Educación y Ciencia
83 - 104
Nueva época Vol. 5 No. 9 (23)
Enero - Junio 2001

Estudio de sesgo y de diferencias entre géneros de los exámenes nacionales de ingreso a la educación media superior y superior del CENEVAL

Agustín Tristán López
Yolanda Leyva Barajas
Rafael Vidal Uribe

Sinopsis

El estudio, realizado sobre los EXANI-I y EXANI-II, presenta la respuesta diferencial entre géneros y los niveles de significancia de las discrepancias entre las respuestas de hombres y mujeres. Se incluye el análisis de estabilidad temporal para los años 1996, 1997 y 1998, para las diferencias de medias entre hombres y mujeres de ambos exámenes.

Este trabajo es complementario al realizado por Tristán, Vidal y Leyva, en donde se demuestra a través del análisis de Rasch que el EXANI-I es una prueba que no produce sesgos que favorecen a alguna parte de la población evaluada y que las diferencias que se observan en función del género son atribuibles de hecho a las diferencias de dominio que tiene cada género y no al instrumento.

Los resultados muestran que no son significativas las diferencias en la dificultad de los reactivos de cada una de las secciones tanto del EXANI-I como del EXANI-II, por lo que las discrepancias encontradas en las medias de respuestas entre los géneros no se pueden atribuir a un sesgo en las pruebas, sino que son un reflejo de los diferentes niveles de dominio que tienen los hombres y las mujeres en ciertas áreas o habilidades que exploran los mismos.

En relación con el desempeño por género, se encontró que las diferencias caen en la categoría de "pequeñas", de acuerdo con la clasificación clásica para este tipo de estudios. En particular, los resultados del EXANI-I indican que hay un mayor dominio de los hombres en las secciones de Habilidad matemática y Geografía, mientras que las mujeres tienen un mayor dominio en Español. En cuanto a los sustentantes del EXANI-II que concursan para incorporarse a alguna institución de educación superior, los hombres obtienen puntuaciones promedio mayores en las secciones de Razonamiento matemático, Mundo contemporáneo y Ciencias sociales, mientras que en el resto de las secciones del examen las diferencias por género no son significativas.

Términos clave: <Investigación> <investigación educacional> <diferencias entre sexos> <exámenes de ingreso> <México>

Abstract

This study about the EXANI-I and EXANI-II exams, presents the differential answer between genders and the level of discrepancy significance between the answers given by men and the answers given by women. It includes the analysis of temporal stability for the years 1996,1997 and 1998, for the mean differences between men and women in both exams.

This project is complementary of a previous one by Tristán, Vidal and Leyva, which demonstrates through the analysis of Rasch that the EXANI-I is a test without a bias that favors one part of the population, and that the differences observed in relation to the gender are ascribed to the differences of command that each gender has rather to the instrument.

The results demonstrate that the differences in the difficulty of the items in each section of the test are not significative, both for the EXANI-I as well as the EXANI-II, hence the discrepancies found in the means of the answers between genders cannot be ascribed to a bias in the tests, but they depict the different levels of command that men and women have in certain areas or skills tested.

Regarding the performance by genders, it was found that the differences can be considered in the category "small" according to the classical categorization for this type of studies. In particular, the results of the EXANI-I indicate that men show greater command in the sections Mathematical skills and Geography, while women perform better in Spanish. About the candidates of EXANI-II to enter higher education institutions, men obtained higher averages in the sections of mathematical reasoning, Contemporary world and Social Sciences, whereas in the rest of the sections there was not found any significative differences between the genders.

Key terms: <Research> <educational research> <sex differences> <entrance examinations> <Mexico>

Introducción

En las últimas décadas se han dado cambios radicales en los roles de las mujeres y los hombres en las sociedades occidentales; tanto en términos de su incorporación al trabajo productivo, como en su participación en la educación superior. Las estadísticas de la matrícula de la educación media superior y superior revelan desde los años 70 un incremento progresivo en la proporción de mujeres.

La importancia de estos cambios ha generado una proliferación de estudios de diferencias entre géneros en especial en lo referente al desempeño académico, principalmente por las implicaciones sociales y políticas que pueden tenerse si llega a probarse una situación injusta que favorezca a uno de los sexos en detrimento del otro, con lo cual puede afectarse la posibilidad de ingresar a un trabajo, a mejores opciones de educación u otros efectos adversos para uno de los sexos (ver por ejemplo los reportes del NAEP de Johnson y Mullis, donde sistemáticamente se incluyen análisis de diferencia de desempeño entre géneros).

Uno de los trabajos pioneros fue el desarrollado por Maccoby y Jacklin (1974), cuyo análisis se sustenta en 1,600 estudios en ocho áreas de logro, personalidad y relaciones sociales, utilizando pruebas de ejecución o rendimiento académico donde participan hombres y mujeres. Entre las conclusiones más relevantes de este análisis, los autores destacan:

- Las mujeres superan a los hombres en el área de habilidad verbal;
- los hombres superan a las mujeres en habilidad espacial-visual;
- los hombres tienen un mejor rendimiento en matemáticas;
- los hombres son más agresivos que las mujeres.

Estas conclusiones discrepan de las de otros investigadores. Por ejemplo, algunos autores han encontrado que no hay diferencias significativas en términos de la habilidad verbal entre hombres y mujeres, mientras que otros investigadores han encontrado que las mujeres superan a los hombres en las habilidades de producción escrita.

Uno de los reportes de investigación más completos y recientes de diferencias de género fue realizado por Nancy Cole en el Educational Testing Service, ETS (1997), quien describe los resultados de 4 años de estudios con datos de más de 400 pruebas diferentes con más de 1,500 conjuntos de datos que involucran millones de estudiantes que integran muestras representativas a nivel nacional en los Estados Unidos. Los resultados de esta autora indican que no hay un cuadro general dominante de un género sobre el otro y que, de hecho, los patrones de diferencias observadas en la ejecución de algunos componentes de habilidades en disciplinas académicas son similares a los intereses y a las actividades fuera de la escuela que tiene cada uno de los sexos, lo cual sugiere que hay una constelación amplia de eventos relacionados con las diferencias observadas. La autora enfatiza que no existe evidencia de que el tipo de formato de los reactivos o los patrones de respuesta (adivinación o velocidad de respuesta) expliquen per se las diferencias observadas.

Los estudios enfocados a analizar la presencia o ausencia de elementos que pudieran incidir en beneficio o perjuicio de la estimación del desempeño entre géneros, grupos sociales, orígenes nacionales, etc., se conocen como 'estudios de sesgo'. Según Cole, un sesgo implica un error sistemático en la medición de un conocimiento o habilidad a favor de un conjunto de individuos de ciertas características comunes. Queda claro que toda medición tiene por objetivo medir afinidades y diferencias, por lo cual la existencia de diferencias entre grupos de personas no implican necesariamente la presencia de un sesgo, a menos que se pruebe que hay 'error sistemático' del instrumento. En este sentido, Cole explica que las diferencias observadas en su estudio no son producto de un error sistemático de medición, sino un reflejo claro de los distintos desempeños que ocurren frente a diversas maneras de medir y con variadas muestras de estudiantes, por lo que concluye que las diferencias obedecen a distintos niveles de ejecución o desempeño

en las áreas disciplinarias y no se deben a los procedimientos de medición.

