Uso del test CAOS en la investigación del proceso de enseñanza-aprendizaje en el curso introductorio de estadística en la UVG

Annelisse Balsells de Martini[1] & Nancy Zurita de Calgua[2]
[1]Catedrática de Estadística, [2]Directora, Departamento de Matemática, Facultad de Ciencias y Humanidades, Universidad del Valle de Guatemala
amartini@uvg.edu.gt - nazurita@uvg.edu.gt

RESUMEN: El presente documento describe los resultados de la aplicación del Test CAOS traducido, contextualizado y validado por parte del Departamento de Matemática, a estudiantes de 12 secciones del curso de Modelos Estadísticos 1 (CU109) en el segundo semestre del año 2014, en la Universidad del Valle de Guatemala. Se analizaron los puntajes porcentuales del test CAOS completo o Nota Total y los puntajes porcentuales del Test CAOS de las preguntas de temas incluidos en el curso, llamado Nota de Contenido. El test se aplicó antes y después del curso. Todos los análisis de confiabilidad con el coeficiente Alfa de Cronbach, del pre y post test, para Nota Total y de Contenido, tuvieron resultados menores a 0.7. En el post test, en las Notas de Contenido y Totales, los estudiantes del curso CU109, del segundo semestre del 2014, en promedio respondieron correctamente poco más del 50% del test CAOS. En el Post test las notas porcentuales promedio de razonamiento fueron significativamente más altas que las de alfabetización. Se observó una diferencia significativa de Nota de Contenido promedio entre pre y post test, siendo esta diferencia de por lo menos 10 puntos porcentuales. Todas las secciones mejoraron significativamente la Nota de Contenido en el post test. La correlación de Spearman entre la nota final del curso CU109 y la nota de Contenido del post test es significativa, pero baja de 0.13. Se recomienda analizar exhaustivamente cada ítem de la prueba, modificar y construir algunos ítems relacionados a los temas de razonamiento o alfabetización estadística para lograr mejorar la confiabilidad del instrumento y hacer estudios posteriores. Además, se recomienda investigar cualitativamente si es posible determinar cuáles son las preguntas del test CAOS que los estudiantes responden incorrectamente y sus posibles causas, en particular el área de alfabetización o cultura estadística.

PALABRAS CLAVE: razonamiento estadístico, alfabetización o cultura estadística, test CAOS, confiabilidad, validez de contenido.

The use of the CAOS test in the research of the teaching-learning process of the Statistics introductory course at UVG.

ABSTRACT: The following article describes the results obtained from the CAOS test application to students grouped in 12 sections and enrolled in the 2014-second semester in the Statistical Model class (CU109) at Universidad del Valle de Guatemala (UVG). The test has been translated and validated by the Mathematics Department at UVG. The complete CAOS percent scores (titled "Total Score"), as well as the percent scores of the CAOS test questions part of the content of the course above (titled "Contents' Score") were analyzed. The test took place before and after the course. All the confidence analysis with Cronbach's Alpha, pre and post test for both the Total Score and the Contents' Score presented results that were less than 0.7. On the post test, in the Content's Score and Total Score, the students typically responded correctly just over 50% of the CAOS test. The mean percent scores of reasoning questions were significantly higher than the literacy ones in the post test. The pre and post test in the Contents' Score showed a significant difference of at least 10%. All 12 sections increased the Contents' Score in the post test. The Spearman's correlation between the Contents' Score in the post test and the final grade in the course was significant, but low with a result of 0.13. It is recommended to exhaustively analyze every item on the test, as well as modify and build other items related to statistical reasoning or literacy in order to increase the reliability of the instrument and perform further studies. Furthermore, it's also recommended to qualitatively investigate if it is possible to determine which are the CAOS test questions that students respond incorrectly and the possible causes behind this, in particular in the area of statistical literacy.

KEYWORDS: Statistical reasoning, statistical literacy, CAOS test, reliability, content validity.

Introducción

Test CAOS

En 2006 se finalizó el desarrollo del test CAOS (Comprehensive Assessment of Outcomes for a first course in Statistics) por parte del proyecto ARTIST (Assessment Resource Tools for Improving Statistical Thinking). El proyecto fue financiado por NSF (National Science Foundation) y los principales investigadores fueron Joan Garfield, Bob delMas, Ann Ooms de la Universidad de Minnesota y Beth Chance del Politécnico de California. El test CAOS se desarrolló a través de un proceso iterativo que duró tres años. Durante este tiempo se adquirieron ítems de parte de instructores, se escribieron ítems no cubiertos por los adquiridos, se revisaron ítems, se obtuvo retroalimentación por parte de asesores y examinadores, y se condujeron dos grandes evaluaciones de validez de contenido. Una primera versión del CAOS fue producida en 2004 y consistió en 34 ítems de selección múltiple. Se utilizó como estudio piloto en el otoño de 2004. Se aplicó a los estudiantes de cursos de estadística introductoria a nivel terciario. Los datos de este estudio piloto se usaron para hacer revisiones que resultaron en una segunda versión (CAOS 2), que consistió en 37 ítems de selección múltiple. Esta segunda versión se probó en la primavera de 2005 con una muestra de 100 estudiantes de nivel secundario y 800 de nivel terciario. El análisis de resultados de esta prueba se usó para hacer cambios adicionales que resultaron en la tercera versión de CAOS (CAOS 3). Un grupo de 30 instructores de estadística seleccionados por el proyecto, realizaron otra ronda de validación del CAOS 3. Esta retroalimentación se usó para agregar o eliminar ítems, como también para hacer revisiones extensas y producir la versión final, llamada CAOS 4. Este test consiste en 40 preguntas de selección múltiple de las cuales 35% corresponden al área de alfabetización o cultura estadística (LITERACY) y la otra parte al razonamiento estadístico (REASONING). En marzo de 2006 se realizó un análisis final de validez de contenido y en el otoño del 2006 se administró a una muestra de 1470 estudiantes de 35 instructores diferentes, provenientes de 33 instituciones de educación superior, de 21 estados de los Estados Unidos de América. El análisis de consistencia interna de esta última versión del test CAOS de 40 ítems de selección múltiple, aplicada como post test después del curso introductorio de estadística, produjo un coeficiente de alfa de Cronbach de 0.82. Esto indica una consistencia interna aceptable para estudiantes a nivel terciario, de un curso introductorio, no matemático, de estadística (DelMas, 2007).

