en la teoría de test, y tiene aún una vigencia representativa en el campo de la evaluación psicológica y educativa. Esta teoría propuesta por Charles Sperman a inicios del siglo XX, usa un modelo matemático sustentado en la curva normal que supone que la habilidad de un sujeto es la sumatoria de los puntajes obtenidos al responder una serie de ítems de una prueba. La TCT se centra en la estimación del puntaje de una persona como si esta hubiera respondido al universo total de preguntas posibles, como este universo es infinito, es necesario hacer una estimación de este puntaje, el cual tendrá cierta cantidad de error (Navas, 1994).

Lo anterior define los siguientes supuestos en los que esta soportada la TCT: el puntaje verdadero es igual a la esperanza matemática o valor esperado de las puntuaciones observadas: la correlación entre el puntaje verdadero en un test y el error en ese test es igual a cero. Es decir, no existe relación entre ambos: la correlación entre los errores dados en dos test diferentes es igual a 0; es decir, los errores son independientes (Navas, 1994).

Índice de Dificultad

El índice de dificultad de un ítem de acuerdo con la TCT, se entiende como la proporción de personas que responden correctamente un reactivo de una prueba. Entre mayor sea esta proporción, menor será su dificultad. Lo que quiere decir que se trata de una relación inversa: a mayor dificultad del ítem, menor será (Backoff, Larrazolo, & Rosas, 2000).

Para calcular la dificultad de un ítem, se divide el número de personas que contestan correctamente el ítem entre el número total de personas que intentan resolver el ítem (correcta o incorrectamente).

$$ID = \frac{Ai}{Ni} \quad \text{.....(Ecuación 1)}$$

Donde: A_i = Número de personas que aciertan el ítem.
N_i = Número de personas que intentaron resolver el reactivo.

La evaluación de los reactivos corresponde a: 0.86 Altamente fáciles, 0.74 a 0.86 Medianamente fáciles 0.53 – 0.73 Dificultad media 0.33 – 0.52 Medianamente difíciles < – 0.32 Altamente difíciles.

ISSN: 2448 - 6574

Índice de Discriminación

Si la prueba y un ítem miden la misma habilidad o competencia, podemos esperar que quien tuvo una puntuación alta en todo el test deberá tener altas probabilidades de contestar correctamente el ítem. También debemos esperar lo contrario, es decir, que quien tuvo bajas puntuaciones en el test, deberá tener pocas probabilidades de contestar correctamente el reactivo. Así, un buen ítem debe discriminar entre aquellos que obtuvieron buenas calificaciones en la prueba y aquellos que obtuvieron bajas calificaciones.

Usualmente, se utilizan dos formas para determinar el poder discriminativo de un ítem: el índice de discriminación y el coeficiente de discriminación. Aunque hay varias maneras equivalentes de calcular el índice de discriminación, en este trabajo utilizaremos la siguiente fórmula:

$$P = \frac{Ac - Ai}{M} \text{..... (Ecuación 2)}$$

Donde: Ac: La frecuencia de aciertos en el grupo superior (convenientes)
Ai: La frecuencia de aciertos en el grupo inferior (inconvenientes)
M: El total de individuos en cada grupo

Entre más alto es el índice de discriminación, el reactivo diferenciará mejor a las personas con altas y bajas calificaciones. Si todas las personas del Ac contestan correctamente un reactivo y todas las personas del Ai contestan incorrectamente, entonces P = 1 (valor máximo de este indicador); si sucede lo contrario, P = -1 (valor máximo negativo); si ambos grupos contestan por igual, P = 0 (valor mínimo de discriminación).

Ebel y Frisbie como se cita en (Backoff, Larrazolo, & Rosas, 2000) nos dan la siguiente regla de "dedo" para determinar la calidad de los reactivos, en términos del índice de discriminación. La Tabla 1, muestra los valores D y su correspondiente interpretación. Asimismo, en la tabla se señalan las recomendaciones para cada uno de estos valores.

Tabla I. Poder de discriminación de los reactivos según su valor D

D=	Calidad	Recomendaciones
> 0,39	Excelente	Conservar
0,30 . 0,39	Buena	Posibilidades de mejorar
0,20 . 0,29	Regular	Necesidad de revisar
0,00 . 0,20	Pobre	Descartar o revisar a profundidad
< -0,01	Pésima	Descartar definitivamente

ISSN: 2448 - 6574

Correlación biserial y coeficiente de correlación biserial

El coeficiente de correlación biserial se calcula para determinar el grado en que las competencias que mide el test también las mide el reactivo. Proporciona una estimación de la correlación producto-momento de Pearson entre la calificación total de la prueba y el continuo hipotético del reactivo, cuando éste se dicotomiza en respuestas correctas e incorrectas.

