Educación y Ciencia *Nueva época Vol. 5 No. 9 (23)* *Enero - Junio 2001*
67- 81

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

Agustín Tristán López
Yolanda Leyva Barajas
Rafael Vidal Uribe

Sinopsis

Una de las preocupaciones de los evaluadores así como de la propia sociedad es saber si la prueba es 'justa' al momento de evaluar a los sujetos. En especial se tiene siempre la inquietud de identificar sesgos, particularmente por sexo, de manera de evitar cualquier fuente de discriminación que pudiera ocurrir en una población de aspirantes a Educación Media Superior, pudiendo darse el caso de que la prueba haga que las mujeres resultaran afectadas por el tipo de reactivos o el planteamiento de la prueba. El trabajo presenta los resultados de análisis de sesgo efectuado por medio de la técnica de Rasch sobre el EXANI I, Examen Nacional de Ingreso al Bachillerato diseñado y promovido por el Ceneval.

En este trabajo se demuestra que el EXANI I es un instrumento bien calibrado, centrado, que no presenta diferencias significativas en la forma de responder entre hombres y mujeres. Con la calibración usando el modelo de Rasch, se muestra que las diferencias que se observan en el desempeño de los sustentantes por sexo no están relacionadas con defectos en la prueba.

El trabajo analiza el desempeño a nivel global y por cada una de las áreas del EXANI I. Se aprecian las diferencias de desempeño en cada caso, en especial se observaron diferencias a favor de los hombres en 7 de las áreas y a favor de las mujeres en 2 de las áreas, quedando un área en la cual la diferencia fue prácticamente nula. De estas diferencias de desempeño por sexo, solamente tres son significativas: español a favor de las mujeres, geografía y habilidad matemática a favor de los hombres.

Las diferencias observadas en cada una de las áreas pueden ser distintas a las reportadas aquí para otras aplicaciones del EXANI I. Lo importante de este estudio es resaltar que las diferencias que se encuentran en el desempeño de hombres y mujeres en el EXANI I son atribuibles a las propias diferencias de dominio que tiene cada género y no a sesgos del instrumento.

Términos clave: <Investigación> <investigación educacional> <diferencias entre sexos> <pruebas estandarizadas> <México>

Abstract

One of the concerns of the testers, and the society, is to know whether the test is 'fair' when administered. There is especial interest in knowing if it is not biased toward one of the sexes, in order to avoid any form of discrimination toward the candidates entering Higher Education Institutions, since there is a possibility that women may be harmed by the way the items, or the test as a whole, are designed. This paper presents the results of a bias analysis done by means of the Rasch technique to the EXANI-I, National High School Admittance Examination designed and promoted by CENEVAL.

This paper shows that the EXANI-I is a well-balanced and well-calibrated instrument, which does not present significant differences in the way women or men answer it. Using the Rasch model for the calibration, it is demonstrated that there is no relationship between gender and defect of the test, observed in the performance of the candidates.

This paper analyses the global performance and also the performance in each area of the EXANI-I. The differences in performance were observed in each case, especially the difference toward men in 7 of the areas and 2 of the areas toward women. Only in one area the difference was practically null. Only three of the differences of performance by gender are significant: Spanish for women, Geography and Mathematical skills for men.

The differences observed in each area could be different to the ones reported here for other applications of the EXANI-I. The importance of this study is to highlight the differences found in the performance of men and women in the EXANI-I are ascribed to the differences in command for each gender and not to bias of the instrument.

Key terms: <Research> <educational research> <sex differences> <standardized test> <Mexico>

Introduccion

Supóngase el siguiente experimento: con una cinta graduada en centímetros se mide la estatura de un conjunto de personas (hombres y mujeres) y se obtiene que la media de estatura de los hombres es de 1.72m y la de las mujeres es de 1.66m. ¿Qué se puede afirmar respecto a la cinta utilizada para medir a las personas? ¿Puede 'calibrarse' la cinta a partir de las mediciones de las personas? Si la cinta es 'justa' ¿No sería de esperar que tanto los hombres como las mujeres tuvieran la misma media de estaturas? Puesto que las estaturas medias de los hombres y de las mujeres son distintas ¿la diferencia de resultados es un indicio de que la cinta mide de manera diferente a los hombres que a las mujeres?