Para poder concluir que hay diferencias en ciertas habilidades y dominios en función del género, es obligatorio demostrar que el instrumento de medida no presenta sesgos significativos que incidan en un error sistemático en la apreciación de los desempeños individuales o grupales. Si se demuestra en una primera fase que el instrumento reporta fielmente las medidas de dominio, habilidad o desempeño de los sujetos, entonces podrá afirmarse, en una segunda fase, que las diferencias obtenidas son propias de los grupos de sujetos (hombres y mujeres en este caso) y no atribuibles al formato de la prueba, al tipo de reactivos utilizados o a algún otro agente externo al propósito de la medición.

Los métodos de análisis de diferencias de sesgo comúnmente empleados se orientan solo a estudiar las diferencias entre medias de dominio o habilidad obtenidas por grupos de mujeres y hombres. Los autores del presente trabajo consideran que este enfoque es limitado, ya que faltaría un estudio sobre el instrumento propiamente dicho y, por lo tanto, no se atendería la necesidad de demostrar los aspectos relativos al sesgo que el instrumento contiene como error sistemático. Por ello el presente trabajo se orienta a la comprensión tanto de las similitudes como de las diferencias en función del género a través de las dos variables implicadas en la medición: la prueba y la población, que se disgregan en términos de la dificultad de los reactivos y de las puntuaciones obtenidas por los sujetos, respectivamente. El estudio se realizó en las diferentes secciones de los Exámenes Nacionales de Ingreso a la Educación Media Superior y Superior (EXANI-I y EXANI-II) del CENEVAL, que han sido analizado cualitativamente por otros procedimientos (Gago).

En un trabajo previo (Tristán, Vidal y Leyva) se demuestra a través de la metodología conocida como funcionamiento diferencial del ítem (DIF, differential item functioning), utilizando el análisis de Rasch, que el EXANI-I es una prueba que no produce sesgos que pudieran favorecer a una de las partes de la población evaluada, de tal modo que las diferencias que se observan entre hombres y mujeres son atribuibles de hecho a las diferencias de dominio que tiene cada género y no a características del instrumento.

Objetivos del estudio de este trabajo y metodología

El estudio que se presenta aquí tiene por objetivos:

a. Determinar el error de medida respecto al género en las pruebas denominadas EXANI-I y EXANI-II.
b. Identificar la existencia o ausencia de sesgo en el instrumento respecto al género.
c. Mostrar las diferencias de desempeño entre hombres y mujeres en dichas pruebas.

Con estos objetivos se persigue identificar los casos en que las diferencias de dominio o desempeño que se encuentren entre hombres y mujeres son producto de diferencias de habilidades y capacidades propias de los géneros y distinguirlos de los casos en que las diferencias son producto de una prueba sesgada hacia uno u otro sexo.

Para este trabajo se utilizó esta metodología:

a. elección del modelo estadístico para determinar el sesgo en las pruebas (Modelo A);
b. elección del modelo estadístico para determinar las diferencias de desempeño entre géneros (Modelo B);
c. elección de las muestras de sustentantes hombres y mujeres;
d. realización de las pruebas estadísticas con los modelos A y B;
e. conclusiones.

Modelos estadísticos para el estudio de género

Contrariamente a lo utilizado tradicionalmente en estudios de género, donde solamente se utiliza un modelo enfocado al estudio de diferencias entre grupos, para este trabajo se utilizaron dos modelos claramente diferenciados: el primer modelo (denominado Modelo A) está enfocado a determinar el sesgo de los exámenes con respecto al

género; este modelo permite determinar si hay errores sistemáticos del instrumento que produzcan un sesgo significativo entre géneros. El segundo modelo (denominado Modelo B) estudia las diferencias estandarizadas entre medias de dominio de hombres y de mujeres; con este modelo se pueden explicar diferencias de desempeño entre géneros. Es evidente que el Modelo B es aplicable solamente si el Modelo A demuestra que el instrumento es insesgado, en caso contrario las diferencias que pudiera revelar el Modelo B no permitirían explicar diferencias de dominio o desempeño entre hombres y mujeres.

Debe insistirse en la necesidad de disponer de dos modelos diferenciados porque la aplicación de una prueba de dominio o habilidades conduce a una matriz de resultados siguiendo el modelo de Guttman, cuyos renglones están formados por las respuestas de los sujetos y cuyas columnas corresponden a las respuestas ante los reactivos (Figura 1).

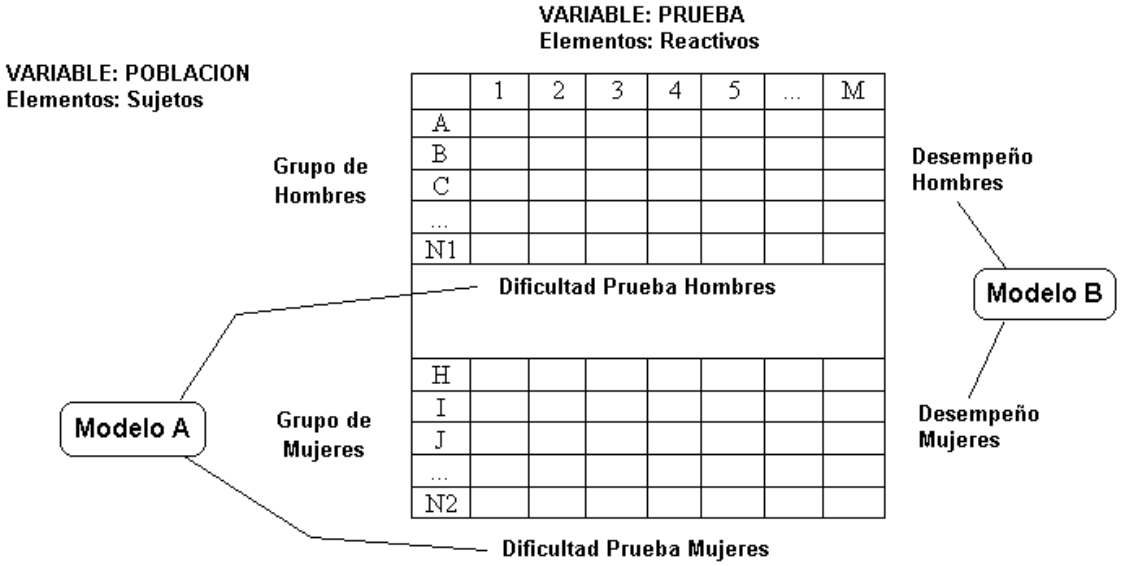

Figura 1.

En este contexto se establecen dos variables: La primera es la variable Prueba cuyos elementos son los reactivos y cuyas medidas son las dificultades de los reactivos y de la prueba. La segunda variable es la Población cuyos elementos son los sujetos o personas que contestan la prueba y cuyas medidas son los desempeños de los sujetos. Las dificultades de los reactivos se obtienen en los marginales inferiores de las columnas, mientras que los desempeños de los sujetos se determinan en los marginales derechos de los renglones. Obsérvese que los análisis tradicionales se enfocan, erróneamente, a determinar el posible sesgo del instrumento utilizando los desempeños de los sustentantes; el error estriba en que se utiliza la variable Población cuando en realidad se está tratando de dictaminar el sesgo de la variable Prueba.

Los estudios con ambos modelos se realizaron para las diferentes secciones del EXANI-I y del EXANI-II. En el caso del Modelo A se trabajó con datos obtenidos de las aplicaciones de 1999, mientras que el Modelo B se utilizó con datos provenientes de aplicaciones de 1996, 1997 y 1998. En este caso el estudio permitió hacer, adicionalmente, un análisis preliminar de estabilidad temporal de las secciones del CENEVAL.