La alfabetización o cultura estadística se refiere a la habilidad clave esperada de ciudadanos de sociedades, en que la información es un andamiaje. Comprende el entender y usar el lenguaje y herramientas básicas estadísticas; sabiendo qué significan términos estadísticos básicos. También comprende el entender el uso de símbolos estadísticos simples y reconocer y ser capaz de interpretar diferentes representaciones de la información. Razonamiento estadístico es la forma en que las personas razonan con ideas estadísticas y hacen sentido de la información estadística. Puede incluir el conectar un concepto con otro (ej. centro y dispersión) o puede combinar ideas acerca de datos y probabilidad. También significa entender y ser capaz de explicar procesos estadísticos, y ser capaz de interpretar resultados estadísticos. Para los docentes, el razonamiento estadístico son las representaciones y conexiones mentales que los estudiantes tienen con respecto a conceptos estadísticos. Expertos reportan que a pesar de los repetidos llamados a cambiar el contenido y pedagogía del curso introductorio de estadística a nivel terciario, no hay evidencia que se hayan hecho cambios sustanciales y que hayan mejorado los resultados de los estudiantes. Esto lo muestra la evidencia de los resultados de este primer curso de estadística en los Estados Unidos de 13,917 estudiantes universitarios de pregrado sometidos a una evaluación llamada "Comprehensive Assessment of Outcomes in Statistics" (CAOS) a lo largo de un período de 6 años. El análisis de los datos indica que el rendimiento de los estudiantes se ha mantenido estable del 2005 al 2011. Esto ha evidenciado la necesidad de cambiar radicalmente el currículo para el curso de estadística introductoria (Garfield, 2012).

Construcción de pruebas

La psicometría es la ciencia que se encarga de la construcción de tests o pruebas psicopedagógicas. Dicha ciencia propone que en la construcción de pruebas debe ser un procedimiento estandarizado en el que los ítems pasan por un proceso de construcción, selección y organización, tal que puedan provocar en una persona ciertas reacciones controladas que se puedan replicar. Para que una prueba tenga dichas características debe de cumplir con los siguientes requisitos:

1. El contenido y la dificultad de los ítems están sistemáticamente controlados (construcción del test).
2. La situación de aplicación del test: el ambiente en el cual se le administra, el material del test, la administración, debe estar bien definida y debe ser reproducida idénticamente para todos los sujetos examinados con el test.
3. El registro del comportamiento provocado en el sujeto examinado debe ser preciso y objetivo. Las condiciones de cómo hacer este registro deben estar definidas y deben ser cumplidas rigurosamente.
4. El comportamiento registrado debe ser evaluado estadísticamente con respecto al de un grupo de individuos llamado grupo de referencia o normativo.
5. Los sujetos examinados son clasificados en función de normas resultantes del examen previo del grupo de referencia o normativo (baremo), lo que permite situar cada una de las respuestas, totales o parciales, en una distribución estadística (contraste).
6. Las respuestas a las cuestiones planteadas dan una medida correcta del comportamiento al que el test apunta (validez).
7. Si las condiciones no cambian, la repetición del examen debe conducir siempre al mismo resultado, o a otro muy próximo (confiabilidad) (Aliaga, 2007).

Validez de contenido

Se define la validez de contenido como la validez que trata de determinar hasta donde los ítems de un instrumento de evaluación representan el dominio del contenido que se desea medir. Para lograr esto los investigadores construyen una alta cantidad de ítems de un dominio determinado (Ruiz, 2002).

No importa que tan buenos sean los ítems la validez siempre es dudosa. Por lo tanto, como no es cuantificable a través de algún índice o coeficiente se recurre al juicio de expertos de la siguiente manera:

- Se seleccionan dos jueces o expertos, por lo menos, a los fines de juzgar, de manera independiente, la "bondad" de los ítems del instrumento, en términos de la relevancia o congruencia de los reactivos con el universo de contenido, la claridad en la redacción y la tendenciosidad o sesgo en la formulación de los ítems.
- Cada experto recibe suficiente información escrita acerca de: (a) el propósito de la prueba; (b) conceptualización del universo de contenido; (c) plan de operacionalización o tabla de especificaciones (en el caso de las pruebas de rendimiento académico).
- Cada juez recibe un instrumento de validación en el cual se recoge la información de cada experto. Dicho instrumento normalmente contiene las siguientes categorías de información por cada ítem: congruencia ítem-dominio, claridad, tendenciosidad y observaciones
- Se recogen y analizan los instrumentos de validación y se toman las decisiones siguientes: (a) los ítems donde hay un 100 por ciento de coincidencia favorable entre los jueces (los ítems son congruentes, están escritos claramente y no son tendenciosos) quedan incluido en el instrumento; (b) los ítems donde hay un 100 por ciento de coincidencia desfavorable entre los jueces, quedan excluidos del instrumento; y (c) los ítems donde sólo hay coincidencia parcial entre los jueces deben ser revisados, reformulados, si es necesario, y nuevamente validados Kerlinger (2002).

Confiabilidad

La confiabilidad de las pruebas pretende determinar si al aplicarlas a distintos grupos de personas ésta brindará resultados parecidos. Se refiera a la replicabilidad y estabilidad. Siempre y cuando se apliquen en condiciones similares o equivalentes. Spearman fue el primero en proponer un coeficiente de confiabilidad basado en la variabilidad de las observaciones. Sin embargo, el coeficiente de confiabilidad más utilizado es el Alfa de Cronbach, que es una mejora del coeficiente de Spearman que utiliza correlaciones entre los ítems. El coeficiente Alfa de Cronbach se calcula por medio de la fórmula:

$$\alpha = \left(\frac{k}{k\text{-}1}\right)\left(1 - \frac{\Sigma S_i^2}{S_{sum}^2}\right)$$

Donde k, es el número de ítems, S_i^2 es la varianza de los ítems y S_{sum}^2 es la varianza total (Ruiz, 2002).

En general, el coeficiente de confiabilidad de cualquier prueba es el cálculo de la correlación promedio de esa prueba con todas las demás de igual cantidad de ítems que son obtenibles por medio del dominio-muestra. La idea es comunicar en qué medida se pueden repetir los resultados obtenidos con un método de medición. Es decir, que tan eficaz es el instrumento. La mayoría de autores proponen que un índice que exceda de 0.7 indica que la prueba puede ser utilizada tanto para investigación como para la toma de decisiones. Sin embargo, una confiabilidad mayor a 0.9 tendría demasiados ítems y requeriría demasiado esfuerzo y tiempo de calibración y de aplicación. Se ha encontrado que un instrumento de 40 ítems genera confiabilidades en rangos aceptables. Sin embargo, en las primeras etapas de la investigación sobre pruebas predictivas se puede ahorrar tiempo y energía con instrumentos de reducida confiabilidad y bastan las confiabilidades de 0.5 a 0.6 (Nunnnally, 1987).

