VALIDACIÓN DE LA PRUEBA CAOS-4 EN EL CONTEXTO MEXICANO

MARÍA DE JESÚS VANESSA MENDOZA RIVERA
Universidad Autónoma de Aguascalientes

DANIEL EUDAVE MUÑOZ
Universidad Autónoma de Aguascalientes

JESÚS ENRIQUE PINTO SOSA
Universidad Autónoma de Yucatán

RESUMEN: La evaluación CAOS-4 (*Comprehensive Assessment of Outcomes in Statistics*), es una de las pocas pruebas de logro que han sido aplicadas a nivel superior y que se refieren al tema de la estadística. En este trabajo se describen los resultados del análisis de reactivos relativos a nivel de dificultad y poder de discriminación que se efectuaron con fines de valorar la calidad de la prueba en la Universidad Autónoma de Aguascalientes. Los resultados muestran que la mayoría de los reactivos tienen una dificultad de medianamente difícil y un poder discriminativo que va de pobre a regular. Así mismo se observa que los reactivos con mayor tasa de respuesta se encuentran en la categoría de datos bivariados.

Palabras clave: Evaluación del aprendizaje, exámenes, pruebas, estudiantes.

n esta ponencia nos interesa dar cuenta de algunos rasgos identitarios presentes en las comunidades estudiantiles de la FES Acatlán y UAM-Azcapotzalco, lo común entre ellos es que son jóvenes estudiantes de licenciatura ubicados en dos de las más importantes instituciones de educación superior pública, al tiempo que no son ajenos a experimentar diversas prácticas escolares que les resultan hostiles. Exponemos algunas formas de violencia que dicen vivir los estudiantes en

el espacio universitario. En la exposición se destaca la relación asimétrica de poder-saber del docente-alumno y la consecuente descalificación que hace éste en el salón de clase a los estudiantes

PALABRAS CLAVE: Educación Superior, Violencia Estudiantes, Docentes.

Introducción

Para llevar a cabo evaluación educativa, se tiene que seguir un proceso riguroso en el que se involucra la elaboración, aplicación y análisis de los instrumentos de medición, a estas etapas fue sometido el examen CAOS.4.

El principal objetivo de un instrumento de medición a gran escala como el que pretende ser CAOS-4, según Aiken, (1996) es ofrecer información acerca de los logros de los estudiantes con la finalidad de hacer un diagnóstico de las capacidades individuales, la efectividad de la enseñanza o bien del programa de estadística que se sigue en los ámbitos educativos y de esta forma contribuir con su mejoramiento.

A pesar de que hay esfuerzos realizados en materia de evaluación educativa en México, según Backhoff, Larrazolo y Rosas (2000) aún falta desarrollar evaluaciones de calidad clasificadas como pruebas de logro que sean estandarizadas y validadas en nuestro país.

CAOS-4, comenzó como parte de un movimiento de reforma de la educación estadística por un grupo de investigadores de la Universidad de Minnesota, se consolidó en un periodo de tres años con el propósito de contar con un instrumento que midiera la comprensión estadística de los estudiantes de nivel superior a través de 10 dimensiones propuestas.

Esta prueba de logro se adecuo a través de una traducción cuidadosa que fue supervisada directamente por miembros del equipo de la Universidad de Minnesota con la finalidad de cuidar la fiabilidad de la traducción al contexto mexicano.

El propósito de este trabajo es dar a conocer los resultados de los reactivos de CAOS-4 aplicado en el contexto de la Universidad Autónoma de Aguascalientes y para lograr el cometido, en primer lugar se hablará de la construcción de la prueba, posteriormente se abordará la metodología y finalmente se presentarán los resultados que se obtuvieron.

Construcción de la prueba

En respuesta a las dificultades que los estudiantes tienen con el aprendizaje y comprensión estadística, en 1900 se inició un movimiento de reforma con la finalidad de transformar la enseñanza de los niveles introductorios de estadística (Cobb, 1992, citado en Ben-Zvi y Garfield, 2004). Como parte de este movimiento surgieron varios institutos, entre ellos el proyecto ARTIST (*Assessment Resource Tools for Improving Statistical Thinking*) para direccionar los cambios de evaluación en educación estadística presentados por Garfield y Gal en1999, quienes destacaron la necesidad de desarrollar elementos de evaluación fiables, válidos y prácticos.