El EXANI-I consta de 128 reactivos, el EXANI-II contiene 120 reactivos, distribuidos en secciones, como se indica a continuación:

EXANI-I	EXANI-II
Habilidad verbal	Razonamiento verbal
Español	Razonamiento numérico
Historia	Mundo contemporáneo
Geografía	Ciencias naturales
Civismo	Ciencias sociales
Habilidad matemática	Matemáticas
Matemáticas	Español
Física	
Química	
Biología	

Para el análisis de las pruebas se utilizaron todas las secciones y todos los reactivos que las forman. La población sobre la cual se han aplicado los EXANI es muy grande, razón por la cual se hizo un muestreo aleatorio, estratificado a las proporciones de hombres y mujeres de la población total (Tabla 1).

Tabla 1.

Tipo de examen	Género	1996	1997	1998
EXANI-I nacional	Femenino	14,853	13,885	7,553
	Masculino	14,261	15,042	7, 301
EXANI-I	Femenino		3,818	22,225
Metropolitano	Masculino		3,572	22,775
EXANI-II Nacional	Femenino		15,674	13,592
	Masculino		13,998	15,347
Totales	Femenino	14,853	33,377	43,370
	Masculino	14,261	32,612	45,423

(*) El EXANI-I Metropolitano de la Ciudad de México incluye al Distrito Federal y a los municipios conurbados del Estado de México.

Modelo A. Estudio de sesgo del examen

El problema que se plantea es el siguiente: Supóngase que se aplica una misma prueba o examen a un grupo de hombres y a un grupo de mujeres y se desea determinar si existe sesgo en el instrumento de medida. Considérese que la prueba aplicada a hombres es la prueba 1 y que la prueba aplicada a mujeres es la prueba 2, si ambas pruebas proporcionan dificultades similares tanto en los reactivos, como en las secciones y en el global, entonces puede afirmarse, sin lugar a dudas, que las pruebas 1 y 2 no difieren. Atendiendo a que la prueba aplicada en ambos casos es la misma, puede afirmarse categóricamente que la prueba no está sesgada.

El razonamiento descrito se traduce en la aplicación de una prueba de hipótesis de diferencia de proporciones para el examen aplicado a los dos grupos de personas (hombres y mujeres), como si se tratara de dos instrumentos diferentes. El estadístico requerido para esta prueba está dado por la expresión *ZH,M*:

$$ZH,M = \frac{Ph - Pm}{1.65\sqrt{pq(\frac{1}{N1} + \frac{1}{N2})}}$$

$$\text{con } p = \frac{N1Ph + N2Pm}{N1 + N2} \text{ , } q=1-p$$

donde *Ph* es el índice de dificultad media del examen de *N1* reactivos aplicados a hombres, *Pm* es el índice de dificultad media del examen de *N2* reactivos aplicados a mujeres. Tanto *Ph* como *Pm* pueden obtenerse a partir del Grado de Dificultad clásico normalizado a 1 (dividiendo GD entre 100). El coeficiente 1.65 del denominador hace referencia al valor de la variable estandarizada Z al 5% de significancia en una prueba de 1 cola (interesa identificar la existencia de sesgo hacia cualquiera de ambos géneros); este valor es más exigente que el usado por otros autores que hacen análisis de dos colas al 5% con un valor de rechazo de 1.96 (Johnson).

Para que el estadístico *ZH,M* sea significativo con un 95% de confianza, su valor absoluto debe ser igual o mayor a 1, en caso contrario el sesgo carece de significación. Si el estadístico proporciona valores inferiores a 1 entonces puede afirmarse que la prueba no induce un sesgo a favor de uno de los grupos (hombres o mujeres en este caso). Si el signo de *ZH,M* es positivo se tiene una mayor influencia hacia los hombres y si es negativo la influencia es mayor hacia las mujeres.

Tabla 2

Valor de *ZH,M*	Interpretación del sesgo al 95% de confianza	Interpretación respecto al examen
$\lvert ZH,M \rvert < 1.0$	No significativo	No produce error sistemático a favor de uno u otro sexo
$\lvert ZH,M \rvert > 1.0$	Significativo	Produce error sistemático a favor de uno u otro sexo

El estadístico *ZH,M* permite estudiar al instrumento, con la ventaja de que no se ve afectado por el tamaño de la población de sustentantes que responden la prueba. Obsérvese que las dimensiones de las pruebas son finitas (dadas por el número de reactivos), pero la variable que se está midiendo (habilidad, conocimientos) es de dimensión infinita; puede decirse entonces que se trata de un muestreo finito y, por lo tanto, la prueba de diferencia de proporciones es representativa de la variable en estudio. La expresión propuesta para *ZH,M* es general, pero puede simplificarse si se toma en cuenta que se trata de una misma prueba aplicada tanto a hombres como a mujeres, por lo que *N1* es igual a *N2*. La simplificación es trivial y, por lo tanto, no se presenta aquí.

Modelo B. Estudio de diferencia de dominio entre géneros

Una vez probado que el instrumento no tiene un sesgo que induce un error sistemático a favor o en contra de uno de los grupos en estudio, se puede

proceder al análisis de las diferencias de desempeño o dominio entre los géneros.

La prueba estadística utilizada aquí es la de diferencia estándar entre medias (Cohen, Cole), por medio del índice estadístico *d*, calculado con la fórmula:

$$d = (\mu h - \mu m) / \sigma$$

donde μh es la media de puntuaciones en hombres, μm es la media de puntuaciones en mujeres y σ es la desviación estándar para toda la población.

El índice *d* es un número independiente de la unidad de medida o de la escala original utilizada para calificar a los sujetos, con lo cual puede interpretarse, por una parte, como el grado de desviación de la hipótesis nula (μh=μm) y, por otra parte, como la magnitud del fenómeno que se requiere detectar. Atendiendo a que el índice *d* se refiere a la diferencia estándar de dos distribuciones que comparten una misma desviación estándar, puede interpretarse convenientemente en función del área no común entre ambas distribuciones, o área en que no se traslapan. Esta área, denominada U1, representa la probabilidad de que difieran ambas distribuciones: un valor U1 nulo implica la igualdad de las distribuciones.

La interpretación de *d* se hace como sigue: Si no existen diferencias entre géneros, *d* será igual a cero. Un valor positivo de *d* indicará un dominio mayor en los hombres, mientras que un valor negativo corresponderá a un dominio mayor en las mujeres. Las diferencias deben ser suficientemente grandes para ser significativas, las cuales se interpretan (Cole, Cohen) de acuerdo con la siguiente convención:

Tabla 3.

Valor absoluto de *d*	Valor de U1	Interpretación de la diferencia	La diferencia es
0.0 < | d | < 0.2	Menor de 14.7 %	Muy pequeña	No significativa
0.2 < | d | < 0.5	Menor de 33.0 %	Pequeña	A revisar, puede tomarse en cuenta
0.5 < | d | < 0.8	Menor de 47.4 %	Mediana	Significativa
0.8 < | d |	Mayor de 47.5 %	Grande	Significativa

De acuerdo con lo presentado hasta aquí, solamente serán de interés los valores superiores a 0.5 que entran en las categorías de diferencias 'medianas' o 'grandes'. Con objeto de hacer un análisis más fino, en este estudio también se tomaron en cuenta los casos de valores del índice *d* en el intervalo de 0.2 a 0.5 o de diferencias 'pequeñas', que se interpretan como categoría 'a revisar'.

