Avances en Medición, 4, 113–128
2006

ESTUDIO DE SESGO Y DE DIFERENCIAS ENTRE GÉNEROS DEL EXAMEN DE INGRESO A LA UNIVERSIDAD AUTÓNOMA METROPOLITANA

Agustín Tristán–López[1]
Instituto de Evaluación e Ingeniería Avanzada, México

Liliana Mendoza–González & Gerardo Gutiérrez–Santiago
Universidad Autónoma Metropolitana, México

Resumen

El objetivo del presente estudio fue analizar el Examen de Ingreso a la Universidad Autónoma Metropolitana (UAM) para identificar si la diferencia entre hombres y mujeres está determinada por la presencia de sesgo a favor de los hombres o por las capacidades individuales de los sustentantes. La muestra estuvo conformada por 36662 sujetos que presentaron el Examen de ingreso a la UAM, el cual consta de seis pruebas que congregan diferentes áreas del conocimiento. Se hicieron tres procedimientos: análisis de sesgo global, análisis de sesgo por reactivo, y estudios de las diferencias debidas al sexo. Los resultados mostraron, que pese a que algunos reactivos presentan sesgo hacia alguno de los géneros, no hay evidencia estadísticamente significativa que señale sesgo global en la prueba. También se observó que el examen resulta difícil para la población y que los hombres tienden a mostrar un mejor desempeño que las mujeres. Al concluir, se descartó la existencia de sesgo en el Examen de Ingreso a la UAM y se determinó que las diferencias en los resultados entre sexo están determinadas por los conocimientos y habilidades de cada sujeto independientemente de su condición de género

Palabras clave: sesgo en las pruebas, DIF

Abstract

This article focuses on the analysis of differential item functioning and sources of bias in the test for selection of students in the Universidad Autónoma Metropolitana. The data from the 36662 candidates was used for this analysis, including three procedures: 1) global test bias, 2) DIF and 3) gender differences. The results show that despite the particular presence of DIF in some cases, there is no significant evidence concerning sources of bias. In general, the test is slighlty difficult for all the students, it shows a trend for men to perform better than women, mostly related to a previous learning and other socio-economic circumstances, rather than a tendency produced by the test.

Key words: test bias, DIF

Introducción

Este estudio tuvo como propósito analizar el Examen de Ingreso a la Universidad Autónoma Metropolitana (UAM, México) para identificar si existe algún tipo de sesgo con relación al sexo de los sustentates y, en su caso, para conocer las diferencias de desempeño entre los hombres y las mujeres con relación a la prueba aplicada. El sesgo se define como un error sistemático en la medición de un conocimiento o una habilidad a favor de un conjunto de

[1] Instituto de Evaluación e Ingeniería Avanzada, S.C.

personas con características comunes, como puede ser el sexo o la etnicidad (Berk,1980; Cole, 1997, 1981; Flaugher,1978). Sin embargo, la presencia de diferencias de desempeño entre grupos de personas no siempre es una prueba de la existencia de un sesgo, debido a que también pueden ser ocasionadas por las diferencias reales en las competencias de las personas o en el desempeño que se pide de acuerdo con un perfil establecido de antemano. Para garantizar que las diferencias observadas en los aspirantes de la UAM reflejan competencias distintas, se estudió si el instrumento presenta o no sesgos significativos y posteriormente se procedió a identificar las diferencias de desempeño entre los sujetos como propiedades intrínsecas del género (Garner y Engelhard, 1999; Gipps y Murphy, 1994, McCornack y McLeod, 1988).

El Examen de Ingreso a la Universidad Autónoma Metropolitana está constituido por seis tipos diferentes de prueba que agrupan diversas áreas del conocimiento y de razonamiento (ver Tabla 1). Estas pruebas contienen, en todos los casos, reactivos o ítems objetivos de opción múltiple de cuatro opciones donde sólo una de ellas es correcta. Los ítems son diseñados por especialistas de la UAM y validados por comités de revisores. En los seis tipos de prueba se evalúa el Razonamiento Verbal (1)[2], Razonamiento Matemático (2), Conocimientos Específicos de cada área (3) y el conjunto se considera como el área global de cada prueba (0).

Tabla 1
Constitución del Examen de Ingreso a la Universidad Autónoma Metropolitana

Tipo	División	Número de ítems			Areas de Conocimiento
		Razonam. Verbal (1)	Razonam. Matemático (2)	Conocim. Específicos (3)	
A - ECAD	Económico-Administrativo	30	30	60	Filosofía / Historia / Lectura, redacción e investigación documental / Matemáticas
B - HUM	Humanidades	30	30	60	Filosofía / Historia / Literatura / Lectura, redacción e investigación documental
C - CSH	Sociales	30	30	60	Filosofía / Historia / Lectura, redacción e investigación documental / Matemáticas
D - CBI	Ciencias Básicas e Ingeniería	15	35	70	Matemáticas / Física / Química
E - CBS	Ciencias Biológicas y de la Salud	30	30	60	Biología / Química / Física / Matemáticas
F - CAD	Ciencias y Artes para el Diseño	30	30	60	Matemáticas / Física / Historia / Dibujo / Diseño y Modelado

[2] En las tablas y gráficas de resultados se emplearán las claves aquí señaladas; Por ejemplo, A-1 indicará la prueba Tipo A - Económico-Administrativo, Area de Razonamiento Verbal (1).

Objetivos del proyecto

1. Determinar si el Examen de Ingreso de la Universidad Autónoma Metropolitana presenta sesgo a favor de uno de los sexos en forma global o diferencial en los ítems, así como su impacto en la medición.

2. Determinar las diferencias de desempeño por género en la prueba citada, en caso de que no haya sesgo o su presencia no sea significativa.