Un componente importante del proyecto ARTIST, fue plantear la evaluación CAOS-4 con la intención de que fuera una evaluación confiable compuesta de un conjunto de ítems que los estudiantes que han finalizado cualquier curso introductorio de estadística pueden comprender. Una segunda meta fue identificar las áreas donde los estudiantes están logrando o no avances significativos en su razonamiento y comprensión estadística.

La evaluación CAOS-4 fue desarrollada a través de un periodo de tres años de tal forma que se obtuvieron ítems de maestros expertos en la materia y también se redactaron algunas preguntas de aquellas áreas no cubiertas por los ítems adquiridos, además se obtuvo retroalimentación de asesores y evaluadores expertos.

La primera versión de CAOS contaba con 34 ítems de opción múltiple; esta versión fue usada en un estudio piloto con estudiantes universitarios de estadística introductoria durante el otoño de 2004. Los datos del estudio piloto fueron usados para hacer una revisión adicional de CAOS, resultando una segunda versión que contenía 37 ítems de opción múltiple. Los resultados de los análisis de datos de la prueba CAOS-2 fueron usados para cambios adicionales, lo que produjo la tercera versión de CAOS, es decir, CAOS-3.

Aunque las clasificaciones indicaron que la prueba media lo que debería medir, los maestros e instructores realizaron sugerencias de cambios. Esta retroalimentación fue usada para sumar y eliminar reactivos de la prueba, además de realizar revisiones extensas para producir la versión final de la prueba, llamada CAOS-4, que actualmente consta de 40 ítems de opción múltiple.

El análisis de final de validez de contenido de CAOS-4 fue realizado por los investigadores de la Universidad de Minnesota en marzo del 2006. La evaluación se aplicó a un total de 1470 estudiantes que estaban terminando un primer curso de estadística en el nivel superior. Con esta muestra se realizó un análisis de consistencia interna de los ítems que componen el instrumento en la que se produjo un alfa de Cronbach de 0.82, considerando aceptable la consistencia interna del instrumento.

Cabe señalar que según delMas, Garfield, Ooms y Chance (2007), todos los elementos de la prueba CAOS-4 fueron pensados para exigir a los estudiantes un pensamiento y un razonamiento estadístico, no para calcular ni usar fórmulas o definiciones de recuerdo, lo que marca la diferencia con muchos exámenes diseñados por los docentes.

Uno de los propósitos más importantes de la prueba CAOS-4, es proporcionar información útil a los profesores de enseñanza de las estadísticas con la finalidad de que puedan determinar si sus estudiantes están aprendiendo a razonar y pensar estadísticamente para de esta forma promover cambios educativos.

Metodología

La evaluación CAOS-4, fue desarrollada con la finalidad de proponer un instrumento de evaluación que pudiera tener una cobertura amplia del contenido estadístico cubierto en un primer curso. La evaluación se obtuvo de la página web ARTIS diseñada por el equipo de Garfield y delMas de la Universidad de Minnesota y con previa autorización de sus diseñadores.

Las áreas temáticas que se abarcan en CAOS-4 son recolección de datos, estadística descriptiva, representación gráfica, diagrama de cajas, distribución normal, datos bivariados, probabilidad, variabilidad de la muestra, intervalos de confianza y finalmente pruebas de significancia; cada una de estas áreas esta representada por al menos 4 reactivos de opción múltiple.

El trabajo de traducción del instrumento se realizó en tres momentos: una primera etapa consistió propiamente en la traducción de la prueba; posteriormente el documento fue revisado por un maestro de enseñanza del idioma inglés el cual validó que la traducción fuera fiel y que las preguntas tuvieran la coherencia del documento original; y finalmente fue

revisado y autorizado por el Dr. delMas de la Universidad de Minessota y aprobó la aplicación de la versión español.

La versión en español estuvo lista a finales del mes de octubre de 2012 para la primera aplicación en la Universidad Autónoma de Aguascalientes, la cual se llevó acabo en el mes de noviembre de ese mismo año.

Posterior al trabajo de traducción de la prueba, se llevó a cabo el proceso de aplicación para validar la traducción del instrumento. La aplicación se efectuó en dos etapas, con la finalidad de asegurar que la lectura de las preguntas fuera comprensible para los sustentantes y al mismo tiempo estimar la duración de la misma en formato de lápiz y papel.

Los participantes que formaron parte del proceso de piloteo se escogieron con características similares a la muestra final empleando las mismas condiciones de aplicación para todos los casos.