Se adoptó el uso del índice *d* porque tiene la ventaja de ser un estadístico estandarizado, empleado en estudios de diferencias de género, para el cual se cuenta con los valores de interpretación sugeridos en las referencias, dados en la Tabla 3. Este índice puede interpretarse como un caso particular de una prueba de diferencia de medias de dos poblaciones.

Estudio simplificado de estabilidad temporal

Cuando se dispone de datos de varios años o de diversas aplicaciones, se puede repetir el cálculo de cualquiera de los estadísticos en estudio con objeto de determinar si los resultados presentan una estabilidad en el tiempo.

El análisis de estabilidad temporal se divide en dos partes. En primer lugar deben calcularse los valores del estadístico en estudio para identificar si son significativos o no. En segundo lugar, sólo cuando los valores son significativos, deberá determinarse si las fluctuaciones son sistemáticas para dictaminar si son estables. Cuando los valores no son significativos es claro que las posibles fluctuaciones que se encuentren en el estadístico carecen de interpretación; en este caso puede decirse que las fluctuaciones son explicables por

el error inherente del instrumento. Un estudio completo de estabilidad temporal requiere la comparación detallada del error de medida contra la significancia del índice *d*, lo cual es motivo de otro trabajo (Tristán, Leyva, Vidal).

Para este trabajo se hizo un estudio simplificado de estabilidad temporal comparando el índice *d* para cada una de las secciones de los EXANI (sin incluir la comparación contra el error de medida de cada una de dichas secciones), de manera de tener una descripción de la forma en que ha evolucionado el índice *d* en los años disponibles. El énfasis se tiene exclusivamente en la identificación de los niveles de significancia en la diferencia entre géneros, bajo el siguiente criterio descriptivo:

[1] Para que se considere la posibilidad de dictaminar una diferencia sistemática se deben cumplir dos condiciones: (a) el signo del índice d debe ser el mismo en todos los años en estudio, (b) debe haber por lo menos un año con valores superiores a 0.2 (para el dictamen 'pequeño') o de 0.5 (para el dictamen 'mediano'). Los casos que cumplen esta condición se denominarán 'potencialmente interesantes'.

[2] Si se cumplen las condiciones del criterio [1], se dirá que la diferencia es sistemática a favor de uno u otro grupo (hombres o mujeres en este caso) si los valores de los diferentes años están en relación no menor al 50% del valor máximo en la sección en estudio. Los casos que cumplen esta condición se denominarán 'de diferencia significativa' al valor del índice *d* elegido.

[3] Los valores absolutos del índice *d* en el rango de 0 a 0.2 (diferencia 'muy pequeña' o 'no significativa') en general se desprecian para el dictamen, independientemente del signo que tengan, salvo que se entre en las condiciones dadas en [1]. Los casos que incurren en esta condición se denominarán 'de diferencia no significativa'.

Resultados

Modelo A. Estudio de sesgo del examen

EXANI-I

Se determinaron los valores de *ZH,M* para cada una de las secciones y la calificación global del EXANI-I tomando una muestra al azar de sustentantes de las aplicaciones de 1999. Para estudiar si existe sesgo o no en el examen, se utilizaron las respuestas tanto de aplicaciones nacionales como del concurso de la zona metropolitana de la ciudad de México.

Todos los valores absolutos de *ZH,M* son muy inferiores a 1 (figura2), por lo que se puede concluir que el EXANI-I a nivel global y en sus secciones no tiene sesgo significativo con relación a los géneros. El caso más desfavorable se obtuvo en la sección de Habilidad matemática con un valor de 0.15, lo cual lo coloca en el 15% del valor de significancia o a un 85% por abajo del valor de 1. Estos datos se presentan en la tabla 1, en donde se puede constatar la inexistencia de sesgo significativo en la forma de respuesta de los sustentantes en todas las secciones del instrumento. Recuérdese que el signo positivo indica 'a favor' de los hombres mientras que el signo negativo indica 'a favor' de las mujeres. Dado que los valores no son significativos, las diferencias observadas pueden ser atribuidas al azar.

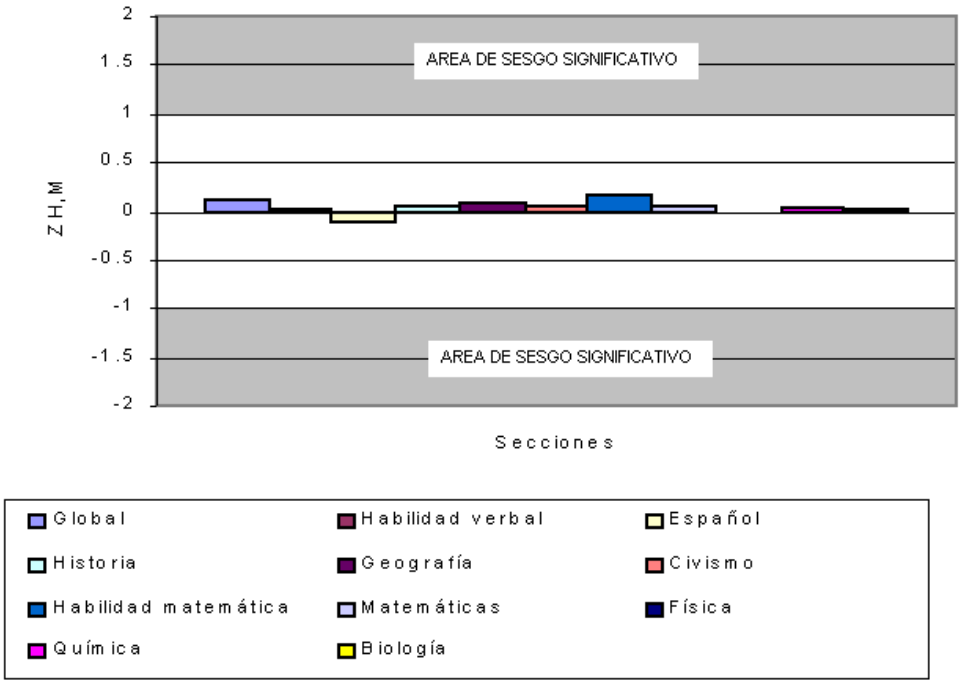

Figura 2

Tabla 4.
Estudio de sesgo del EXANI-I

Seccion	Núm. de reactivos	Hombres Ph	Mujeres Pm	*ZH,M*
Global	128	0.487	0.475	0.115
Habilidad verbal	24	0.532	0.530	0.011
Español	10	0.432	0.472	-0.109
Historia	10	0.467	0.446	0.058
Geografía	10	0.542	0.512	0.081
Civismo	10	0.539	0.521	0.048
Habilidad matemática	24	0.498	0.462	0.151
Matemáticas	10	0.382	0.366	0.043
Física	10	0.452	0.452	-0.001
Química	10	0.478	0.469	0.024
Biología	10	0.472	0.466	0.017

EXANI II

Se determinaron los valores de *ZH,M* para cada una de las secciones y el global del EXANI-II tomando a una muestra de sustentantes de las aplicaciones de 1999.