Este trabajo se centró en validar el test CAOS para evaluar el rendimiento de los estudiantes después del curso introductorio de estadística (Modelos Estadísticos 1 CU109) que estudiantes de las Facultades de Ciencias y Humanidades, Ingeniería y Ciencias Sociales toman en la Universidad del Valle de Guatemala. Por otro lado, determinar la confiabilidad del test CAOS que se tradujo y se contextualizó. Además, comparar los resultados del CAOS antes y después del curso introductorio de estadística, como comparar los resultados entre diferentes secciones que tomaron el curso en el segundo semestre de 2014. Por último, se buscó analizar si hay una correlación entre el resultado del CAOS después del curso y la nota obtenida en el mismo.

Metodología

Debido al interés del departamento de matemática de la Universidad del Valle de Guatemala (UVG) en adoptar el test CAOS, se tradujo y se contextualizó el test en el año 2013. El interés consistió en medir a qué nivel de razonamiento o alfabetización estadística se está llevando a los estudiantes del curso actual de Modelos Estadísticos 1 (CU109) de la UVG.

En la primera fase de la investigación, en 2014, se validó el test traducido con el apoyo de los profesores de estadística del departamento de Matemática de la UVG. Para esta validación cada profesor leyó cada pregunta y comentó si estaba bien traducida, redactada y se entendía lo que se pretendía preguntar. Con los comentarios de los profesores se llegó a un consenso de cómo debería quedar cada pregunta. Después de esto se hizo una clasificación de las preguntas para determinar cuáles eran de alfabetización estadística y cuáles de razonamiento estadístico.

El test se les administró dos veces a todos los estudiantes del segundo semestre del año 2014, inscritos en el curso de Modelos Estadísticos 1; el primer día de clases (Pre) y al finalizar el curso

(Post). Los estudiantes en este semestre estuvieron distribuidos en 12 secciones de diferentes tamaños y arregladas básicamente por carreras.

De las 40 preguntas del test solamente se tomaron 33 para calcular la nota porcentual de los estudiantes. Las 7 preguntas que no se tomaron en la evaluación, se refieren a temas que no se abordaron en el curso de Modelos Estadísticos 1 de la UVG en el segundo semestre del 2014. A esta nota se le llamó Nota Contenido del curso CU109. Con el propósito de comparación futura de nuestros estudiantes con estudiantes en el extranjero, también se conservó la nota porcentual sobre las 40 preguntas del test. A esta nota se le llamó Nota Total. Para hacer comparaciones se analizaron las notas de 255 estudiantes que tomaron ambos tests, tanto el Pre, como el Post.

Para el análisis primero se calculó el coeficiente de confiabilidad alfa de Cronbach a la Nota de Contenido y a la Nota Total, tanto en el pre test, como en el post test. El alfa de Cronbach para el Post Test Nota Total se comparó con el reportado por el CAOS en inglés en la versión final. Luego se hicieron pruebas de normalidad de Kolmogorov-Smirnov y de Shapiro-Wilk a las siguientes notas: Pre Test Nota Total, Pre Test Nota Contenido, Post Test Nota Total, Post Test Nota Contenido, como también para las cuatro notas, pero por sección. Además para las notas del Post Test siguientes: Nota Total Alfabetización, Nota Contenido Alfabetización, Nota Total Razonamiento y Nota Contenido Razonamiento. Todo esto con el fin de definir si usar pruebas paramétricas o no paramétricas en ciertas partes del análisis. Con el propósito de comparación y donde el tamaño de muestra permitía aplicar el Teorema del Límite Central, se construyeron intervalos de confianza para las medias de las siguientes notas: Pre y Post Test de Nota Total y Pre y Post de Nota Contenido, Post Test Nota Total Alfabetización, Post Test Nota Total Razonamiento, Post Test Nota Contenido Alfabetización y Post Test Nota Contenido Razonamiento. Para comparar entre el Pre y el Post Test para todos los estudiantes en general, se pudo hacer una prueba t pareada de comparación de medias, ya que aunque las notas en estos grupos no eran normales, la gran cantidad de estudiantes (más de 30), permite utilizarla. Para determinar si no todas las notas promedio Total y Contenido por sección son iguales, se realizó una prueba no paramétrica de Kruskal-Wallis en el Pre y en el Post Test. Mientras que para las comparaciones entre Pre y Post por sección, tanto para Notas Total, como Nota de Contenido, debido a que varias secciones tenían menos de 30 estudiantes y no todas con distribución normal, se utilizaron pruebas no paramétricas de ranking con signos de Wilcoxon. Por último, para analizar la correlación entre la nota final del curso CU109 y la Nota de Contenido en el Post Test, se realizó una prueba no paramétrica de correlación de Spearman.

Resultados y discusión

1. Coeficiente de confiabilidad de Cronbach:

El coeficiente de confiabilidad Alfa de Cronbach (Tabla 1) en los cuatro análisis pre y post se encontró menor a 0.7. Dado que en esta investigación se buscaba medir el nivel de razonamiento o alfabetización estadística en estudiantes de primer año de la UVG, a través de la traducción de una prueba de inglés a español (de la prueba CAOS), se podía esperar que el coeficiente Alfa de Cronbach fuera bajo y menor al que se le calculó a la prueba original (0.83). La razón anterior se debe a que es la primera investigación que se hace con dichos reactivos y sugiere que hay que hacer un análisis exhaustivo de cada ítem y descartar los que no correlacionan directamente con la prueba. Podemos decir que la prueba se construyó con validez de contenido ya que los catedráticos que imparten el curso de Modelos Estadísticos 1 relacionaban los ítems a lo que se enseña en el aula pero no se podría asegurar que se puede replicar los resultados obtenidos. Adicional a esto, se puede observar que el índice aumenta de pre a post. Esto se debe a que hay un impacto en los estudiantes al recibir en curso en aumentar sus conocimientos estadísticos, por lo que sus respuestas tienden a ser más homogéneas.

Tabla 1. Coeficientes de confiabilidad de Cronbach

Coeficiente de confiabilidad de Cronbach de Pre Test Nota Total: 0.4
Coeficiente de confiabilidad de Cronbach de Pre Test Nota Contenido: 0.36
Coeficiente de confiabilidad de Cronbach de Post Test Nota Total: 0.6
Coeficiente de confiabilidad de Cronbach de Post Test Nota Contenido: 0.53

2. Tests de normalidad de notas:

A las notas de Pre Test Nota Total, Pre Test Nota Contenido, Post Test Nota Total y Post Test Nota Contenido, se le aplicaron las pruebas de normalidad de Kolmogorov-Smirnov y la de Shapiro-Wilk. En ninguno de los casos hubo evidencia que las notas fueron normales.

También se comprobó la normalidad con las mismas dos pruebas de las notas por sección, para cada uno de los 4 grupos de notas. Algunas secciones cumplieron con una distribución normal y en la mayoría no hubo evidencia de normalidad.