Se consideran aceptables los ítems con valores superiores a 0.25. La ecuación para obtener este indicador, es la siguiente:

$$r_{bp} = \frac{\bar{x}_p - \bar{x}}{\sigma} * \sqrt{\frac{p}{q}} \quad \text{..... (Ecuación 3)}$$

Donde: *p*: La proporción de individuos que acertaron
q: La proporción de individuos que fallaron
X_p: La media en X de los sujetos cuya proporción es *p*
X: La media del test
S_x: La desviación típica del test

Metodología

Diseño de la investigación

De acuerdo (Hernández, Fernández, & Baptista, 2014) el estudio se caracteriza por tener un alcance descriptivo y un diseño no experimental del tipo transeccional descriptivo debido a que las variables no serán manipuladas. Se fundamenta en que solo se emplearán los datos recolectados posterior a la aplicación de dos pruebas de conocimientos a estudiantes de una generación escolar.

Selección de la muestra

Se consideró a la totalidad de los resultados obtenidos en la aplicación censal de los exámenes de diagnóstico en las competencias en Matemáticas y Lenguaje y comunicación en alumnos de segundo semestre de la generación 2017-2020 de los 39 planteles del Conalep, como a continuación se indica:

- Examen de Matemáticas: 14,392 estudiantes
- Examen de Lenguaje y Comunicación: 13,773 estudiantes

ISSN: 2448 - 6574

Método de recolección de datos

A partir de los exámenes desarrollados por el Conalep Estado de México, los cuales se denominaron: Examen de Evaluación Diagnóstica en Matemáticas y Examen de Evaluación Diagnóstica en Lenguaje y Comunicación. Ambos tuvieron como objetivo valorar el dominio de contenidos adquiridos en el primer semestre.

Las pruebas contienen reactivos de tipo politómico o mejor conocidos como opción múltiple, los cuales se caracterizan por su versatilidad para evaluar conocimiento factual (puramente memorístico), habilidades intelectuales de alto orden, o disposiciones actitudinales y valorativas. Con ese tipo de preguntas, siempre que sean bien utilizadas, se puede medir una gran cantidad de atributos sofisticados de los estudiantes (Instituto Nacional de Evaluación Educativa, 2013)

Los instrumentos incluyen los siguientes aspectos evaluados:

Matemáticas: "Sentido numérico y pensamiento algebraico" con 13 reactivos, "Cambios y relaciones", con 2 reactivos. Conformando un total de 15 reactivos.

Lenguaje y Comunicación:"Texto expositivo" con 11 reactivos, "Manejo y construcción de la información" 3 reactivos, "Texto literario" con un reactivo.

Procedimiento

Los pasos para la administración y calificación del examen fueron definidos por el área académica del Conalep Estado de México como a continuación se indica: (1) se compartió a planteles el link el examen el cual fue aplicado a través de un formulario de Google Drive; (2) los responsables de la aplicación del examen en planteles colocaron las pantallas de inicio en los laboratorios de informática; (3) los estudiantes respondieron el examen sin ningún tipo de ayuda (calculadora, diccionario libros, etc.). Durante este proceso, se encontró siempre presente una persona capacitada que resolvió cualquier problema o duda sobre el manejo de la parte computarizada del examen.

Las respuestas de los estudiantes fueron extraídas de la aplicación y conjuntadas en una base de datos para su procesamiento estadístico. Se generó una hoja de cálculo en Excel para su interpretación de las respuestas a un formato binario (0 y 1). Utilizando el software de hoja de cálculo Excel, se calcularon los valores p (dificultad), D (índice de discriminación) y rpbis para todos los reactivos del examen.

El índice de dificultad se calculó con la ecuación (1), el índice de discriminación con la ecuación (2) y el coeficiente de discriminación con la ecuación (3).

ISSN: 2448 - 6574

Resultados

La tabla 02 muestra el índice de dificultad, índice de discriminación y coeficiente de correlación biserial de los 15 reactivos de ambas pruebas.