Todos los valores de *ZH,M* son notablemente inferiores a 1, por lo tanto se puede afirmar que los exámenes no tienen sesgos o no son significativos con relación a los géneros. El caso más desfavorable se obtuvo en la sección de Mundo Contemporáneo con un valor de -0.6036, lo cual lo coloca a un 40% por abajo del valor de significancia de 1. Los resultados se presentan en la tabla 2 y se representan en la figura 3, en donde puede apreciarse en forma gráfica la inexistencia de sesgo en la forma de respuesta de los sustentantes en todas las secciones del EXANI II.

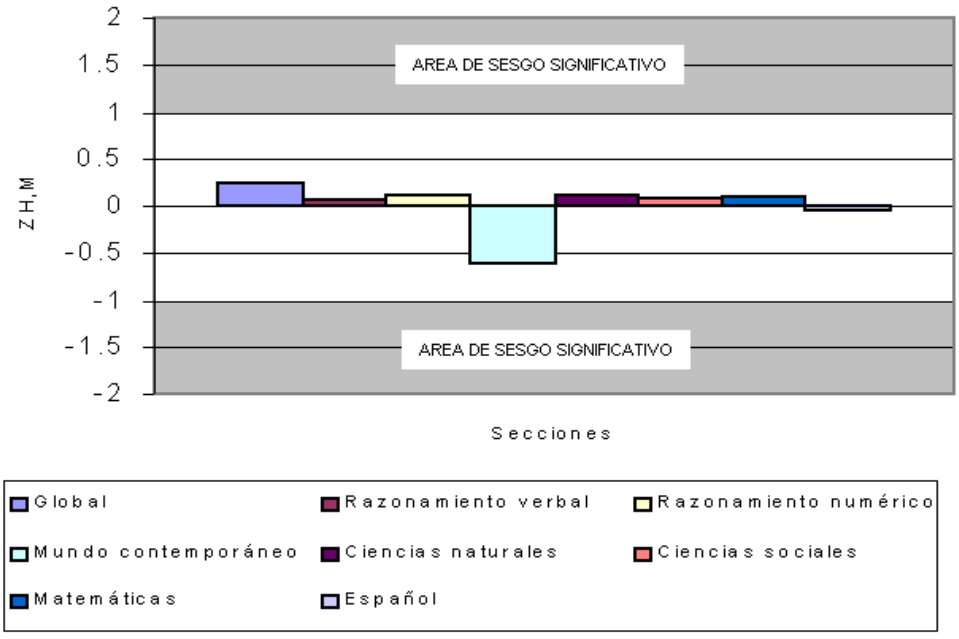

Figura 3

Tabla 5.
Estudio de sesgo del EXANI-II

Seccion	Núm. de reactivos	Hombres Ph	Mujeres Pm	*ZH,M*
Global	114	0.378	0.353	0.234
Razonamiento verbal	20	0.428	0.411	0.066
Razonamiento matemático	18	0.538	0.506	0.118
Mundo contemporáneo	16	0.332	0.506	-0.604
Ciencias naturales	15	0.301	0.267	0.128
Ciencias sociales	16	0.332	0.309	0.084
Matemáticas	14	0.309	0.282	0.095
Español	15	0.358	0.370	-0.041

Modelo B. Estudio de dominio por género

Una vez comprobado que el instrumento no tiene sesgo con relación al género de los sustentantes, se procedió a estudiar las diferencias de dominio en cada una de las secciones, utilizando el índice *d*. Aprovechando las bases de datos disponibles de varios años se hicieron los cálculos del índice *d* y se estudió por el esquema simplificado la estabilidad temporal de cada una de las secciones de los EXANI.

EXANI-I Nacional y Metropolitano

Como resultado de los cálculos realizados con el índice *d* se tiene que, en ninguno de los casos, hay diferencias medianas o grandes para *d*. En las tablas 6 y 7 se presentan los datos de los años 1996, 1997 y 1998 para el EXANI-I Nacional; y de 1997 y 1998 para el examen Metropolitano. Obsérvese que el índice *d* está en el rango de diferencias 'muy pequeñas' o 'pequeñas' (figuras 4 y 5); en particular se aprecia que la sección de Español es la que sistemáticamente muestra las diferencias más marcadas, con un dominio promedio superior de las mujeres, con valores absolutos de 0.196 a 0.225. También se puede apreciar que Geografía se ha mantenido en los tres años en los primeros y segundos lugares con valores de 0.205 a 0.247 es decir que hay un dominio promedio mayor en los hombres. Algunas otras secciones presentan sistemáticamente valores muy bajos de *d*. Tal es el caso de Matemáticas, Química y Biología, por ejemplo.

Por otra parte se tienen secciones con dominio promedio superior en los hombres pero que presentan fluctuaciones en cuanto a su ubicación en función del año, a este respecto aparecen Habilidad matemática e Historia. Un caso interesante de fluctuaciones más marcadas se tiene con Física tanto en las muestras nacionales como en las de la zona metropolitana del D. F., que en 1996, se encontraba en primer lugar, mientras que ocupa el segundo lugar en 1997 para la muestra nacional, el cuarto lugar en el mismo año para la zona metropolitana y los últimos sitios en 1998.

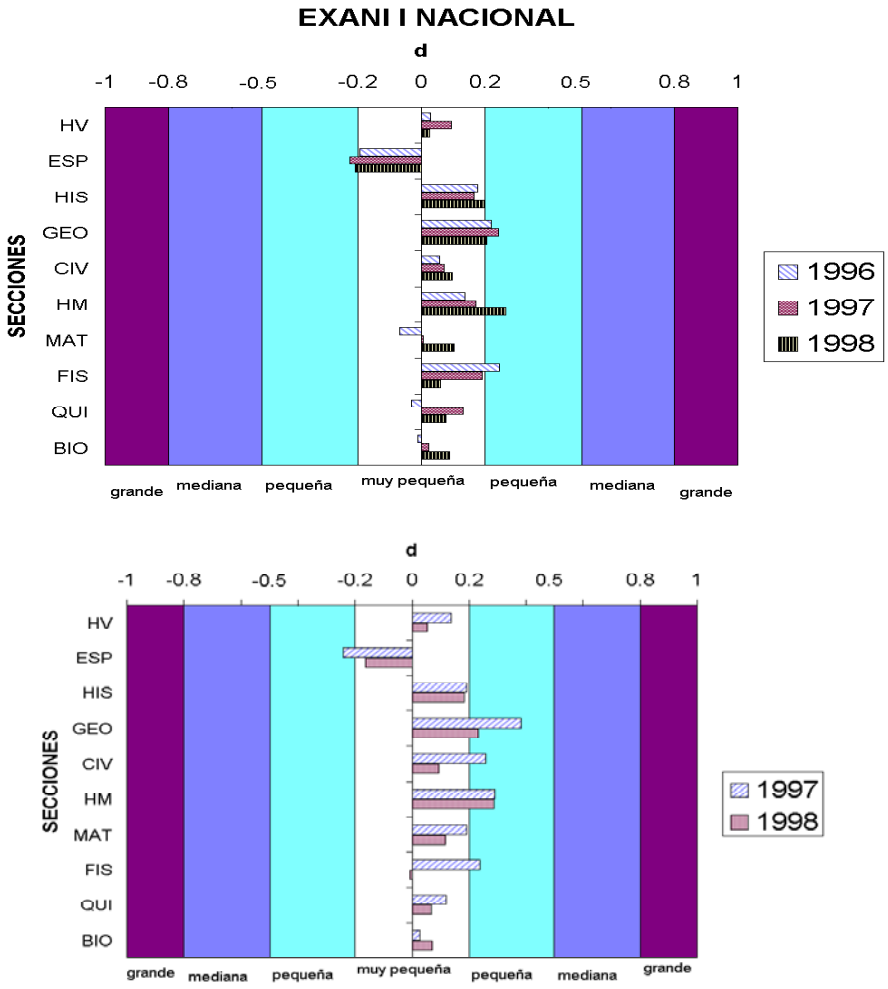
EXANI I NACIONAL
d
-1
-0.8
-0.5
-0.2
0
0.2
0.5
0.8
1
SECCIONES
HV
ESP
HIS
GEO
CIV
HM
MAT
FIS
QUI
BIO
grande
mediana
pequeña
muy pequeña
pequeña
mediana
grande
1996
1997
1998
d
-1
-0.8
-0.5
-0.2
0
0.2
0.5
0.8
1
SECCIONES
HV
ESP
HIS
GEO
CIV
HM
MAT
FIS
QUI
BIO
grande
mediana
pequeña
muy pequeña
pequeña
mediana
grande
1997
1998