Las preguntas que se hacen con relación a este experimento pueden parecer extrañas y de hecho son inconsistentes. En primer lugar, no es posible calibrar una cinta métrica a partir de mediciones de las personas. El patrón de medida para la cinta podría ser el "metro patrón" o los modernos instrumentos electrónicos de referencia. Para saber si una cinta está bien calibrada se debe comparar contra un patrón fijo de referencia, pero los objetos que van a medirse con la cinta no pueden ser, al mismo tiempo, patrones. Es posible que una cinta esté mal hecha, bastará con hacer la comparación de ella contra el patrón para decidir si está bien o mal, en cambio no es posible, a partir de la medición de las personas saber si hay error en la cinta.

La cinta puede ser justa (estar bien calibrada), pero esto no implica que las medias de estaturas de hombres y mujeres deban ser iguales. Podemos estar a favor de la igualdad entre los sexos, pero hay diferencias evidentes, por diversas razones: configuración física, hábitos alimenticios, funciones fisiológicas, aspectos sociales, cuestiones familiares, tradición cultural.

Si fuera deseable que tanto los hombres como las mujeres tuvieran la misma estatura media y esto no se obtiene, no se debe condenar a la cinta que revela las diferencias, sino irse a las causas del problema y, en su caso, buscar soluciones con cambios de alimentación, realización de ejercicios enfocados a incrementar el crecimiento, etc.

Pueden establecerse las mismas dudas en el caso de una prueba del dominio cognoscitivo, en cuyo caso las respuestas serán también las mismas. Supóngase que la media de dominio de los hombres en una pruebas de conocimientos es de 900 y la media de las mujeres es de 1000; se estaría tentado a decir que la prueba es 'injusta' para los hombres, ya que arroja una media más baja para ellos, pero al afirmar que hay injusticia se estarían cometiendo dos errores simultáneamente: el primero al hacer la hipótesis de que los hombres y las mujeres deben salir igual en el examen y el segundo error al pretender "calibrar" la prueba en función de los resultados de los sustentantes. Se olvida que la calibración del instrumento se debe hacer con un 'patrón' independiente y que no se puede calibrar la prueba usando a los sujetos como elemento de referencia.

Este trabajo presenta un estudio de género del EXANI I (Examen Nacional de Ingreso al Bachillerato) que ofrece el CENEVAL, utilizando la técnica de Rasch. Se ofrecen los resultados de los sujetos y la calibración del instrumento, pudiéndose observar las diferencias de desempeño en cada una de las áreas o temas explorados por la prueba.

Descripción del EXANI I

El EXANI I (Examen Nacional de Ingreso al Bachillerato) del CENEVAL, se ofrece desde 1994 en México y está enfocado a medir las habilidades y los conocimientos de los jóvenes que están por ingresar al Nivel Medio Superior. Se ha aplicado a más de 230 instituciones, en aproximadamente 1000 aplicaciones, y ha sido utilizado como instrumento de selección para ingreso al bachillerato de más de 2,150,000 sustentantes, de los cuales aproximadamente el 50% son hombres y el 50% son mujeres.

Hasta la fecha (noviembre de 1999) se han construido unas 65 versiones del EXANI I, con un total aproximado de 5,000 reactivos que han sido calibrados, revisados, mejorados o eliminados, en su caso, buscando contar con buenos instrumentos de

medición de los conocimientos y las habilidades de los aspirantes al bachillerato.

La prueba está constituida por 128 reactivos organizados en 11 temas o áreas, incluyendo el resultado global, como sigue:

Tabla 1.
Organización temática del EXANI-I

No.	Nombre del área o tema	Número de reactivos
1	Global	128
2	Habilidad verbal	24
3	Español	10
4	Historia	10
5	Geografía	10
6	Civismo	10
7	Habilidad matemática	24
8	Matemáticas	10
9	Física	10
10	Química	10
11	Biología	10

Objetivos del estudio de este trabajo y metodología

El estudio que se presenta aquí tiene por objetivos:

a. Mostrar la calibración del EXANI I y su independencia respecto al género.
b. Mostrar algunas diferencias de desempeño entre hombres y mujeres en dicha prueba.