Era importante saber acerca de la normalidad de los diferentes grupos de notas, ya que dependiendo de esto, se elegiría la prueba adecuada para hacer comparaciones de las notas entre el Pre Test y el Post Test para los estudiantes en general y por secciones. En el caso de querer comparar entre el pre y el post para todos los estudiantes en general, se pudo hacer una prueba t pareada de comparación de medias, ya que aunque las notas en estos grupos no eran normales, la gran cantidad de estudiantes (más de 30), permite utilizarla. Mientras que para las comparaciones entre Pre y Post por sección, debido a que varias secciones tenían menos de 30 estudiantes y no todas con distribución normal, se utilizaron pruebas no paramétricas de ranking con signos de Wilcoxon.

3. Intervalos de confianza de media de notas en el Post Test para las Notas Totales y de Contenido:

Observando los resultados de notas del Post test en la Tabla 2, se puede ver que el intervalo de la Nota promedio Total va de

Tabla 2. Intervalos de confianza de media de Nota Total y Contenido en el Post test.

Post				
Descriptivas				
			Estadístico	Error estándar
Nota Total	Promedio		50.9396	.62321
	Intervalo 95% de confianza de la media	Límite inferior	49.7131	
		Límite superior	52.1661	
Nota Contenido	Promedio		53.9760	.64690
	Intervalo 95% de confianza de la media	Límite inferior	52.7029	
		Límite superior	55.2491	

Tabla 3. Intervalos de confianza de media de Nota Total y Contenido en el Post test, separadas en alfabetización y razonamiento.

Post				
Descriptivas				
			Estadistico	Error estándar
Notas Total Alfabetización	Media		42.4497	.75010
	Intervalo 95% de confianza de la media	Límite inferior	40.9735	
		Límite superior	43.9258	
Notas Contenido Alfabetización	Media		46.4308	.84384
	Intervalo 95% de confianza de la media	Límite inferior	44.7701	
		Límite superior	48.0914	
Notas Total Razonamiento	Media		55.5111	.74460
	Intervalo 95% de confianza de la media	Límite inferior	54.0457	
		Límite superior	56.9765	
Notas Contenido Razonamiento	Media		57.7486	.77997
	Intervalo 95% de confianza de la media	Límite inferior	56.2137	
		Límite superior	59.2836	

49.7% a 52.2%, con media de 50.9%. El intervalo de Nota promedio Contenido va de 52.7% a 55.2%, con media de 54%. Esto sugiere que en ambos casos los estudiantes del curso CU109, del segundo semestre del 2014, respondieron correctamente poco más del 50% del test CAOS.

Se calcularon intervalos de confianza en el Post Test para las Notas Totales y de Contenido, pero separadas en alfabetización y razonamiento. Ver Tabla 3.

En el test de 40 preguntas o Notas Total hay 14 preguntas de alfabetización y 26 de razonamiento. En el test de 33 preguntas o Notas Contenido, hay 11 preguntas de alfabetización y 22 de razonamiento. Separando las notas porcentuales en alfabetización y en razonamiento (ver Tabla 3), se observa que las medias de notas porcentuales en razonamiento, tanto en Notas Total, como Notas de Contenido, son más altas que en la parte de alfabetización. No solamente son más altas, sino significativamente más altas, ya que los intervalos de las medias de alfabetización no se traslapan con los correspondientes intervalos de las medias de razonamiento. Se le debe prestar atención a estos resultados, ya que ambas áreas son importantes y el curso debería desarrollar ambas áreas por igual.

Además, a éstas, se les aplicaron las dos pruebas de normalidad:

Tabla 4. Pruebas de normalidad.

Post	Tests de Normalidad	
	Kolmogorov-Smirnova	Shapiro-Wilks
	Sig.	Sig.
Nota Total Alfabetización	.000	.000
Nota Contenido Alfabetización	.000	.000
Nota Total Razonamiento	.000	.002
Nota Contenido Razonamiento	.000	.000

Como se puede ver (Tabla 4), ninguno de estos juegos de notas parece tener una distribución normal, ya que todos los valores p son menores a 0.05 de significancia. Aunque no parecen tener una distribución normal, los tamaños de muestra (n>30) permiten aplicar el Teorema del Límite Central y hacer la inferencia usando los intervalos de confianza sobre la media.

4. Intervalos de confianza de notas promedio para todos los estudiantes en general en el Pre y en el Post test, tanto Notas Total, como Notas Contenido:

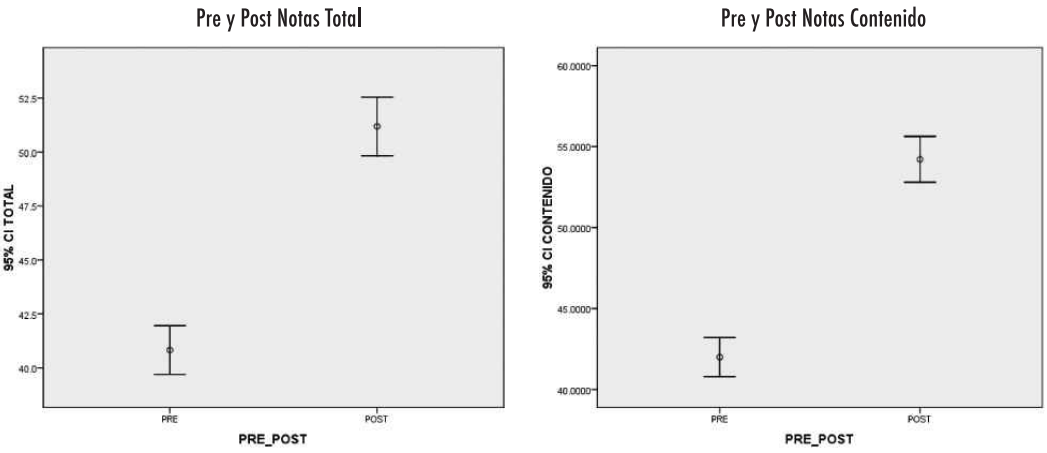
Tabla 5. Intervalos de confianza de la media de Nota Total y Contenido en el Pre test.

Pre				
Descriptivas				
			Estadístico	Error estándar
Nota Total	Promedio		40.9815	.55677
	Intervalo 95% de confianza de la media	Límite inferior	39.8853	
		Límite superior	42.0777	
Nota Contenido	Promedio		42.1437	.59293
	Intervalo 95% de confianza de la media	Límite inferior	40.9763	
		Límite superior	43.3110	

Tabla 6. Intervalos de confianza de la media de Nota Total y Contenido en el Post test.

Post				
Descriptivas				
			Estadístico	Error estándar
Nota Total	Promedio		50.9396	.62321
	Intervalo 95% de confianza de la media	Límite inferior	49.7131	
		Límite superior	52.1661	
Nota Contenido	Promedio		53.9760	.64690
	Intervalo 95% de confianza de la media	Límite inferior	52.7029	
		Límite superior	55.2491	

Gráfico 1. Intervalos de confianza de notas promedio en el Pre y Post.