Figura 5

Código	Sección	Código	Sección
HV	Habilidad Verbal	HM	Habilidad Matemática
ESP	Español	FIS	Física
HIS	Historia	QUI	Química
GEO	Geografía	BIO	Biología

Tabla 6. EXANI I NACIONAL

	1996				1997				1998			
	FEME	MASC			FEME	MAS			FEME	MAS		
	13885	15042			14853	14261			7301	7553		
SECCIÓ	x	x	σ	*d*	x	x	σ	*d*	x	x	σ	*d*
HV	50.47	52.07	16.88	.095	47.95	48.50	18.90	.029	52.69	53.13	16.60	.027
	52.31	47.31	22.24	-.225	46.57	42.80	19.32	-.196	52.48	47.91	21.84	-.209
HIS	46.05	49.74	20.31	.182	41.21	44.57	18.19	.185	45.11	49.18	20.16	.202
GEO	44.46	49.35	19.81	.247	42.94	47.62	20.92	.224	45.57	49.66	19.94	.205
CIV	47.29	48.70	21.78	.065	47.25	46.13	19.12	-.059	48.07	50.06	20.26	.098
HM	47.71	51.17	19.78	.175	41.39	44.55	22.26	.142	48.40	53.68	19.83	.266
MAT	42.42	42.55	22.22	.006	39.40	40.84	20.90	.069	41.66	43.92	21.95	.103
FIS	41.57	45.80	21.95	.193	35.17	40.26	20.38	.250	42.38	43.65	21.13	.060
QUI	46.83	49.61	20.85	.133	47.27	46.63	19.48	-.033	48.99	50.59	20.99	.076
BIO	49.32	49.82	21.59	.023	42.68	42.92	20.17	.012	49.15	50.83	20.39	.088

Tabla 7. EXANI I METROPOLITANO

	1997				1998			
	FEME	MASC			FEME	MASC		
N	3818	3572			22225	22775		
SECCIÓN	x	x	σ	*d*	x	x	σ	*d*
HV	48,73	50.90	15.74	.138	55.53	56.52	15.02	.066
ESP	55,65	50.22	22.60	-.240	55.30	51.74	21.93	-.162
HIS	52,31	56.34	21.17	.190	44.13	47.78	20.08	.182
GEO	46,22	54.22	21.52	.372	49.92	54.65	20.78	.228
CIV	48,65	54.22	21.38	.261	52.52	54.31	19.38	.092
HM	53,03	58.72	19.90	.286	50.92	56.88	21.05	.283
MAT	50,01	54.22	22.68	.186	43.11	45.43	20.04	.116
FIS	47,29	52.52	22.26	.235	44.53	44.38	20.68	-.007
QUI	53,13	55.59	20.76	.118	53.82	55.27	21.45	.068
BIO	57,96	58.53	21.86	.026	52.27	53.68	20.27	.070

A partir de los resultados obtenidos se aplican los criterios para el análisis simplificado de estabilidad temporal sobre el índice *d*. En todos los casos se obtuvieron valores por debajo de 0.5, lo cual indica que las diferencias máximas sólo podrían calificarse como 'pequeñas'. Siendo valores reducidos puede afirmarse, de entrada, que los resultados obtenidos para *d* no muestran, en general, valores significativos, de tal modo que no es factible decidir de manera absoluta que hay una diferencia de dominio 'mediana' o 'grande' entre géneros.

Se clasificaron las secciones 'potencialmente interesantes', para ello se utilizó a 0.2 como valor de referencia. Este es el caso de las secciones de Español, Historia, Geografía, Habilidad matemática y Física en el caso del EXANI-I Nacional. Para el examen Metropolitano las secciones 'potencialmente interesantes' son: Español, Geografía, Civismo y Habilidad matemática. En este último examen se tiene una incompatibilidad de criterio en la sección de Física, ya que presenta un cambio de signo en el valor de *d*. Conviene observar aquí que la sección de Historia está en una situación particular, porque el máximo obtenido es 0.202 a nivel nacional, mientras que en los otros años y en la aplicación metropolitana los valores son inferiores a 0.2, de tal modo que pudiera catalogarse a Historia como una sección límite que haría considerarla como 'no significativa'. De cualquier modo se conserva para el dictamen posterior.

A continuación se identifican los casos de las secciones 'potencialmente interesantes' que pueden clasificarse como 'de diferencia significativa'. En esta última clasificación permanecen en el EXANI-I Nacional la sección de Español (que muestra un mayor dominio femenino) y las secciones de Historia, Geografía y Habilidad matemática (con mayor

ISSN 0188 3364

dominio masculino). En cuanto al examen Metropolitano solo se conservan la sección de Español (con dominio femenino) y las secciones de Geografía y Habilidad matemática (dominantemente masculinas).

En conclusión, se puede establecer una razonable conjetura de mayor dominio femenino en el caso de Español y mayor dominio masculino en Habilidad Matemática y Geografía. El caso de Historia hace suponer un ligero dominio masculino, pero dado que los valores se encuentran en el intervalo de 'pequeños', el dictamen no es concluyente. Desde luego, las diferencias deberán corroborarse en futuras aplicaciones del EXANI-I.

Si se analizan los máximos y las tendencias, es notable que solo Geografía presenta valores consistentemente superiores a 0.2 tanto en la aplicación nacional como en el examen Metropolitano. En segundo lugar se tiene que mencionar que la sección de Español muestra de manera consistente un mayor dominio femenino. En menor medida se puede hablar de un dominio masculino en Habilidad matemática y en Historia.

Los casos de las secciones de Física y Civismo muestran un comportamiento poco estable en el tiempo, lo cual permite sugerir la realización de estudios diferenciados en estas secciones con revisiones de los contenidos evaluados para poder explicar las diferencias observadas y que no pueden ser explicadas por cuestiones de sesgo inherente al instrumento.

Junto con estas tendencias, resulta interesante observar que si bien Habilidad matemática es de predominio masculino, la sección de Matemáticas, junto con las de Química y Biología, no reportan diferencias entre hombres y mujeres, con cambios de signo en los valores de *d* en los diferentes años. Siendo valores pequeños en todos los casos, es posible afirmar de inmediato que el dominio en estos temas no presenta predominio masculino o femenino.

Por otra parte, es interesante el hecho de que no existan diferencias significativas en la sección de Habilidad Verbal, mientras que en estudios realizados principalmente en Estados Unidos, las mujeres presentan ejecuciones superiores a las de los hombres en esta área.