Estos objetivos pretenden determinar si los resultados encontrados en los dominios o desempeños de los hombres y de las mujeres están asociados con las habilidades y capacidades propias de los sustentantes o si, por el contrario, son ocasionados por sesgos inherentes a la prueba. Si la prueba está sesgada y favorece a uno de los sexos, entonces deben tomarse medidas para corregir las calificaciones que se hacen de los sustentantes, de manera de equilibrar el sesgo inducido por la prueba. En cambio, si la prueba no muestra sesgos, entonces las diferencias observadas son un reflejo de las capacidades de los grupos de sustentantes.

La metodología seguida en el presente trabajo consiste de estos pasos:

(a) elección del marco teórico: paradigma de reactivo a utilizar;
(b) elección del software para realizar la calibración usando el paradigma elegido;
(c) elección de las muestras de sustentantes hombres y mujeres;
(d) calibración del instrumento y análisis de diferencias en el funcionamiento de la prueba para los hombres y las mujeres;
(e) medición de los sujetos y análisis de diferencias en el desempeño de los hombres y de las mujeres;
(f) Conclusiones.

A continuación se detallan brevemente los aspectos principales de esta metodología.

Descripción del estudio

Marco teórico

En función de los trabajos más modernos presentados en la literatura, se utilizó el modelo de Rasch que permite calibrar los reactivos de una prueba y medir a los sujetos, por medio de un modelo logístico que transforma las medidas a una escala lineal, independiente de la población e independiente de la prueba.

El análisis de Rasch es una técnica rigurosa dentro de la denominada Teoría de la Respuesta al Item, ya que brinda la menor flexibilidad posible, al manejar un solo parámetro de ajuste a una curva teórica de comportamiento ideal de los reactivos. Es sabido que el denominado 'modelo de tres parámetros' es mucho más flexible y, por lo tanto, permite que el modelo ajuste mejor a los reactivos de una prueba dada. El modelo de Rasch, siendo más riguroso, obliga por lo tanto a que los reactivos sean redactados, revisados y analizados de manera más escrupulosa, para evitar rechazarlos.

No se pretende aquí hacer una presentación del análisis de Rasch, para ello el lector deberá acudir a las referencias, pero con objeto de hacer más completo este trabajo se presentan algunos aspectos generales del modelo de Rasch.

El modelo de Rasch es un modelo estocástico que proporciona una función con la cual se obtiene la probabilidad de respuesta 'p' de una persona de medida 'B' ante un reactivo de dificultad 'D', por medio de la expresión:

$$P = \frac{e^{(B-D)}}{1+e^{(B-D)}}$$

La función "p" toma valores de 0 a 1, con una distribución logística en el intervalo ($-\infty$, $+\infty$) para B y D. La unidad para B y D es el lógito, definido como el logaritmo del momio de aciertos (para B) o de fallas (para D). La medida de un sujeto o de un reactivo se define como el valor para el cual p=0.5, que es un valor independiente de la muestra e independiente de la prueba, por lo que se trata de valores calibrados.

Una prueba debe estar centrada, para ello el valor D debe ser lo más cercano al origen 0. Si la prueba está centrada, las medidas de los sujetos tienen un origen común de referencia y reflejarán los desempeños entre sujetos o entre grupos.

Para identificar si un reactivo se comporta de acuerdo con el modelo probabilista de Rasch, se debe estimar la bondad de ajuste. Para ello se utilizan generalmente dos elementos de medida en términos de los residuos cuadráticos:

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

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

Donde X representa el puntaje observado, E es el puntaje esperado, s^2 es la varianza de puntuaciones y N es el número de sujetos.