En el Gráfico 1 se compara gráficamente los intervalos de confianza de notas promedio del Pre y Post Test para las Notas Total y las Notas Contenido.

Comparando los intervalos de nota promedio Pre Test con Post Test, de la Nota Total (ver Tablas 5 y 6), se observa que en la Nota Total el Pre Test va de 40-42, mientras que el Post Test de 50-52. Esto muestra que hay diferencia significativa entre ellos de por lo menos 8 puntos porcentuales. También se puede observar esta diferencia en el Gráfico 1 (izquierda). De igual forma, en las tablas 5 y 6 y en el Gráfico 1 (derecha), se puede apreciar la diferencia significativa de Nota promedio de Contenido ente Pre y Post test de por lo menos 10 puntos porcentuales.

5. Comparación de las notas promedio de Contenido del Pre con el Post Test para todos los estudiantes en general con una prueba t pareada:

Tabla 7. Prueba t pareada de Notas de Contenido del Pre y Post test.

Estadístico muestras emparejadas

	Media	N	Desviación estándar	Error estándar Media
Pre Notas Contenido	41.996435	255	9.7266711	.6091078
Post Notas Contenido	54.212715	255	11.4981311	.7200409

Correlaciones muestras emparejadas

	N	Correlación	Sig.
Pre Notas Contenido & Post Notas Contenido	255	.411	.000

Prueba muestras emparejadas

	Diferencias emparejadas					t	gl	Sig. (2-colas)
	Media	Desviación estándar	Error estándar Media	Intervalo 95% confianza de diferencias Inferior	Superior			
Pre Notas Contenido Post Notas Contenido	-12.216600	11.6120529	.7271750	-13.6483407	-10.7842202	-16.800	254	**.000**

De todas las comparaciones que se hubieran podido hacer entre notas promedio, se eligió hacer la comparación entre Pre y Post Test de las Notas promedio de Contenido, tanto en general, como por sección, ya que esto puede dar información valiosa acerca de si la aplicación del Curso CU109 durante el semestre logró modificar o mejorar esta nota. No interesa la comparación de la Nota Total, ya que en ésta se incluyen ítems que no son parte del contenido del curso. Además de la inferencia hecha con los intervalos de confianza sobre las notas promedio de Contenido, se realizó la prueba t pareada de los 255 estudiantes que tomaron los dos test, para comparar la notas promedio de Contenido entre el Pre y el Post test. De la Tabla 7, ya que el valor p es menor que alfa de 0.05, se puede llegar a concluir que hay diferencia significativa de las notas promedio de Contenido, entre Pre y Post, mostrándose una mejor nota promedio en el Post de por lo menos 10%.

6. Intervalos de confianza para notas promedio por sección para el Pre y el Post Test, tanto Nota Total, como Nota Contenido:

Observando el Gráfico 2 de los intervalos de las medias de Notas Totales y Notas de Contenido por sección en el Pre Test, llama la atención la sección 110. Ésta se muestra por encima de varios intervalos y no se traslapa con todos los intervalos de las demás secciones. Esto se puede deber a un grupo con mayores conocimientos previos o con una mejor actitud. Esto podría estudiarse en investigaciones cualitativas posteriores. Luego, si se observan los intervalos de notas promedio en el Gráfico 2 del Post Test, nuevamente el intervalo para la media de Notas Totales y Notas de Contenido para la sección 110, se muestra por encima de varios de los otros intervalos, marcando una diferencia significativa. En el Post Test la diferencia es más marcada. Después del curso, siempre para la sección 110, en

Gráfico 2. Intervalos de confianza de las notas promedio en el Pre y Post test por sección.

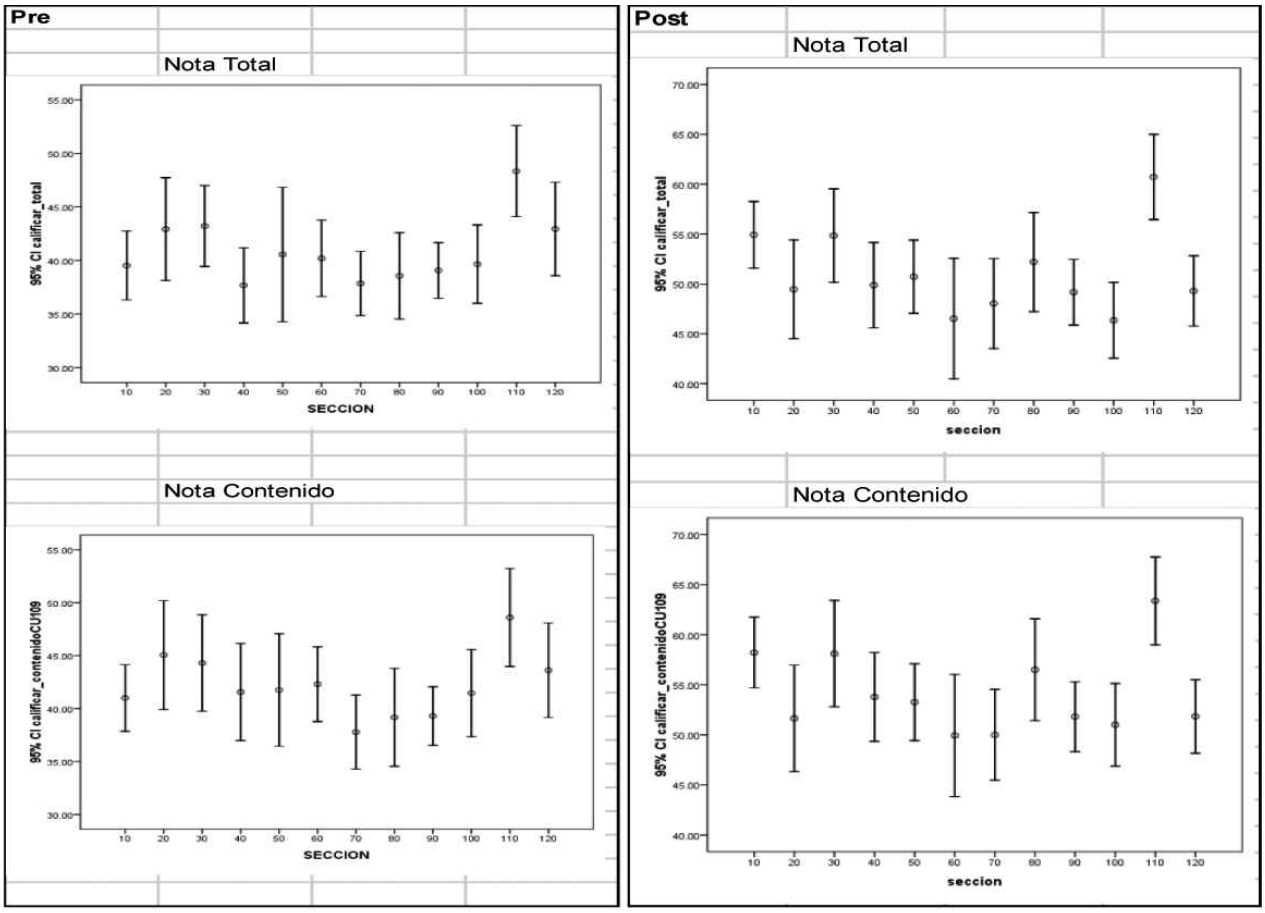


Tabla 8. Prueba Kruskal-Wallis de notas de Pre y Post test.