Población analizada

La población de examinados estuvo conformada por los 36662 sujetos que presentaron el Examen de Ingreso a la Universidad Autónoma Metropolitana como aparece en la tabla 2

Tabla 2.
Distribución de la población analizada

Tipo	División	Mujeres	Hombres
A - ECAD	Económico-Administrativo	4348	2986
B - HUM	Humanidades	166	113
C - CSH	Sociales	6765	3468
D - CBI	Ciencias Básicas e Ingeniería	1497	3793
E - CBS	Ciencias Biológicas y de la Salud	5189	2687
F - CAD	Ciencias y Artes para el Diseño	2610	3040

Hipótesis

[H_1]: El Examen de Ingreso a la Universidad Autónoma Metropolitana presenta sesgo a favor de uno de los sexos.

[H_2]: Si el examen presenta defectos de diseño que producen sesgo, entonces las diferencias que pudieran existir en el desempeño entre hombres y mujeres se deben a dichos defectos de diseño y no a diferencias en las competencias reales de los sustentantes.

Método

Para atender los objetivos propuestos, la metodología consistió de estas etapas: (1) Análisis global de la prueba para identificar si el instrumento contiene sesgo por diseño, (2) análisis del funcionamiento diferencial de los ítems para identificar su impacto en la medición del desempeño de los sustentantes y (3) análisis de las diferencias de desempeño entre los sexos. Esta metodología fue seguida anteriormente en estudios similares (Tristán, Leyva y Vidal, 2001a, 2001b) realizados en México con buenos resultados

1. Análisis global de la prueba.

Para el análisis del sesgo global se emplearon dos métodos reconocidos en la literatura: el modelo clásico y el modelo logístico.

a) *Modelo clásico*: Dado que el examen fue aplicado a dos poblaciones distintas por razones de sexo, se estudió el sesgo con la prueba de hipótesis de diferencia de proporciones con el estadístico *ZH,M* como aparece definido en la ecuación 1 para el 5% de significancia de una cola y valor *p* calculado como aparecee en la ecuación 2; donde *Ph* es el índice de dificultad media del examen de *N1* reactivos aplicado a los hombres y *Pm* es el índice de dificultad media del examen de *N2* reactivos aplicados a las mujeres. En este caso particular *N1=N2* ya que se trataba de la misma prueba aplicada todos los examinados.

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

$$p = \frac{N1Ph + N2Pm}{N1 + N2}\ ,\ q = 1 - p \tag{2}$$

b) *Modelo logístico*: Se eligió el modelo de Rasch (1960) porque no requiere de agrupaciones parciales, atendiendo a que su función logística es suficiente para la descripción de los constructos y para definir las medidas de los reactivos (Linacre, 1996; Hambleton, Swaminathan & Rogers, 1991; Wright y Stone, 1988a). Tiene la facilidad de que las variables independientes del modelo (*B*=habilidad del sujeto y *D*=dificultad del ítem) se representan en el mismo eje de dominio, conocimientos o habilidad en el intervalo (–∞, ∞); las unidades en que se expresan *B* y *D* se denominan "lógitos". De acuerdo con el modelo estocástico, la probabilidad de respuesta acertada "*p*" de una persona con medida de habilidad *B* ante un reactivo de dificultad *D* esta dada por (Rasch, 1960; Wright & Stone, 1879, 1998a):

$$p = \frac{e^{(B-D)}}{1 + e^{(B-D)}} \tag{3}$$

Para identificar si los reactivos cumplen satisfactoriamente con las hipótesis de construcción del modelo se estima la bondad del ajuste al modelo (Infit y Outfit) con ayuda del programa Winsteps de Linacre & Wright (1994-2005), el cual proporciona los dos elementos de medida definidos en las ecuaciones 4 y 5; donde *X* es el puntaje observado, *E* es el puntaje esperado, s^2 es la varianza de puntuaciones y *N* es el número de sujetos. Ambas medidas de ajuste se comparan contra el valor *1.0* como valor deseable, estableciéndose un rango de aceptación de *0.8* a *1.2*.

$$OUTFIT = \sum \frac{(X - E)^2}{s^2 N} \tag{4}$$

$$INFIT = \frac{\sum (X - E)^2}{\sum s^2} \tag{5}$$

Para detectar el sesgo se usaron las medidas de la dificultad *D* obtenidas del modelo de Rasch, a partir de un mismo conjunto de reactivos aplicados a los dos grupos establecidos. Si *DH* es la dificultad del ítem para los hombres y *DM* es la dificultad del mismo ítem para las mujeres, entonces la distancia relativa del área entre las curvas del ítem para los dos grupos es $DR = \sum(DH-DM)$. Tal diferencia se consideró significativa cuando la distancia relativa fue superior a *0.10* (Draba, 1977), o sea cuando $|DR| > 0.10$.

Con los valores de medida Rasch de los reactivos para mujeres y para hombres se puede hacer una representación gráfica (Grimby, 1998; Wright, Mead & Draba, 1976; Wright y Douglas, 1977) como la que se muestra en la figura 1, en la que se representa la nube de puntos correspondientes a los reactivos medidos en la prueba para los dos grupos. La recta de NO-SESGO corresponde con la recta de pendiente *m=1* donde coinciden las medidas de los reactivos para hombres y para mujeres. En consecuencia, los reactivos que se ubican a la derecha de dicha recta resultan más difíciles para las mujeres que para los hombres ([Clase 1]: sesgo desfavorable para las mujeres); los reactivos ubicados a la izquierda resultan más difíciles para los hombres ([Clase 2]: sesgo desfavorable para los hombres).