EXANI-II Nacional

Habiéndose demostrado que el instrumento no presenta diferencias de sesgo, las diferencias observadas entre las medias de desempeño pueden ser atribuidas a diferencias de dominio entre los géneros. En el caso de EXANI-II en todas las secciones se obtuvieron valores en el rango de 'muy pequeñas' o 'pequeñas', en general inferiores a 0.5. Estas diferencias caen en un rango de no significancia o 'a revisar'.

Existe un orden bastante estable en cuanto a las diferencias encontradas en los dos años de estudio (tablas 8). En particular, la sección de Mundo contemporáneo es la que muestra mayor dominio masculino, seguida de Ciencias sociales y Razonamiento numérico. En contraposición a los datos observados en el EXANI-I en donde se encuentra que la sección de Español muestra sistemáticamente un mayor dominio promedio por parte de las mujeres, en el EXANI-II los valores de *d* para los dos años, aunque igualmente favorables a las mujeres, caen en el rango de no significativos.

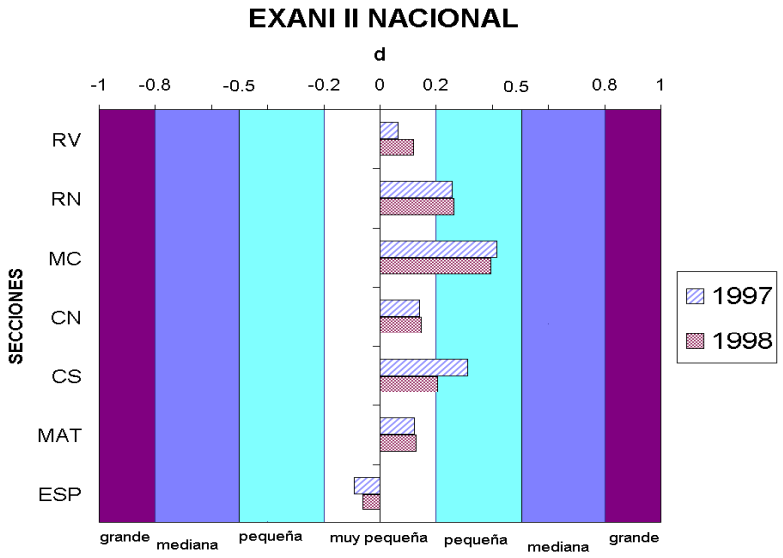

Figura 6.

Tabla 8. EXANI II NACIONAL

	1997				1998			
	FEME	MASC			FEME	MASC		
N	15674	13998			15347	13592		
SECCIÓN	x	x	σ	*d*	x	x	σ	*d*
RV	44.52	45.50	15.00	.065	41.52	43.29	14.66	.121
RM	34.08	38.66	17.03	.269	34.72	38.81	16.48	.248
MC	41.05	47.18	14.68	.417	41.53	47.46	14.94	.397
CN	33.09	35.13	14.17	.144	31.29	33.33	13.59	.150
CS	30.79	35.31	13.98	.323	30.88	33.71	13.28	.213
MAT	32.12	34.44	18.13	.126	30.25	32.64	17.58	.136
ESP	34.14	32.56	17.18	-.116	32.90	31.86	16.87	-.062

Respecto a la estabilidad temporal de las secciones del EXANI-II, se pueden clasificar como 'potencialmente interesantes' las de Mundo contemporáneo, Razonamiento numérico y Ciencias sociales. Estos tres casos se pueden clasificar a su vez como secciones 'de diferencias significativas' al nivel de 0.2, que es un nivel de significancia 'pequeño' de mayor dominio masculino. Las demás secciones no presentan diferencias significativas en los valores del índice *d*.

En este caso se puede hacer el mismo comentario que el que se presentó para EXANI-I, con relación a la pequeña diferencia dada por *d*. Tratándose de valores no significativos la posición relativa de las secciones puede ser diferente para otras aplicaciones del EXANI-II. Pudiera esperarse, sin embargo, un comportamiento razonablemente similar solamente para Mundo Contemporáneo, esto deberá comprobarse en futuras aplicaciones del examen.

Comparaciones entre los exámenes

No se pretende hacer la comparación de resultados entre EXANI-I y EXANI-II, ya que se trata de exámenes de características diferentes no solamente en su diseño, dimensiones y especificaciones, sino principalmente por las poblaciones objetivo a las cuales están dirigidos. Las secciones de estos instrumentos pudieran ser clasificadas como sigue:

Tabla 9.

EXANI-I	EXANI-II
Habilidad verbal	Razonamiento verbal
Habilidad matemática	Razonamiento matemático
Español	Español
Matemáticas	Matemáticas
Física	Ciencias naturales
Química	
Biología	
Historia	Ciencias sociales y Humanidades
Geografía	
Civismo	
	Mundo contemporáneo

La clasificación de las secciones es arbitraria en este momento, solo para fines de este estudio, tomando en cuenta que no se tienen reactivos de anclaje que permitan una comparación más estricta u orientada a propósitos específicos de contraste entre los dominios que se miden con cada uno de los instrumentos. La sección de Mundo contemporáneo no se consideró para la comparación entre los exámenes, por no disponerse de una sección similar en el EXANI-I.

A partir de esta organización de las pruebas, se prepararon los datos de la Tabla 10, donde se tienen los valores consolidados de las medias de desempeño para hombres y mujeres y las desviaciones estándar en cada caso. El índice *d* se calculó considerando el valor medio de las desviaciones estándar de ambos exámenes, esto es obligatorio para poder hacer comparables las distribuciones de las diferentes poblaciones.

Tabla 10.

	EXANI-II				EXANI-I			
	FEME	MASC	σ	*d*	FEME	MASC	σ	*d*
Habilidad verbal	43.04	44.41	14.83	0.085	49.88	50.91	17.63	0.063
Habilidad matemática	34.39	38.73	16.76	0.233	45.95	49.95	20.58	0.214
Español	33.51	32.21	17.02	-0.068	50.64	46.18	21.22	-0.233
Matemáticas	31.20	33.55	17.86	0.119	41.23	42.50	21.72	0.064
Ciencias naturales	32.22	34.25	13.89	0.118	44.88	46.74	20.78	0.107
Ciencias sociales	30.83	34.53	13.64	0.219	45.38	48.42	20.08	0.180

Los valores del índice *d* determinados para las secciones combinadas de EXANI-I y EXANI-II, muestran tendencias similares para ambas pruebas. Las secciones de Habilidad verbal, Matemáticas y Ciencias naturales no presentan diferencias significativas entre géneros. Se puede clasificar en significancia 'pequeña' de manera sistemática a la sección de Habilidad matemática, a favor de los hombres. La sección de Ciencias sociales es ligeramente más alto en EXANI-II que lo que se presenta en EXANI-I. El caso más interesante es el de la sección de Español, con dominio mayor de las mujeres, que es de importancia significativa en EXANI-I y su efecto se atenúa notablemente en EXANI-II (figura 7).