Estas medidas de ajuste tienden a 1 en el valor esperado, estableciéndose un rango de aceptación de 0.8 a 1.2. Los reactivos o áreas con valores por debajo de 0.8 o por arriba de 1.2 pueden representar problemas de planteamiento o problemas en los patrones de respuestas de las personas. Mientras más cercano a 1 se encuentre el reactivo su comportamiento será más parecido al paradigma estocástico del modelo de Rasch.

Cuando se dispone de medidas de dos grupos en un área (B1 y B2 respectivamente), se considera que se tiene una diferencia significativa si la distancia relativa en lógitos es superior al 0.10, lo cual se expresa:

distancia relativa =B2-B1 , significativa si DR > 0.10

El CENEVAL utiliza una escala de 700 a 1300 puntos 'CNE', por tratarse de una escala de

interpretación simple que fue elegida, entre otras razones, porque la medida de una persona nunca toma valores negativos. La escala CENEVAL puede obtenerse directamente a partir de las medidas 'B' en lógitos, con la expresión:

CNE = 1000 + 100 B

Para lo que se presenta en este trabajo se trabajará directamente con las medidas en lógitos, por ser las cantidades que proporciona el software empleado.

Método

Software elegido

Se eligió el programa BIGSTEPS (Universidad de Chicago) por ser el software que está siendo utilizado de manera más generalizada por el Rasch Measurement Special Interest Group de la American Educational Research Association. Existe otro software comercial, así como otros programas desarrollados especialmente para ajuste logístico, pero se juzgó conveniente el uso de un programa neutral y soportado por los investigadores de la Universidad de Chicago.

Muestreo

El software fija el número máximo de sujetos que es posible utilizar para efectuar el análisis de Rasch, dependiendo del número de reactivos que tiene la prueba y de la capacidad de memoria RAM disponible en la computadora. Para el caso del EXANI I que tiene 128 reactivos, el máximo de sujetos por población que puede emplearse es 9000. Una muestra representativa de los sustentantes que se han presentado en el EXANI I no requiere ser superior a unos 300 sujetos para un error del 5%, sin embargo se aprovechó que BIGSTEPS permite poblaciones de hasta 9000 sujetos por lo que se prepararon dos muestras: una de 9000 hombres y otra de 9000 mujeres que contestaron el EXANI I. El error que se tiene con muestras de 9000 sujetos es despreciable y, por lo tanto, los estudios que se presentan aquí son altamente significativos.

Las muestras de 9000 sujetos se prepararon tomando sustentantes en forma aleatoria, con ayuda del generador de números aleatorios del programa "EXPLORER" ©, que permite construir muestras aleatorias de cualquier tamaño a partir de una base de datos cualquiera.

Calibración del instrumento

Una vez que se tuvieron disponibles las muestras tomadas al azar, se prepararon las corridas de BIGSTEPS y se procesaron las respuestas de los sujetos, obteniéndose en primer lugar la calibración del instrumento. Las gráficas de la Figura 1 están tomadas de la salida del programa. Contienen polígonos de frecuencias en orientación vertical, contrastando del lado izquierdo las medidas de los sujetos y del lado derecho la calibración de los reactivos. Con ayuda del programa se obtuvieron gráficas similares para todas las áreas de la prueba, tanto para hombres como para mujeres. Aquí solamente se presentan las gráficas del área 1 (Global), posteriormente se proporcionan los resultados en forma tabular.