Pre		
Estadístico de prueba [a,b]		
	Notas Total	Notas Contenido
Chi-cuadrado	23.642	20.744
gl	11	11
Sig. Asimp.	.014	.036
a. Test Kruskal Wallis		
b. Variable agrupación: SECCION		

Post		
Estadístico de prueba [a,b]		
	Notas Total	Notas Contenido
Chi-cuadrado	35.909	33.369
gl	11	11
Sig. Asimp.	.000	.000
a. Test Kruskal Wallis		
b. Variable agrupación: SECCION		

el post test la nota promedio Total va de 56 a 65 y para la de Contenido de 59 a 68. Esta sección presenta las mejores notas, tanto en el Pre, como en el Post test. A excepción de esta sección, todos los demás intervalos de notas promedio se traslapan entre sí. Indicando que el resto de las secciones no presenta resultados significativamente diferentes. Debido a la falta de normalidad de las notas por sección, para poder confirmar lo antes mencionado se realizaron pruebas de Kruskal-Wallis en el Pre y en el Post test.

En la Tabla 8, debido a que el valor p de las cuatro pruebas es menor a alfa de 0.05, se puede concluir que las notas por sección, tanto Totales, como de Contenido, no son todas iguales, tanto en el Pre, como también en el Post Test. En otras palabras, si hay diferencias significativas de notas entre las 12 secciones, tanto en el Pre, como en el Post test. Puede ser que la sección 110 sea la que haga que la prueba de Kruskal-Wallis lleve a concluir que hay diferencias significativas. Esto no se comprobó.

Para visualizar las notas promedio por sección en el Pre y en el Post se graficaron los intervalos de confianza sobre las notas promedio. A continuación se muestran los gráficos por sección para las Notas Totales (ver Gráfico 3) y para las Notas de Contenido (ver Gráfico 4).

En los Gráficos 3 y 4 se puede observar que en algunas secciones los intervalos de nota promedio de Pre y Post si se traslapan y en otras secciones no.

Debido a que las notas no siguen una distribución normal en la mayoría de las secciones y los tamaños de sección son de 35 o menos (hay una de 13), no es recomendable hacer inferencia con los intervalos de confianza sobre las notas promedio, ya que el tamaño de la muestra no es suficientemente grande para que se cumplan los requerimientos del teorema del límite central. Además hay datos atípicos en muchas de las secciones.

Para comparar las Notas de Contenido, entre el pre y post, pero por sección, se hicieron pruebas no paramétricas de Ranking con Signos de Wilcoxon. A continuación se presentan los resultados (ver Tabla 9).

En la columna a la derecha de la Tabla 9, se puede observar que todos los valores p de todas las secciones son menores a alfa de 0.05, por lo que se puede concluir que en todas las secciones hay una diferencia significativa de la notas de Contenido, siendo mejores las notas en el Post test. Esto es importante porque indica que todas las secciones mejoraron significativamente la Nota de Contenido en el Post Test.

7. Coeficiente de correlación de Spearman de las notas de todos los estudiantes de Contenido con las notas finales del curso CU109:

Era de interés investigar si había correlación entre las notas finales del curso CU109 y las notas de Contenido del Post test. Para esto se calculó y se hizo la prueba de significancia de correlación de Spearman. Como se puede ver en la Tabla 10, la correlación es significativa, pero se considera baja, 0.13. La evidencia de que la nota del curso CU109 tenga correlación con la nota de Contenido del Post test, aunque significativa, es muy poca, como para tratar de explicar que la mejora de la nota de Contenido del Pre al Post test se deba en parte al curso.

Recomendaciones

Luego de analizar y discutir los resultados de esta primera aplicación de la prueba CAOS traducida al español se hacen las siguientes recomendaciones.

1. Hacer análisis exhaustivo de cada ítem de la prueba, modificar y construir algunos ítems relacionados a los temas de razonamiento o alfabetización estadística. Aplicarlo, analizarlo y modificarlo las veces que sea necesaria para lograr mejorar la confiabilidad del instrumento. Una vez confiable se podrá utilizar para tomar decisiones sobre los estudiantes de la UVG y hacer comparaciones en las diferentes cohortes de estudiantes y el impacto del curso sobre los resultados. Por otro lado, medir el énfasis que el curso de Modelos Estadísticos 1 da en las dos áreas, alfabetización y razonamiento, que mide el test CAOS.

Gráfico 3. Intervalos de confianza de Nota promedio Total en Pre y Post test por sección.