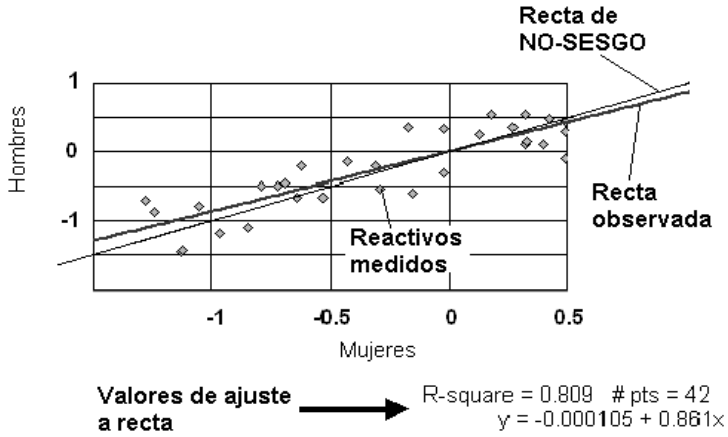

Figura 1. Interpretación del análisis logístico de sesgo

Con ayuda del modelo de Rasch se obtuvo la escala métrica asociada a la prueba. El modelo involucra una hipótesis de independencia respecto a los sujetos, que indica que la escala debe ser prácticamente igual si la prueba no está sesgada con respecto a hombres o a mujeres. Una prueba sesgada debe mostrar diferencias significativas en la métrica como muestra la Figura 2 con datos hipotéticos.

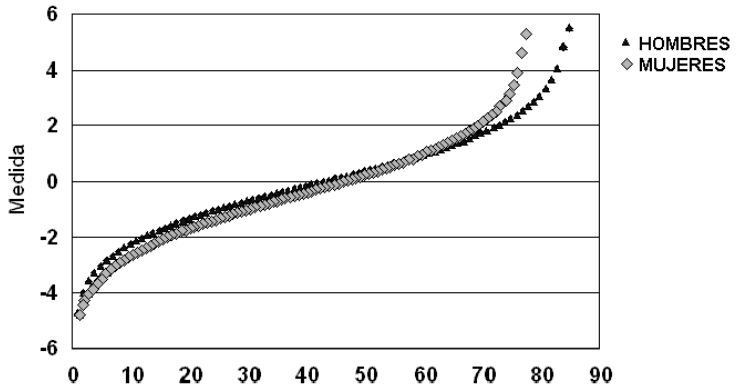

Figura 2. Escala métrica de una prueba sesgada aplicada a dos grupos de personas

2. Análisis de funcionamiento diferencial de los reactivos.

Para estudiar el sesgo por reactivo se emplearon dos procedimientos: el modelo de Rasch (Draba, 1977; Wright y Stone, 1988b; Wright, Mead y Draba, 1976) y una prueba χ^2 para detectar el posible sesgo en cada ítem individualmente (Lord, 1980; Hambleton y Rogers,1989; Holland y Wainer, 1993; Dorans y Holland, 1993; Thissen, Steinberg y Wainer, 1993), pero se decidió no usar otros modelos logísticos, Mantel-Haenszel, etc., por estar contenidos en cierto modo dentro del modelo de Rasch.

a) *Modelo de Rasch*. Este modelo permite identificar los casos puntuales donde se presenten sesgos hacia uno u otro sexo. Dadas las medidas de las dificultades que presenta un mismo ítem en hombres y en mujeres (Notadas como *DH* y *DM*, respectivamente) y los respectivos errores estándar *sh* y *sm* del reactivo para los grupos, se establece el estadístico *t* que se define en la ecuación 6, el cual sigue la distribución *t* de Student con *N-2* grados de libertad, siendo *N* el número de reactivos en estudio

$$t = \frac{DH - DM}{\sqrt{sh^2 + sm^2}} \qquad (6)$$

Este estadístico se aplicó combinadamente con un criterio de diferencia de desempeño interítem dado por el numerador *SHIFT=DH–DM*, que se considera significativo si su valor absoluto es mayor que 0.5 (criterio empírico conocido como half logit rule). Cuando el valor de SHIFT es positivo el reactivo es más difícil para los hombres. En este caso el modelo de Rasch se prefirió respecto al modelo de la Teoría de la Respuesta al Item y sobre otros modelos descriptivos (Hambleton y Rogers, 1989; Johnson, 1994).

b) Se aplicó un *modelo de funcionamiento diferencial del item* (DIF, siglas de differential item functioning) propuesto para identificar el sesgo reactivo por reactivo, que emplea un modelo logístico uniparamétrico equivalente al modelo de Rasch, y una prueba de hipótesis con base en la distribución χ^2. El estadístico aparece definido en la ecuación 7; donde *bh* y *bm* son las medidas del reactivo para hombres y mujeres, respectivamente, y *var(bh)* y *var(bm)* son sus varianzas. La diferencia entre los grupos se considera significativa si $\chi^2 > \chi^2_{1,.05}$ para 1 grado de libertad con una significancia del 5%

$$\chi^2 = \frac{(bh - bm)}{\text{var}(bh) + \text{var}(bm)} \qquad (7)$$

3. Diferencias de desempeño por sexo.

Para identificar las diferencias por géneros se emplearon dos métodos: un modelo clásico y el logístico

a) *Modelo clásico*: Se utilizó la prueba de diferencia estándar entre medias (Cohen, 1988; Cole, 1981), mediante el estadístico *d* definido en la ecuación 8, donde μh y μm son las medias de puntuaciones de los hombres y de las mujeres respectivamente, y σ es la desviación estándar de toda la población. El estadístico *d* es un número independiente de la unidad de medida con la que se califica a los sujetos. Si *d* es positivo la diferencia favorece a los hombres, es decir, existe sesgo desfavorable para las mujeres. La interpretación de este estadístico se realiza en función del área no común, *U1*, entre las distribuciones de la prueba aplicada a hombres y la de la prueba aplicada a mujeres. El área *U1* corresponde a la probabilidad de que ambas distribuciones sean estadísticamente diferentes.

$$d = \frac{\mu h - \mu m}{\sigma} \qquad (8)$$

b) *Modelo logístico*: Se utilizó de nueva cuenta el modelo de Rasch, identificando las diferencias entre las medidas de los dos grupos. Siendo *BH* y *BM* la medida de hombres y de mujeres respectivamente, $DR = \Sigma(BH\text{-}BM)$ se considera significativa si su valor absoluto es mayor de 0.1 ($|DR| > 0.10$)

Igualmente se empleó el estadístico *d*, pero utilizando las medidas de los sujetos en lugar de los porcentajes de aciertos. Estimando el error estándar de medida *ESM* por medio del modelo de Rasch, se definió un intervalo de aceptación para los valores de *DR* y del estadístico *d*. Estos se consideran significativos si su valor absoluto es mayor que *1.96 ESM*.