Tabla 02. índice de dificultad, índice de discriminación y coeficiente de correlación biserial

	Matemáticas			Lenguaje y Comunicación		
Ítem	Índice de Dificultad	Índice de Discriminación	Coeficiente de Correlación Biserial	Índice de Dificultad	Índice de Discriminación	Coeficiente de Correlación Biserial
1	60,7	0,48	0,40	40,10	0,44	0,36
2	68,7	0,37	0,33	56,93	0,62	0,50
3	72,5	0,42	0,33	37,87	0,47	0,38
4	80,5	0,29	0,33	84,56	0,14	0,17
5	71,7	0,31	0,30	75,92	0,33	0,34
6	61,4	0,36	0,31	78,07	0,20	0,21
7	71,0	0,35	0,33	74,60	0,35	0,34
8	48,0	0,51	0,40	67,55	0,48	0,41
9	81,3	0,27	0,33	72,29	0,17	0,15
10	82,7	0,22	0,28	46,05	0,50	0,40
11	69,9	0,35	0,30	70,56	0,32	0,31
12	84,3	0,19	0,24	63,15	0,34	0,30
13	77,0	0,30	0,31	79,64	0,17	0,17
14	55,9	0,39	0,29	92,16	0,05	0,10
15	70,9	0,21	0,20	81,48	0,13	0,15

Índice de dificultad: Los resultados arrojados del examen de Matemáticas el 86%, de los reactivos se ubican en un nivel de dificultad alto y 14% moderado. En el caso de Lenguaje y Comunicación, indican que el 73% de los reactivos tienen un índice de dificultad alto, el 20% y un 10% fácil. La tabla 03 muestra la interpretación de la dificultad de los ítems.

ISSN: 2448 - 6574

Tabla 03

	Lenguaje y Comunicación	Matemáticas
Ítem	Interpretación	Interpretación
Ítem 1	Moderado	Difícil
Ítem 2	Moderado	Difícil
Ítem 3	Fácil	Difícil
Ítem 4	Muy difícil	Difícil
Ítem 5	Difícil	Difícil
Ítem 6	Difícil	Difícil
Ítem 7	Difícil	Difícil
Ítem 8	Difícil	Moderado
Ítem 9	Difícil	Difícil
Ítem 10	Moderado	Difícil
Ítem 11	Difícil	Difícil
Ítem 12	Difícil	Difícil
Ítem 13	Difícil	Difícil
Ítem 14	Muy difícil	Moderado
Ítem 15	Difícil	Difícil

Interpretación de la dificultad de los ítems.

Índice de discriminación: En el caso de Matemáticas el 20% de los reactivos presenta una excelente calidad, el 46% en una calidad buena, el 26.6 % una calidad regular y el 8% una mala calidad. En el examen de Lenguaje y Comunicación se identificó el 33% de los reactivos con una excelente calidad, el 26.6% con una buena calidad, el 4.8% con calidad regular y 26.6 con mala calidad. La tabla 04 muestra la evaluación de la calidad de los ítems.

ISSN: 2448 - 6574

	Lenguaje y Comunicación		Matemáticas	
Ítem	Calidad	Recomendación	Calidad	Recomendación
Ítem 1	Excelente	Conservar	Excelente	Conservar
Ítem 2	Excelente	Conservar	Buena	Posibilidades de mejorar
Ítem 3	Excelente	Conservar	Buena	Posibilidades de mejorar
Ítem 4	Pobre	Descartar o revisar a profundidad	Regular	Necesidad de revisar
Ítem 5	Buena	Posibilidad de mejorar	Buena	Posibilidades de mejorar
Ítem 6	Pobre	Descartar o revisar a profundidad	Buena	Posibilidades de mejorar
Ítem 7	Buena	Posibilidad de mejorar	Buena	Posibilidades de mejorar
Ítem 8	Excelente	Conservar	Excelente	Conservar
Ítem 9	Pobre	Descartar o revisar a profundidad	Regular	Necesidad de revisar
Ítem 10	Excelente	Conservar	Regular	Necesidad de revisar
Ítem 11	Buena	Posibilidad de mejorar	Buena	Posibilidades de mejorar
Ítem 12	Buena	Posibilidad de mejorar	Pobre	Descartar o revisar a profundidad
Ítem 13	Pobre	Descartar o revisar a profundidad	Buena	Posibilidad de mejorar
Ítem 14	Pobre	Descartar o revisar a profundidad	Buena	Posibilidad de mejorar
Ítem 15	Pobre	Descartar o revisar a profundidad	Regular	Necesidad de revisar

Tabla 04 Calidad de ítems

Coeficiente de correlación biserial: En Matemáticas el 86.6% muestra una correlación por encima de .25 dentro de la clasificación aceptable. Para el caso de Lenguaje y Comunicación el 60% de los reactivos se situaron por encima del .25 de la correlación.