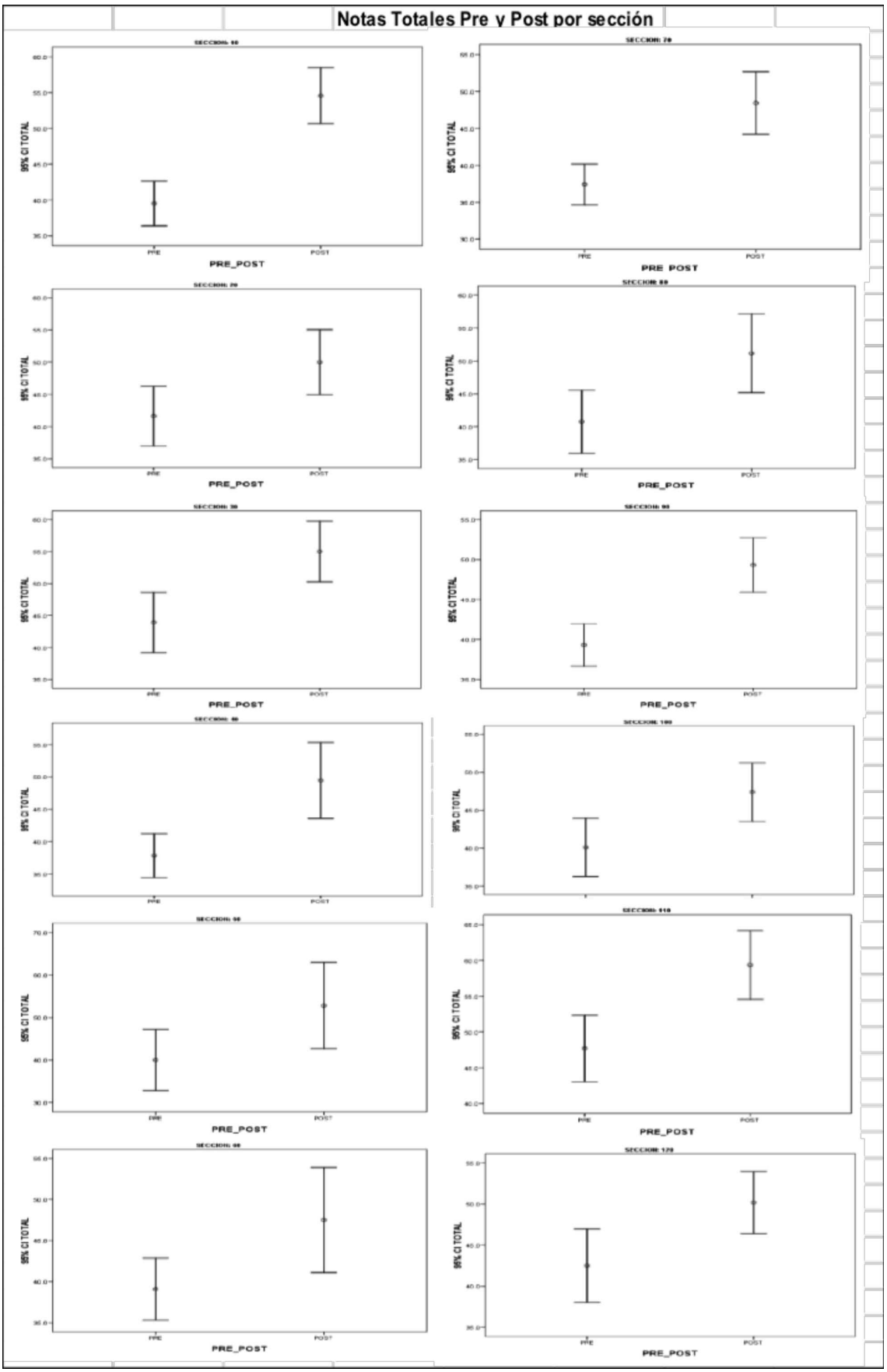


Gráfico 4. Intervalos de confianza de Nota promedio de Contenido en Pre y Post test por sección.

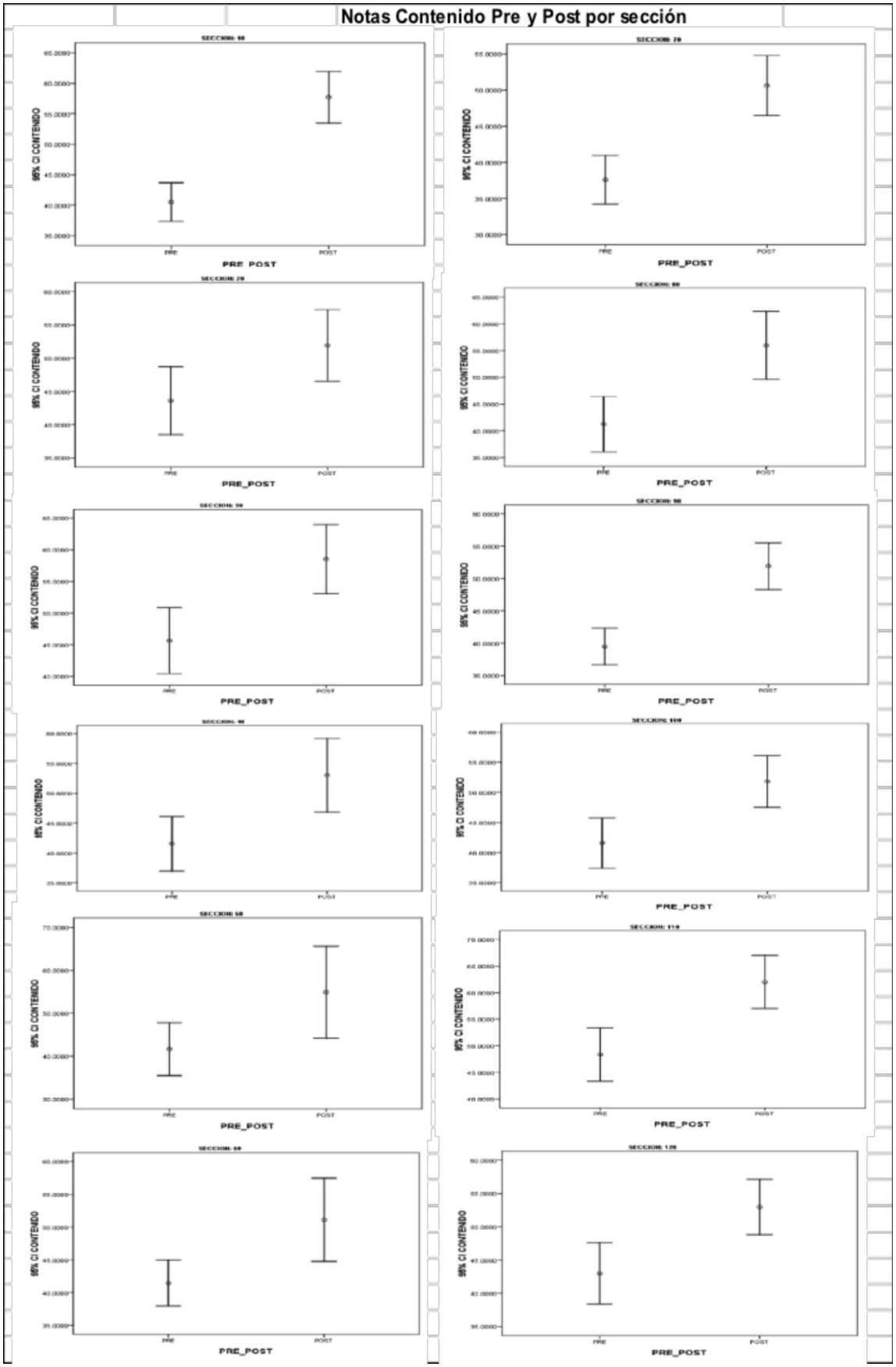


Tabla 9. Pruebas de ranking con signo de Wilcoxon de Notas de Contenido por sección.