El criterio de selección de los estudiantes de la primera aplicación de CAOS-4 fue bajo el criterio de un muestreo no probabilístico intencional. Se convocó a los estudiantes del Centro de Ciencias Sociales y Humanidades para participar en una primera etapa con 68 estudiantes del sexto semestre de las licenciaturas de Psicología y Asesoría Psicopedagógica. En esta primera etapa se verificó la duración de la aplicación que oscilaba entre 40 y 50 minutos.

Durante la primera aplicación se detectó la necesidad de incorporar un apartado en el que se incluyera la instrucción de circular en el cuadernillo de preguntas y/o escribir en la hoja de respuestas cualquier término o enunciado que no entendieran al momento de leer el ítem.

En una segunda etapa se incorporó la licenciatura en Mercadotecnia con una participación de 30 estudiantes de quinto semestre; con base en los resultados de esta aplicación, se determinó conservar la instrucción de circular y/o escribir en la hoja de respuestas cualquier término o enunciado que no se entendiera al momento de leer el ítem en la aplicación final llevada a cabo en noviembre de 2012 (ver gráfica 1).

Nueve carreras de la Universidad Autónoma de Aguascalientes del Centro de Ciencias Sociales y Humanidades de los semestres quinto, séptimo y noveno participaron

en este estudio. El muestreo que se utilizó fue no probabilístico intencional y el tamaño de la muestra fue 206 estudiantes, de los cuales 28.6% de la muestra eran del sexo masculino y el resto 71.4% del sexo femenino.

Resultados

Para determinar la calidad de los reactivos de la prueba propuesta se determinó el índice de dificultad y el poder de discriminación que se reportan en los siguientes párrafos y a través de la estadística descriptiva se obtuvo una media de 13.92, una mediana de 14, la desviación estándar alcanzada fue de 3.20, un valor mínimo de 5 con un valor máximo de 23 reactivos contestados correctamente, lo cual dista mucho de acercarse al reporte de aplicación de CAOS-4 presentado en 2006 por Garfield, delMas, Chance, Poly y Ooms, donde los datos que se reportaron fueron los siguientes: una media de 55.77, una mediana de 53.75, la desviación estándar de 16.13, con un valor mínimo de 18 y un máximo de 100.

El índice dificultad promedio de la prueba CAOS-4 es de 0.37, es decir, los reactivos caen en la categoría de medianamente difícil para los 206 estudiantes que participaron en la aplicación final de la prueba. En la tabla 1 se muestra el promedio obtenido por dimensión de la dificultad de los ítems.

Haciendo una comparación de las 10 áreas temáticas del examen, la dificultad de la evaluación oscila entre 0.06 hasta 0.69 siendo las preguntas más fáciles las que corresponden a datos bivariados y las más difíciles las que corresponden a la categoría de variabilidad de la muestra.

Clasificando los reactivos según su nivel de dificultad, podríamos agruparlos de la siguiente manera: 22.5% tienen una dificultad media en el examen; 40% de los reactivos se clasifican en la categoría de medianamente difíciles; y 37.5% están en el rango de altamente difíciles.

Por otro lado con la finalidad de identificar los reactivos que no están discriminando en la evaluación CAOS-4 se ejecutó la prueba *t* de *student* para muestras independientes, la cual indicó en los resultados, que de los 40 ítems redactados en la prueba original los reactivos 1, 4, 6, 7, 9, 10, 15, 16, 20, 23, 26, 28, 30, 31 y 40 no están discriminan adecuadamente por tener un nivel de significancia de $p > 0.05$. Los reactivos a los que se tiene que poner especial atención para mejorar la prueba CAOS-4 corresponden a las

categorías de: representación gráfica, distribución normal, probabilidad, diagrama de cajas, estadística descriptiva, variabilidad de la muestra, datos bivariados, pruebas de significancia, intervalos de confianza.

El poder de discriminación promedio de todo la prueba fue de 0.13, lo que indica que la totalidad de la evaluación debe revisarse a profundidad, además es importante mencionar que de los 40 ítems dos de ellos están discriminando negativamente.

Para identificar aquellos reactivos que necesitan alguna modificación, se utilizó un criterio óptimo de calidad para aceptar o rechazar reactivos retomado por Pérez, Acuña, y Arratia en 2008, es decir, aquellos con valores $p < 0.20$ es necesario descartarlos o bien revisarlos a profundidad y aquellos que cuentan con un valor $p > 0.30$ son considerados buenos reactivos. Utilizando estos datos, se puede notar que los reactivos que no cumplen con las mínimas normas de calidad se distribuyen en casi todas las áreas temáticas.