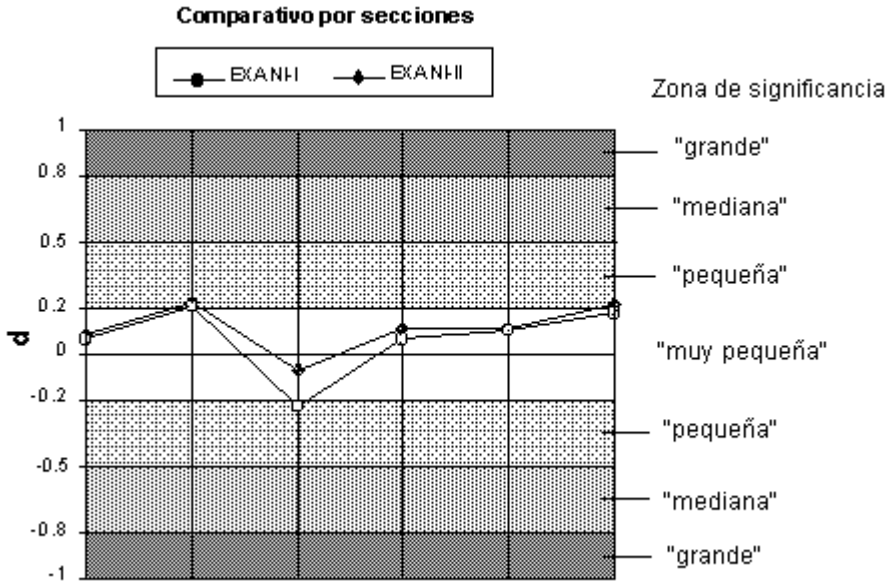

Figura 7

Las diferencias en valor absoluto entre secciones se presentan en la figura 8, pudiéndose observar que la máxima diferencia ocurre en la sección de Español con un valor de 0.16, mientras que las secciones restantes no muestran diferencias superiores a 0.06.

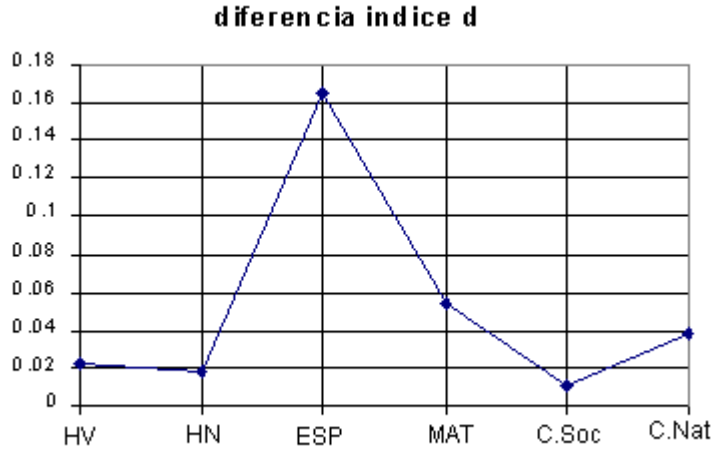

Figura 8

La similitud de comportamiento en el índice *d* en las secciones de los EXANI es un elemento que deberá corroborarse en aplicaciones futuras de ambos exámenes. Puede resultar conveniente incluir reactivos o secciones de anclaje que faciliten las comparaciones para otro tipo de estudios relacionados con el desempeño académico, no solamente de los géneros sino de otras agrupaciones que se desee establecer.

Conclusión

Este estudio se enfocó a analizar las diferencias que se pueden encontrar en las aplicaciones de los exámenes que se elaboran en el CENEVAL, denominados EXANI-I y EXANI-II, de manera de determinar si las diferencias de resultados entre hombres y mujeres, en cada una de las secciones exploradas en los cuestionarios, son un reflejo de las diferencias reales de dominio o producto del sesgo de la prueba. Este hecho es de interés porque diversos grupos sociales (los sustentantes, las instituciones y hasta grupos de observadores) requieren ser evaluados con justicia y en condiciones que no afecten a los resultados, ya que de estas condiciones puede depender alguna decisión institucional o legal.

El trabajo se realizó a través de dos modelos: un primer modelo se enfoca propiamente al estudio del sesgo de la prueba, el segundo modelo tiene por objetivo determinar las diferencias entre géneros; el segundo modelo es aplicable en tanto se haya probado que la prueba es insesgada a favor de uno u otro de los sexos en estudio. Complementariamente se hizo un estudio simplificado de estabilidad temporal de las secciones que forman a los exámenes y una comparación entre las secciones de los exámenes.

Se demuestra en primera instancia que los exámenes no tienen sesgo por género, habiéndose obtenido diferencias muy bajas. En consecuencia puede concluirse que las diferencias observadas en las medias de respuestas de los sustentantes están asociadas a diferencias de dominio y no son atribuibles

a errores sistemáticos producidos por las pruebas; las diferencias de dominio podrían ser explicadas en función de algunas otras variables como por ejemplo los patrones de preferencias diferenciadas por áreas disciplinarias, intereses personales, etc.

Las diferencias encontradas entre los géneros pueden ser clasificadas como 'pequeñas' o 'muy pequeñas', en ninguno de los casos se encontraron diferencias medianas o grandes. Este patrón se verificó con un análisis simplificado de estabilidad temporal y con una comparación entre las secciones de los exámenes.

Dentro de los casos de significancia 'pequeña' se tienen las secciones de Geografía y Habilidad Matemática, en EXANI-I, y las secciones de Mundo contemporáneo, Razonamiento numérico y Ciencias sociales, en EXANI-II, todas ellas con mayor dominio masculino. Posiblemente el patrón más interesante es el de la sección de Español, en EXANI-I, que sistemáticamente resulta con valores de dominio promedio más altos a favor de las mujeres, patrón que se atenúa de manera importante en los resultados del EXANI-II.

Dado que las diferencias se encuentran en los rangos pequeños o de no significación, los patrones hallados aquí pudieran cambiar entre las diversas aplicaciones de los cuestionarios. Si las especificaciones de los exámenes analizados aquí permanecen constantes se puede considerar que las fluctuaciones esperadas no serán importantes desde el punto de vista estadístico.

Referencias

Cohen, J. (1988). Statistical Power Analysis for the Behavioral Sciences. Second Edition. LEA. Hillsdale, N. J.

Cole, N. (1997). THE ETS GENDER STUDY: How Females and Males Perform in Educational Settings. Educational Testing Service. Princeton, N. J.

Draba, R. E. (1977). The identification and interpretation of item bias. Education Statistics Laboratory Memorandum 25, Mars. University of Chicago. 11 PP

Gago, A.(1999). Diferencias de género en el EXANI-I. 3a Reunión Nacional de Instituciones Usuarias del EXANI-I. 24-26 Nov. 1999, Mazatlán, México

Guttman, L. (1950). The basis of scalogram analysis. En Stouffer y col. Measurement and prediction, New York. John Wiley.

Johnson E.G. (1994). Estimation error variance by gender and race/ethnicity. The NAEP 1992 technical report. ETS, NCME, ERI, Julio 1994, REPORT N.23-TR20, PP 967 Y SIGS

Mullis I.V.S. Y COL. (1993). Overall mathematics achievement for demographic groups for the nation and the states. The NAEP 1992 mathematics report. ETS, NCME, ERI, April 1993, REPORT N.23-ST02

Tristán, A., Vidal, R. y Leyva, Y. (1999). La calificación de Juan es 900, la de María es 1000. ¿Hay diferencia en su dominio o el examen está mal hecho? Un estudio de género**.** Informe interno del CENEVAL, Dirección Técnica (en publicación).

Tristán A., Leyva, Y. y Vidal, R. (1999). Modelo para el análisis de la estabilidad temporal de un examen. Aplicación al EXANI-I y al EXANI-II. Informe Interno del Ceneval. Dirección Técnica (pendiente de publicación).