Resultados

Sesgo global:

De acuerdo con los valores del modelo clásico se observó que las pruebas no están centradas por diseño, lo que hace que resulten medianamente difíciles para todos los sujetos; los valores del índice de dificultad (media de aciertos) son inferiores a 0.5 o 50% (ver figura 3). La media de los índices de dificultad de cada parte de la prueba resulta más grande para los hombres que para las mujeres lo cual indica que la prueba es ligeramente más fácil para los primeros. La diferencia media es de *0.02875* (*2.8%*) en la prueba TIPO F, mientras que la diferencia más grande es de *0.05* (*5%*) en la prueba TIPO B.

Todas las diferencias del índice de dificultad son positivas (figura 4), lo cual implica una tendencia a favor de los hombres, pero dicha tendencia no es significativa, porque los valores de *ZH,M* son inferiores a *1.0* (el valor más grande se encuentra en el área Global) y no se

tiene un patrón sistemático entre los tipos de prueba, por ejemplo, Razonamiento Verbal es el área que muestra mayores diferencias en las pruebas tipo A y B, mientras que resulta ser la más pequeña en las C, D y F.

El segundo estudio, realizado con el modelo logístico, presentó las nubes de reactivos, en algunos casos, sin diferencia entre la recta observada y la de NO-SESGO. Con excepción de los casos B-3, E-2 y F-1 (ésta última se muestra en la figura 5) que presentaron una pendiente que cae en el rango de diferencia mediana, todas las demás áreas no muestran diferencias importantes, lo que implica que no se puede concluir que hay un sesgo relevante en la prueba.

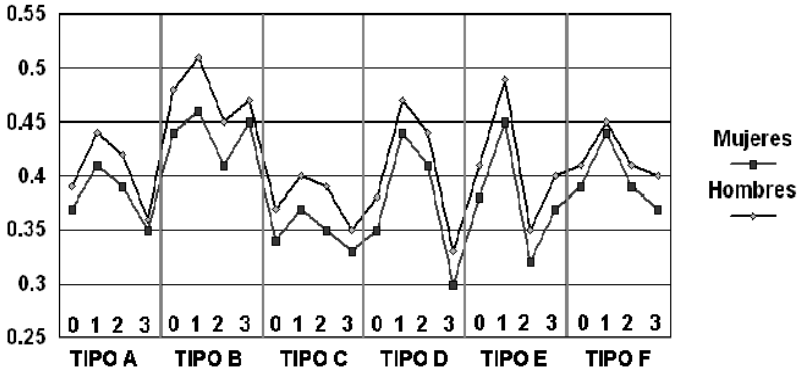

Figura 3. Resultados de *ZH,M* o índice de dificultad media para cada tipo de prueba y área de evaluación

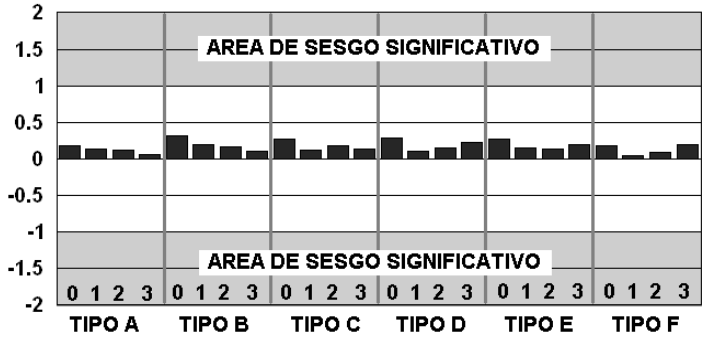

Figura 4. Valores de *ZH,M* por tipo de prueba para estudio de sesgo global

En la tabla 3 se presenta el resumen de los valores de ajuste a la recta en cada prueba y subprueba y se señalan los casos más desfavorables de acuerdo con la clase de recta (1 ó 2). Las diferencias globales de las pruebas tampoco son significativas desde el punto de vista del análisis de Rasch. Como se puede ver en la figura 6 todas las pruebas y sus áreas están prácticamente centradas alrededor de *0*, con una diferencia menor a una milésima. Desde luego se puede apreciar que hay áreas donde las diferencias ocurren a favor de los hombres y otras a favor de las mujeres, lo cual no se podía apreciar con el análisis clásico usando el

estadístico *ZH,M*. Sin embargo, en todos los casos se trata de valores muy pequeños y no son sistemáticos, como se puede observar en la figura 7.

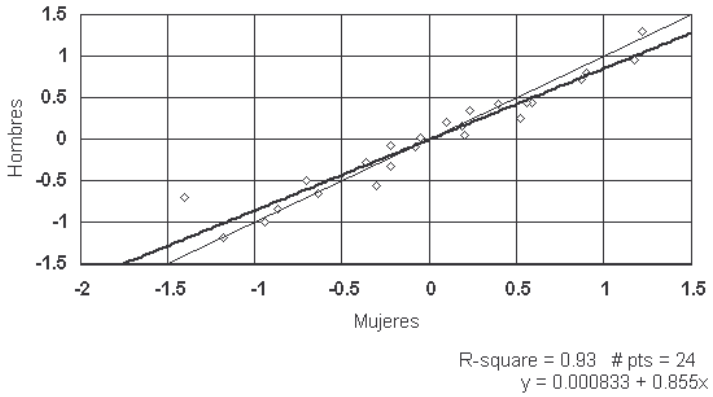

Figura 5. Diferencias observadas en la prueba F-1 con rango de diferencia mediana