En síntesis, el análisis realizado bajo la teoría de la TCT, muestra que ambos exámenes cuentan con un índice de dificultad por arriba del 60%, para efectos de evaluación, los ítems son en promedio “Muy Difíciles” para los estudiantes. En el caso del índice de discriminación, el examen de matemáticas solo presenta un reactivo el cual requiere ser cambiado, sin embargo, en Lenguaje y Comunicación el mismo caso se presenta en 5 de los reactivos. Los resultados de la correlación biserial nos indican en el caso de matemáticas que solamente dos reactivos se encuentran fuera del rango de aceptación, es decir que estos ítems no miden conocimientos de la misma área disciplinar que la mayoría del examen. En Lenguaje y comunicación este coeficiente demuestra que 6 reactivos se encuentran fuera del rango establecido.

ISSN: 2448 - 6574

Conclusiones

La mejor manera de conocer la calidad de los reactivos de un examen es a través de su análisis empírico una vez puestos a prueba. De acuerdo con la Teoría Clásica del Test aplicada a la evaluación educativa los indicadores fundamentales para realizar el análisis son el índice de dificultad, el índice de discriminación y la correlación biserial puntual. Esta teoría sigue vigente gracias a la sencillez de sus supuestos matemáticos.

Los resultados del estudio demuestran que el examen realizado en Conalep Estado de México con fines de evaluación diagnóstica a estudiantes de segundo semestre fue en promedio difícil para la mayoría de los estudiantes. Lo anterior debido a posibles omisiones metodológicas en la distribución de los niveles de dificultad de los reactivos o bien a que se hayan contemplado conocimientos fuera de los objetivos de aprendizaje en los programas de estudio, esto según la interpretación de los resultados del índice de discriminación y la correlación biserial.

Lo anterior también demuestra las necesidades de formación de las áreas de evaluación educativa para hacer frente a las necesidades del Plan Nacional de Evaluación Educativa y el Programa de Evaluación y Mejora Educativa del Estado de México, en razón de que los resultados de las evaluaciones de los aprendizajes deben servir para la correcta toma de decisiones que permita mejora la calidad educativa.

ISSN: 2448 - 6574

Referencias bibliográficas

Backhoff, E., Ibarra, M. A., & Rosas, ,. M. (1995). Sistema computarizado de exámenes (SICODEX). *Revista Mexicana de Psicología*, 55-62.

Backoff, E., Larrazolo, N., & Rosas, M. (2000). Nivel de dificultad y poder de discriminación del Examen de Habilidades y Conocimientos Básicos (EXHCOBA). *Revista Electrónica de Investigación Educativa*, 1.

CESPD. (s.f.). *cespd.edomex.gob.mx*. Obtenido de http://cespd.edomex.gob.mx/eval_proyectos_eb_peeme

Hernandez Sampieri, R., Fernández Collado, C., & Baptista Lucio, M. (2014). *Metodología de la Investigación .* Ciudad de México: McGraw-Hill.

INEE. (septiembre de 2005). Obtenido de www.inee.gob.mx.

INEE. (2015). *Política Nacional de Evaluación Educativa.* Ciudad de México: Documentos Rectores INEE.

INEE. (07 de Diciembre de 2017). *Gob.mx.* Obtenido de https://www.gob.mx/cms/uploads/attachment/file/278497/Calendario_de_Evaluaciones_INEE-SEP_2018.pdf

INEE. (diciembre de 2017). *www.inee.edu.mx.* Obtenido de http://publicaciones.inee.edu.mx/buscadorPub/P1/F/105/P1F105.pdf

Instituto Nacional de Evaluación Educativa. (03 de Marzo de 2013). *http://www.inee.edu.mx.* Obtenido de http://www.inee.edu.mx/images/stories/Publicaciones/Documentos_tecnicos/De_pruebasymedicion/construccion_reactivos/Completo/mtconstrecexcalemarca.pdf

Lewis R., A. (2013). *Test psicológicos y evaluación .* México: Pearson.

Navas, J. M. (1994). Teoría Clásica del Test versus Teoría de Respuesta al Ítem. *Psicológica*, 15.