La dimensión con mayor tasa de respuesta fue la de datos bivariados la cual tuvo un promedio de 54% de respuestas correctas, seguida de la dimensión de intervalos de confianza con un promedio de 40.3%

Las dimensiones con menor tasa de respuesta se refieren a probabilidad con tan sólo el 26.6%, diagrama de cajas con el 28.03% y representación gráfica con el 29.98%. Es decir, los estudiantes no han logrado porcentajes altos en las dimensiones reportadas.

Conclusiones

En conclusión, ninguna dimensión logró ser contestada con más del 50% de los ítems de forma correcta en el contexto de las licenciaturas seleccionadas de la Universidad Autónoma de Aguascalientes.

La evaluación de CAOS -4 tuvo una serie de revisiones que buscaban determinar que la traducción realizada fuera la adecuada para el contexto mexicano, específicamente para estudiantes de la Universidad Autónoma de Aguascalientes por lo que la primera aplicación formal de esta evaluación fue fiel a la prueba original. Al someter la prueba a un proceso de confiabilidad interna a través de alfa de Cronbach los resultados no fueron alentadores ya que estos estuvieron muy por debajo de lo que Aiken (1996) ha reportado para las pruebas de aprovechamiento. Sin embargo, esto no quiere decir que la prueba

deba descartarse, como ha referido Morales (2007) un coeficiente de confiabilidad bajo no indica necesariamente que el instrumento sea malo y que no es posible utilizarlo, si no más bien, debe mejorar en cuanto a su calidad de confiabilidad y validez ya que para este caso son pocas las evaluaciones que en México intentan adentrarse a conocer la comprensión estadística de los estudiantes desde el punto de vista del razonamiento y el pensamiento y que algunos autores de otros países ya han comenzado a estudiar.

Finalmente, CAOS-4 debe ser revisada a la luz de los planes de estudio para adecuar cada uno de los ítems que no han logrado un índice de discriminación y de dificultad óptimos de tal forma que la prueba sea estandarizada para evaluar no solo la alfabetización, sino también el razonamiento y la comprensión estadística en el contexto mexicano.

Al realizar el análisis de discriminación de los ítems se obtiene como resultado que la prueba no puede ser usada con todos sus reactivos ya que gran parte de ellos deben ser revisados a profundidad para mejorar la calidad de la prueba. Tomar en cuenta las diez dimensiones de CAOS-4 para reformular las preguntas supondría llevar un nuevo proceso que tiene como fin último llegar a lograr la confiabilidad y validez de cada uno de los componentes de la prueba y así consolidar una evaluación a gran escala que mida la competencia estadística.

Tablas y figuras

Gráfica 1. Frecuencia de palabras.