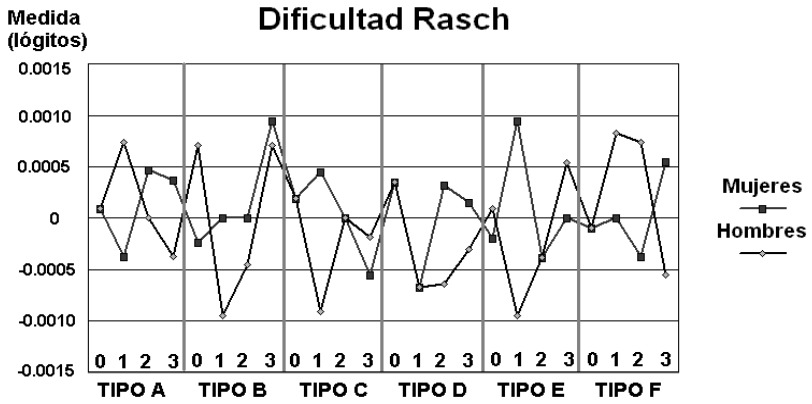

Figura 6. Diferencia global en lógitos para las pruebas

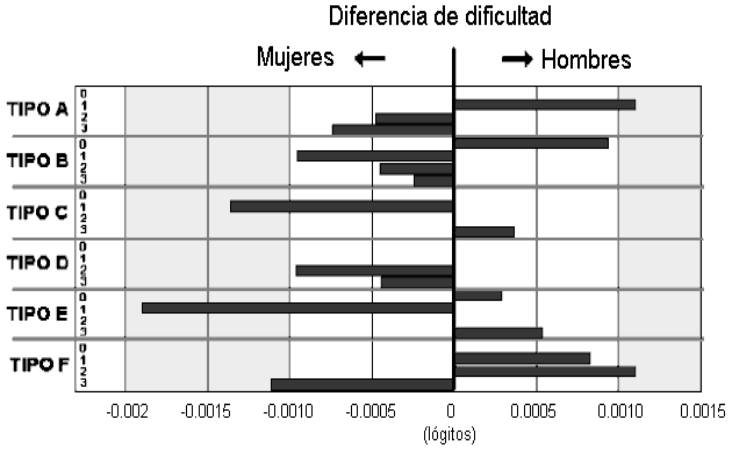

Figura 7. Diferencia global en lógitos para los diferentes tipos de pruebas

Tabla 3
Valores de ajuste a la recta

Tipo de examen	Area	Pendiente	Clase	Interpretación	Tipo de examen	Area	Pendiente	Clase	Interpretación
A	0	0.995	[1]	Pequeña	D	0	0.992	[1]	Pequeña
A	1	1.010	[2]	Pequeña	D	1	0.905	[1]	Pequeña
A	2	1.050	[2]	Pequeña	D	2	0.959	[1]	Pequeña
A	3	0.949	[1]	Pequeña	D	3	1.020	[2]	Pequeña
B	0	0.905	[1]	Pequeña	E	0	0.982	[1]	Pequeña
B	1	0.908	[1]	Pequeña	E	1	0.929	[1]	Pequeña
B	2	0.955	[1]	Pequeña	E	2	0.890	[1]	Mediana
B	3	0.861	[1]	Mediana	E	3	1.040	[2]	Pequeña
C	0	1.020	[2]	Pequeña	F	0	0.954	[1]	Pequeña
C	1	1.060	[2]	Pequeña	F	1	0.855	[1]	Mediana
C	2	1.030	[2]	Pequeña	F	2	0.971	[1]	Pequeña
C	3	0.977	[1]	Pequeña	F	3	0.988	[1]	Pequeña

Desde el punto de vista global se pueden identificar las áreas más alejadas de la Recta de NO-SESGO. En la Figura 7 se presentan las zonas que definen un semi-rango de aceptación de *0.001*, se confirma que las pruebas más alejadas son A1 y F2, que resultan más difíciles para los hombres, y E1, C1 y F3, más difíciles para las mujeres (ver figura 8)

La escala métrica de la prueba y de sus partes, aplicada a hombres y mujeres, determinó que la prueba no presenta sesgos generales o por área con respecto a uno u otro sexo, ya que las curvas coinciden en todos los casos. En la figura 9 se presentan las métricas del examen global (0) y de las partes 1, 2 y 3 de la prueba tipo A; puede verse que las curvas son prácticamente idéntica en las pruebas