Prueba de rankings con signo de Wilcoxon

Rankings

SECCION			N	Ranking promedio	Suma de Rankings
10	Post Notas Contenido - Pre Notas Contenido	Rankings negativos	2[a]	2.50	5.00
		Rankings positivos	28[b]	16.43	460.00
		Vínculos	0[c]		
		Total	30		
20	Post Notas Contenido - Pre Notas Contenido	Rankings negativos	6[a]	7.42	44.50
		Rankings positivos	15[b]	12.43	186.50
		Vínculos	2[c]		
		Total	23		
30	Post Notas Contenido - Pre Notas Contenido	Rankings negativos	3[a]	3.33	10.00
		Rankings positivos	12[b]	9.17	110.00
		Vínculos	1[c]		
		Total	16		
40	Post Notas Contenido - Pre Notas Contenido	Rankings negativos	1[a]	4.50	4.50
		Rankings positivos	12[b]	7.21	86.50
		Vínculos	1[c]		
		Total	14		
50	Post Notas Contenido - Pre Notas Contenido	Rankings negativos	0[a]	.00	.00
		Rankings positivos	8[b]	4.50	36.00
		Vínculos	0[c]		
		Total	8		
60	Post Notas Contenido - Pre Notas Contenido	Rankings negativos	3[a]	9.50	28.50
		Rankings positivos	18[b]	11.25	202.50
		Vínculos	1[c]		
		Total	22		
70	Post Notas Contenido - Pre Notas Contenido	Rankings negativos	2[a]	7.00	14.00
		Rankings positivos	24[b]	14.04	337.00
		Vínculos	1[c]		
		Total	27		
80	Post Notas Contenido - Pre Notas Contenido	Rankings negativos	0[a]	.00	.00
		Rankings positivos	12[b]	6.50	78.00
		Vínculos	1[c]		
		Total	13		
90	Post Notas Contenido - Pre Notas Contenido	Rankings negativos	1[a]	3.00	3.00
		Rankings positivos	24[b]	13.42	322.00
		Vínculos	4[c]		
		Total	29		
100	Post Notas Contenido - Pre Notas Contenido	Rankings negativos	4[a]	7.88	31.50
		Rankings positivos	17[b]	11.74	199.50
		Vínculos	1[c]		
		Total	22		
110	Post Notas Contenido - Pre Notas Contenido	Rankings negativos	1[a]	3.00	3.00
		Rankings positivos	23[b]	12.91	297.00
		Vínculos	0[c]		
		Total	24		
120	Post Notas Contenido - Pre Notas Contenido	Rankings negativos	5[a]	9.30	46.50
		Rankings positivos	20[b]	13.93	278.50
		Vínculos	2[c]		
		Total	27		

a. Post Notas Contenido < Pre Notas Contenido

b. Post Notas Contenido > Pre Notas Contenido

c. Post Notas Contenido = Pre Notas Contenido

Estadísticos de Prueba[b]

SECCION		Post Notas Contenido - Pre Notas Contenido
10	Z	-4.687[a]
	Sig. Asimp. (2-colas)	**.000**
20	Z	-2.471[a]
	Sig. Asimp. (2-colas)	**.013**
30	Z	-2.848[a]
	Sig. Asimp. (2-colas)	**.004**
40	Z	-2.871[a]
	Sig. Asimp. (2-colas)	**.004**
50	Z	-2.527[a]
	Sig. Asimp. (2-colas)	**.012**
60	Z	-3.031[a]
	Sig. Asimp. (2-colas)	**.002**
70	Z	-4.114[a]
	Sig. Asimp. (2-colas)	**.000**
80	Z	-3.062[a]
	Sig. Asimp. (2-colas)	**.002**
90	Z	-4.310[a]
	Sig. Asimp. (2-colas)	**.000**
100	Z	-2.928[a]
	Sig. Asimp. (2-colas)	**.003**
110	Z	-4.224[a]
	Sig. Asimp. (2-colas)	**.000**
120	Z	-3.141[a]
	Sig. Asimp. (2-colas)	**.002**

a. Basado en rankings negativos

b. Prueba de rankings con signo de Wilcoxon

Tabla 10. Coeficientes de Correlación de Spearman Nota final del curso – Notas Contenido Post test.

Correlaciones de Spearman				
			Nota final curso CU109	Post Notas Contenido
Ro de Spearman	Nota final curso CU109	Coeficiente de Correlación	1.000	**.136***
		Sig. (2-colas)	.	**.019**
		N	297	297
	Post Notas Contenido	Coeficiente de Correlación	**.136***	1.000
		Sig. (2-colas)	**.019**	.
		N	297	297
*. Correlaciónn es significativa a nivel de 0.05 (2-colas).				

2. Investigar cualitativamente si es posible determinar cuáles son las preguntas del test CAOS que los estudiantes responden incorrectamente y sus posibles causas. Es importante notar que las preguntas del test CAOS que corresponden a alfabetización (LITERACY), fueron las que menor puntaje obtuvieron por lo que se sugiere iniciar en esa área.

Agradecimiento

Se agradece al Departamento de Matemática y el equipo de catedráticos del curso de Modelos Estadísticos 1 (CU109) de la Universidad del Valle de Guatemala que apoyó para la validación y la aplicación de las pruebas. Además, al Departamento de Psicología Educativa de la Universidad de Minnesota en Estados Unidos de América al permitir la traducción y contextualización del Test CAOS para implementarlo en la UVG.

Bibliografía

Aliaga, J. (2007). *Psicometría: tests psicométricos, confiabilidad y validez.* Tomada el 27 de julio de 2016 en: http://datateca.unad.edu.co/contenidos/401517/1U2Libro EAPAliaga .pdf

DelMas, R., Garfield, J., Ooms, A., & Chance, B. (2007). Assessing Students` Conceptual Undersatanding after a First Course in Statistics. *Statistical Educational Research Journal*,6 (2), 28-58

Garfield, J., & Ben-Zvi, D. (2010). *Developing Students´Statistical Reasoning. Connecting Research and Teaching Practice.* Springer.

Garfield, J., DelMas, R., & Zieffler, A. (2012). *Developing statistical modelers and thinkers in an introductory, tertiary-level statistics course.* EEUU: Springer.

Kerlinger, F. y H. Lee. (2002), *Investigación del comportamiento.* Mc-Graw-Hill. 4ta Edición.

Nunnally, J. (1987) *Teoría psicométrica.* Editorial Trillas. 1ra Edición.

Ruiz Bolívar, C. (2002). *Instrumentos de Investigación Educativa.* Venezuela: Fedupel.

Tishkovskaya, S., & Lancaster, G. A. (2012). *Statistical Education in the 21st Century: a Review of Challenges, Teaching Innovations and Strategies for Reform.* Obtenido de www.amstat.org/publications/jse/v20n2/tishkovskaya.pdf