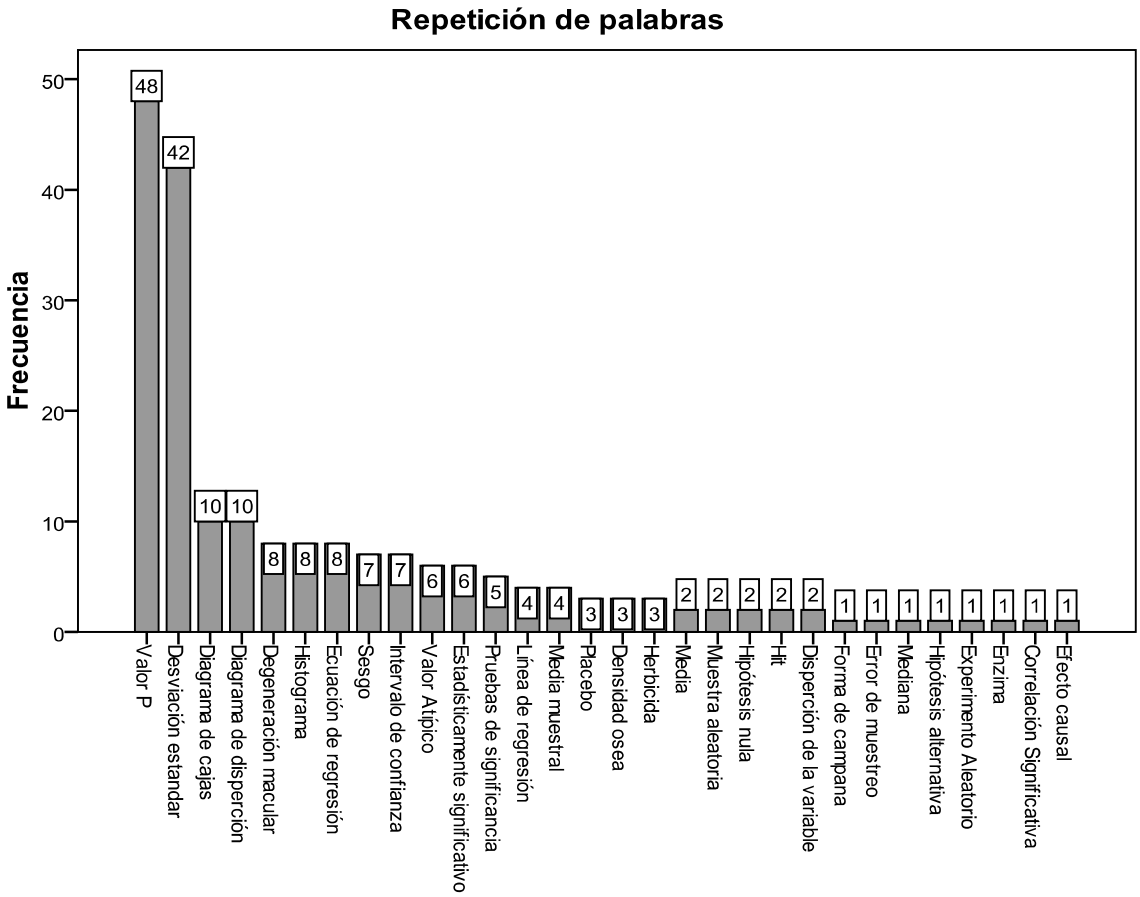

Tabla 1. Promedio de índice de dificultad y poder de discriminación de la prueba

Dimensión	*Poder de discriminación*	*Índice de dificultad*
Recolección de datos	0.14 Pobre, revisar a profundidad	0.38 (medianamente difícil)
Estadística descriptiva	0.10 Pobre, revisar a profundidad	0.4 (mediana mente difícil)
Representación gráfica	0.13 Pobre, revisar a profundidad	0.31 (altamente difícil)
Diagrama de cajas	0.11 Pobre, revisar a profundidad	0.33 (altamente difícil)
Distribución normal	0.13 Pobre, revisar a profundidad	0.35 (medianamente difícil)
Datos bivariados	0.18 Pobre, revisar a profundidad	0.55 (dificultad media)
Probabilidad	0.16 Pobre, revisar a profundidad	0.28 (altamente difícil)
Variabilidad de la muestra	0.18 Pobre, revisar a profundidad	0.34 (medianamente difícil)
Intervalos de Confianza	0.04 Pésima, descartar	0.40 (medianamente difícil)
Pruebas de significancia	0.11 Pobre, revisar a profundidad	0.39 (medianamente difícil)

Total 40 ítems	0.13 Pobre, revisar a profundidad	0.37 medianamente difícil

Bibliografía

Aiken, L.R. (1996). *Tests psicológicos y evaluación*. México: Prentice Hall. Hispanoamericana.

Backhoff, 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,* 2(1). Consultado el 21 de abril de 2013 en: http://redie.uabc.mx/vol2no1/contenido-backhoff.html

Ben-Zvi, D., y Garfield, J. (2004). Statistical Literacy, Reasoning, and Thinking: Goals, Definitions, and Challenges. En Ben-Zvi, D., y Garfield, J. (Eds.), *The Challenge of Developing Statistical Literacy, Reasoning and Thinking* (pp. 5)

Garfield, J., delMas B., Chance, B., Poly, C., Ooms, A. (2005). Comprehensive Assessment of Outcomes for a first course in Statistics (CAOS). *Web ARTIST Project.* Consultado el 3 de octubre de 2011 en: https://app.gen.umn.edu/artist/

Garfield, J., delMas B., Chance, B., Poly, C., Ooms, A. (2006). Summary Statistics for a National Sample of Undergraduates. *Web ARTIST Project.* Consultado el 3 de octubre de 2011 en: https://app.gen.umn.edu/artist/

delMas, R., Garfield, J., Ooms, A., & Chance, B. (2007). "Assessing students' conceptual understanding after a first course in statistics". *Statistics Education Research Journal* 6(2):28-58.

Pérez, J., Acuña, 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. *Revista-Cuadernos* 53 (2). Consultado el 1 de febrero de 2013 en: http://www.revistasbolivianas.org.bo/pdf/chc/v53n2/v53n2a03.pdf