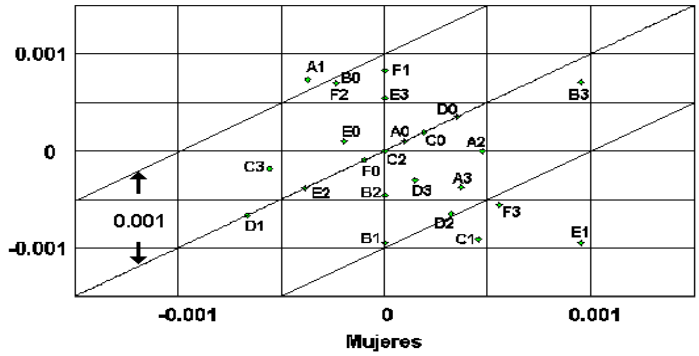

Figura 8 Diferencia de dificultad en las áreas

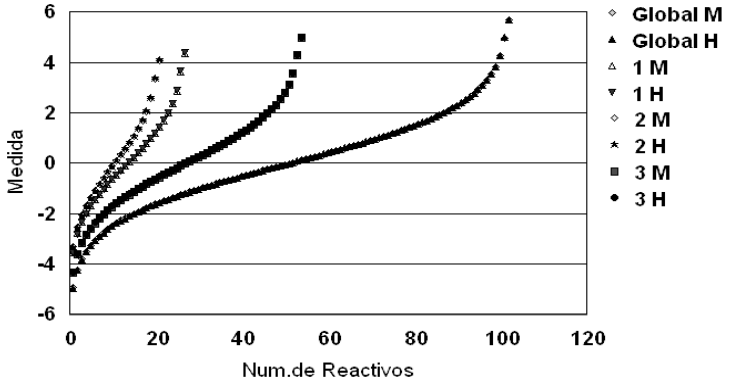

Figura 9. Escala métrica de la prueba Tipo A

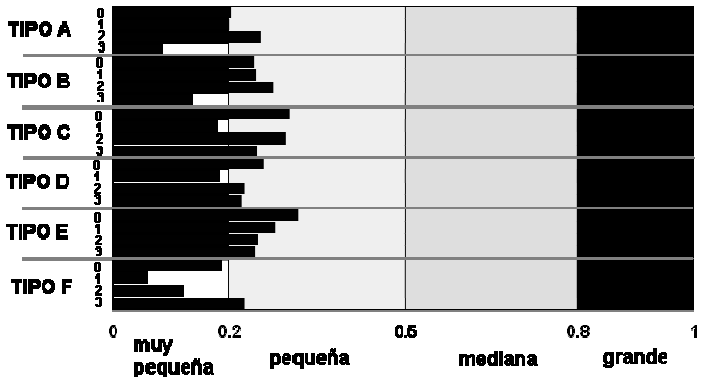

Figura 10.Diferencias de dominio por género

Sesgo por reactivo

Dada la cantidad de elementos obtenidos con el estadístico *Zt*, el criterio de diferencia *SHIFT* y la prueba de χ^2 al analizar cada tipo de prueba con sus respectivas áreas, se prefirió presentar los casos significativos para los tres criterios de sesgo en la tabla 4. Aquí se pueden ver todos los reactivos que muestran valores significativos con alguno de los criterios (*SHIFT*, *Zt* y χ^2), aunque no representen un sesgo para toda la prueba.

Diferencias entre géneros

La primera comparación de dominio mostró que existe una tendencia favorable, no significativa, para los hombres, éstos demuestran mayor dominio que las mujeres; sin embargo, en todos los casos las diferencias son inferiores a *5%* (en términos de Grado de Dificultad emitido en la escala 0-100%). Todas estas diferencias quedan dentro del error de medida de la prueba. Similarmente con el estadístico *d* se observó que en todos los casos las diferencias están a favor de los hombres, pero todas las diferencias se encuentran en el rango de "pequeña" o "muy pequeña" (ver figura 10). De otra parte el estudio de diferencias de género que se realizó con las medidas de Rasch, presentó el plano de medidas de mujeres y hombres comparadas contra la recta de NO-SESGO (Figura 11), y todas las medidas cayeron

en la zona cercana a dicha recta, con un rango no mayor a 0.23. Atendiendo a que el error estándar de medida para los sujetos alcanzó un máximo de 0.61, se concluyó que las diferencias observadas no son significativas. Finalmente, en la figura 12 se presentan las significancias de las diferencias desde el enfoque del estadístico *d,* pero esta vez en función de las medidas de los sujetos, en lugar de los porcentajes de aciertos. Todos los valores de este estadístico cayeron en las categorías de "muy pequeña" o "pequeña", aunque se apreció una tendencia de mayor dominio por parte de los hombres. Empleando el error estándar se comprobó que las diferencias no son significativas.

Tabla 4
Casos significativos del sesgo por reactivo

Num	BM	BH	ESM M.	ESM H.	SHIFT	Zt	χ^2	Desfavorable para Mujeres	Desfavorable para Hombres
Tipo A, Area 0					**Σ = -1.27**	**Σ Zt = - 25.4**		**X**	
40	-0.41	-1.07	0.03	0.04	-0.66	-13.2	174.24	X	
118	0.11	-0.5	0.03	0.04	-0.61	-12.2	148.84	X	
Tipo B, Area 0					**Σ = + 0.65**	**Σ Zt = + 2.6**			**X**
40	-0.2	-0.84	0.16	0.21	-0.64	-2.42	5.88	X	
102	-1.32	-0.67	0.18	0.2	0.65	2.42	5.84		X
90	-0.23	0.36	0.16	0.2	0.59	2.3	5.31		X
76	-0.5	0.08	0.16	0.2	0.58	2.26	5.13		X
10	-0.36	0.18	0.16	0.2	0.54	2.11	4.45		X
25	-0.17	-0.71	0.16	0.21	-0.54	-2.05	4.18	X	
116	0.43	-0.1	0.17	0.2	-0.53	-2.02	4.08	X	
Tipo C, Area 0					**Σ = - 3.48**	**Σ Zt = - 69.6**		**X**	
82	-0.82	-1.92	0.03	0.04	-1.1	-22	484	X	
7	-1.23	-1.99	0.03	0.04	-0.76	-15.2	231.04	X	
1	-1.59	-2.16	0.03	0.04	-0.57	-11.4	129.96	X	
3	0.34	-0.2	0.03	0.04	-0.54	-10.8	116.64	X	
42	0.03	-0.48	0.03	0.04	-0.51	-10.2	104.04	X	
Tipo D, Area 0					**Σ = - 0.63**	**Σ Zt = - 9.39**		**X**	
55	-0.12	-0.75	0.06	0.03	-0.63	-9.39	88.2	X	
Tipo E, Area 0					**Σ = - 1.17**	**Σ Zt = - 22.17**		**X**	
5	-0.55	-1.19	0.03	0.04	-0.64	-12.8	163.84	X	
50	0.84	0.31	0.04	0.04	-0.53	-9.37	87.78	X	
Tipo F, Area 0					**Σ = + 0.20**	**Σ Zt = + 2.25**			**X**
4	-1.59	-0.86	0.05	0.04	0.73	11.4	129.98		X
78	0.71	0.14	0.05	0.04	-0.57	-8.9	79.24	X	
62	0.93	0.38	0.05	0.04	-0.55	-8.59	73.78	X	
93	0.96	1.55	0.05	0.05	0.59	8.34	69.62		X

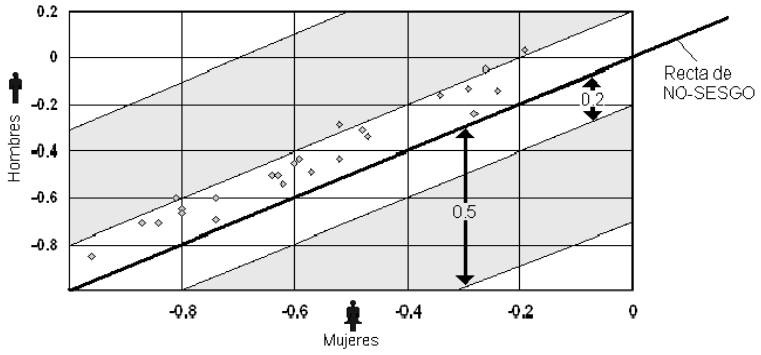

Figura 11 Diferencia de dominio

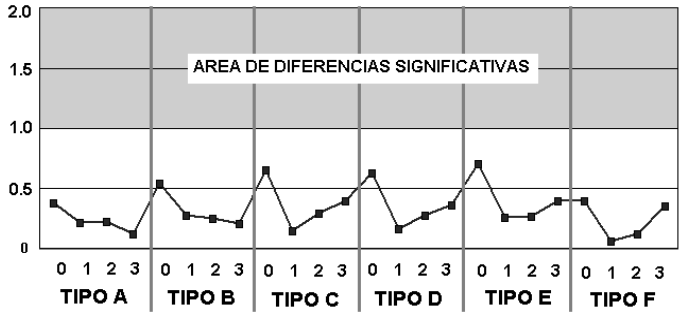

Figura 12. Identificación de significancia en las diferencias de género

Conclusiones y sugerencias

Con base en los resultados obtenidos a través del conjunto de análisis estadísticos realizados en torno al Examen de Ingreso a la Universidad Autónoma Metropolitana, se determinó que las diferencias entre hombres y mujeres están dadas probablemente por los conocimientos y habilidades de cada sujeto independientemente de su condición de género. No se encontraron evidencias estadísticamente significativas que indiquen un sesgo global a favor de los hombres (que desfavozca a las mujeres), por lo tanto, se rechazan las hipótesis [H1] y [H2].

Pese a que no se encontró un sesgo general a favor de los hombres, sí se identificaron reactivos que se comportan de manera diferencial entre los sexos, demostrando con ello que la presencia de reactivos con sesgo en una u otra dirección no influyen significativamente en el comportamiento global de la prueba. Sin embargo, no deja de ser importante realizar una revisión particular de los reactivos que presentaron sesgo utilizando la tabla de especificaciones. Atendiendo al desempeño de los aspirantes, el estudió muestra que el desempeño es mejor (aunque no significativo) por parte de los hombres en todos los casos, inclusive en las pruebas donde hay un sesgo favorable para las mujeres. Esta diferencia de desempeño tiene valores inferiores al error de medida propio del instrumento y se clasifica como "pequeño" o "muy pequeño".

En conclusión, las diferencias generales mostradas en todas las áreas a favor de los hombres pueden atribuirse a un mejor desempeño de los aspirantes varones, en función de sus antecedentes o de las experiencias de aprendizaje del ciclo anterior asociadas con el perfil de ingreso a la Universidad y no a un sesgo inducido por los reactivos de la prueba. Es evidente que, a nivel general, no hay problemas significativos con la prueba que ameriten hacer ajustes a los resultados de los estudiantes, pero también es claro que sí puede haber alguna repercusión desde el punto de vista particular de un aspirante cuando éste se encuentra cerca del punto de corte, porque puede verse afectado en el proceso de admisión, pudiendo quedar por abajo del punto de corte por la presencia de uno o dos reactivos sesgados para la medición. Estas conclusiones son muy similares a las reportadas por Garner & Engelhard (1999) o por McCornack & McLeod (1988) en otros exámenes estandarizados.

Con base en los resultados obtenidos pueden plantearse estas preguntas generales: ¿qué mecanismo se debe establecer para calificar "igualitariamente" a ambos sexos? y ¿qué procedimientos son los más convenientes entre los siguientes: (a) validar los reactivos del examen y conservar solo los que no presenten sesgo, (b) ponderar el número de ítems o la calificación con base en el sexo, (c) normalizar por sexo?. Por otra parte, atendiendo a la permanencia de los estudiantes en la institución (mayor deserción de los hombres que de las mujeres) y a la terminación de sus estudios (mayor eficiencia terminal entre las mujeres que entre los hombres), se puede plantear una pregunta relacionada con la política institucional: ¿cuál es la cuota de ingreso por sexo que desea la Universidad? Es decir, si se tiene una mayor eficiencia académica y terminal con las mujeres, podría alguien afirmar que conviene que entren más mujeres que hombres, el problema que se genera con esto es ¿cómo establecer cantidades de aspirantes aceptados por sexo, admitiendo mayor cantidad de mujeres que se hombres, sin que se convierta en una medida política que afecte los derechos de los estudiantes hombres?. Para responder a estas preguntas, se sugieren las siguientes acciones que pueden perfeccionar la prueba y, en su caso, ayudar a identificar causas reales de sesgo, no necesariamente asociadas con el género:

a) Revisar los reactivos que presentaron sesgo para señalar y clasificar los elementos que producen el error sistemático a favor cualquiera de los géneros, con el fin de corregirlos y tratar de evitarlos en el diseño de nuevos reactivos.

b) Revisar la tabla de especificaciones de la prueba para incluir áreas "neutras" y cancelar las de tendencia "masculina" o "femenina" (Gipps & Murphy, 1994) si es que existen. Debe anotarse que la presencia de un ítem con sesgo hacia uno de los sexos no implica que toda el área contenga un sesgo por sí misma.

c) Estudiar el posible sesgo debido a otras variables asociadas con el desempeño de los aspirantes a la UAM (escolaridad de los padres, ingreso familiar, tipo de escuela de procedencia, ubicación geográfica, etc.) y determinar los factores más significativos que incidan en las diferencias de desempeño.

d) Realizar un estudio factorial para identificar los reactivos asociados con sexo, si es que existe asociación, así como la participación de cada factor en el comportamiento de los sujetos.

e) Realizar un estudio específico de asignación de los aspirantes, tomando en cuenta la presencia de los reactivos con sesgo, después de "balancear" el instrumento para que el valor de sesgo total Zt sea prácticamente cero. En este caso se determina el número de estudiantes que ingresan por cada sexo con la prueba original. Una vez que la prueba se iguala entre los dos grupos de estudiantes (o proceso de "equating" por sexo), identificar cuántos estudiantes ingresan por cada sexo. Determinar las diferencias de admisión y los puntajes en los cuales se presentan las diferencias.

f) Repetir el análisis de sesgo por reactivo en futuras aplicaciones de manera de tomar una decisión sobre la permanencia de los reactivos para su calificación o eliminación.

Todos estos estudios sugieren que la prueba sea revisada desde otros enfoques complementarios, con objeto de disipar las objeciones que pudieran hacerse respecto a un posible sesgo favorable a los hombres. En particular resulta indispensable contar con el conjunto de ítems al momento de la revisión. Para este estudio no se dispone del banco de ítems por cuestiones de confidencialidad impuestas por la Institución.

Referencias

Berk R.A. (1980) *Handbook of methods for detecting test bias*. The Johns Hopkins University Press. Baltimore, EUA

Cohen, J. (1988) *Statistical power analysis for the behavioral sciences*. LEA, Hillsdale, N.J. Cap.2, 20-40.

Cole, N.S. (1997) *The ETS gender study: How females and males perform in educational settings*. Princeton NJ: Educational Testing Service, 1-33

Cole N.S. (1981) *Bias in testing*. American Psychologist. 36(10) 1067-1077

Dorans N.J., Holland P.W. (1993). DIF Detection and description: Mantel-Haenszel and Standardization, en P. W. Holland & H. Wainer (Eds). *Differential item functioning*. LEA, Hillsdale NJ. 35-66

Draba R.E. (1977) *The identification and interpretation of item bias*. Educational Statistics Laboratory. Memorandum 25. Marzo. University of Chicago. 1-11

Flaugher R.L. (1978) *The many definitions of test bias*, American Psychologist. Jul 1978. 671-679

Garner M. & Engelhard G. (1999) *Gender differences in performance on multiple-choice and constructed response mathematics items*, Applied measurement in edication, 12(1), 29-51

Gipps C. y Murphy P. (1994) *A fair test? Assessment, achievement and equity*. Open University Press. Buckingham EUA, Cap. 2, 28-64

Grimby G. (1998) *Useful reporting of DIF* .Rasch Measurement Transactions*,* 12:3, 651

Hambleton R.K., Swaminathan H., Rogers H.J. (1991) *Fundamentals of item response theory*, SAGE publications, California, EUA, Cap. 8. 109-122

Hambleton R.K., Rogers H.J. (1989) *Detecting potentially biased test items: Comparison of IRT Area dn Mantel-Haenszel Methods*, Applied Measurement in Education, 2(4), 313-314.

Holland P.W., Wainer H. eds (1993) *Differential item functioning* LEA, Hillsdale NJ. EUA, 453 pp.

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, Appendix N. 967-984

Linacre J.M. (1996) *DIF in polytomous items*. Rasch Measurement Transactions, 10:3, 520

Linacre, J.M, Wright B.D. (1994-2005) *A user´s guide toBIGSTEPS/WINSTEPS*. Chicago. MESA Press

Lord M. (1980) *Applications of item response theory to practical testing problems*. LEA, Hillsdale, N.J.. Cap. 14, 212-224

McCornack R.L. y McLeod M.M. (1988) *Gender bias in the prediction of college course performance*, Journal of Educational Measurement, 25 (4), 321-331.

Rasch, G. (1960). *Probabilistic models for some intelligence and attainment test*. Copenhague: The Danish Institute for Educational Research. Republicado (1980) MESA Press, Chicago. 199 pp.

Thissen D., Steinberg L., Wainer H. (1993). *Detection of differential item functioning using the parameters of Item Response Theory* en P. W. Holland & H. Wainer (Eds). *Differential item functioning*. Hillsdale NJ: LEA, 67-113

Tristán L.A., Leyva B. Y. Vidal U.R. (2001a), *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*. Educación y Ciencia, México. Enero-junio. Vol. 5, No.9 (23), 67-81

Tristán L.A., Leyva B. Y. Vidal U.R. (2001b), *Estudio de sesgo y de diferencias entre géneros de los exámenes nacionales de ingreso a la educación média superior y superior del Ceneval*. Educación y Ciencia, México. Enero-junio. Vol. 5, No.9 (23), 83-104

Wright B.D., Mead R. y Draba R. (1976) *Detecting and correcting test item bias with a logistic response model.* Research Memorandum N. 22. The University of Chicago,Statistical Department of Education, EUA, 1-24

Wright B.D., Douglas G.A. (1977) *Best procedures for sample-free item analysis*. Applied Psychological Measurement, 1 (2) 281-295

Wright B.D., Stone M.H. (1979) *Best test design*, Chicago: MESA press

Wright B.D., Stone M.H. (1998a) *Diseño de mejores pruebas*, México. Ceneval

Wright B.D., Stone M.H. (1988b) *Identification of item bias using Rasch measurement*. Research memorandum N. 55, Mesa Press, 1-11

Manuscrito recibido en septiembre de 2005
Aprobado para publicación en noviembre de 2005