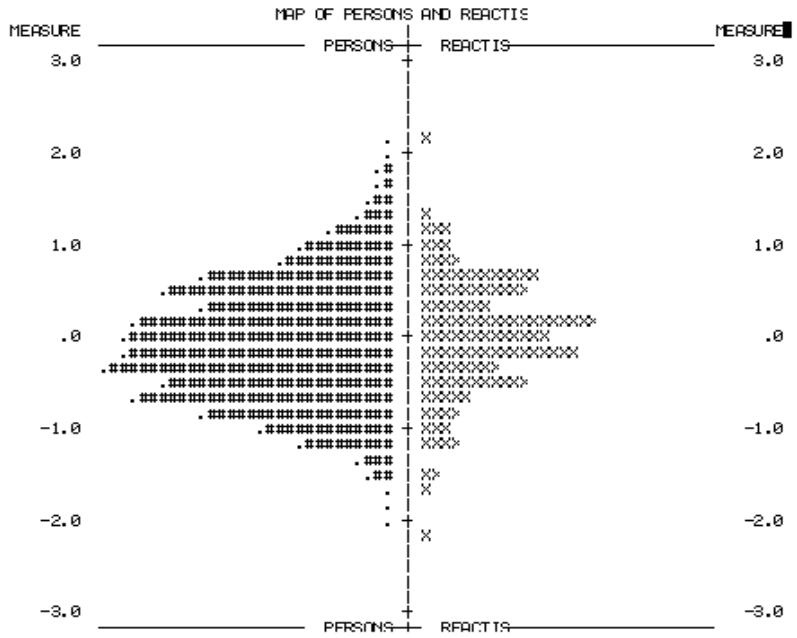

Figura 1a. Calibración Área 1 (Global) / mujeres.

ISSN 0188 3364

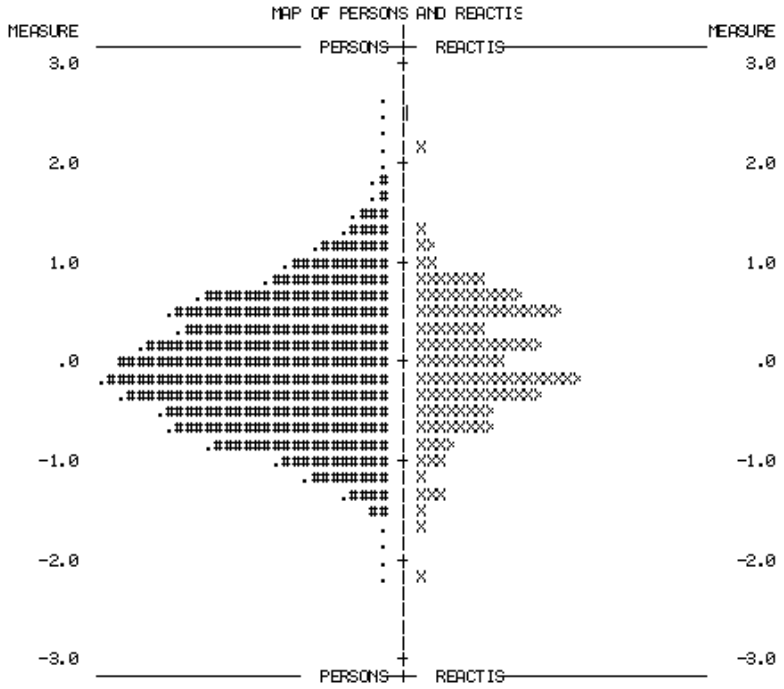

Figura 1b. Calibración Área 1 (Global) / hombres

Si se compara la calibración de la prueba aplicada a hombres y a mujeres, se podrá apreciar que el comportamiento del instrumento es prácticamente idéntico entre ambas poblaciones para todas las áreas que lo forman. Los valores cuantitativos de calibración se tienen en la tabla 2, donde se asientan la dificultad 'D' (en lógitos) de cada área de la prueba aplicada a hombres y a mujeres, así como los valores de ajuste INFIT y OUTFIT para ambos géneros.

Tabla 2.
Dificultad y ajuste de la prueba

Area	Dificultad D Mujeres	Dificultad D Hombres	INFIT Mujeres	INFIT Hombres	OUTFIT Mujeres	OUTFIT Hombres
1. Global	0.000313	-0.000000332	0.9992	0.9990	1.0030	1.0020
2. Habilidad verbal	-0.0000681	-0.00000000482	0.9983	0.9975	0.9999	0.9945
3. Español	0.000054	0.00000000769	1.0000	1.0010	0.9990	1.0030
4. Historia	-0.001	-0.001	1.0000	1.0000	1.0000	1.0030
5. Geografía	0.0012	-0.001	0.9990	0.9969	0.9960	0.9950
6. Civismo	0.000996	0.0000000215	0.9930	0.9950	1.0160	1.0140
7. Habilidad matemática	0.00125	0.000417	0.9995	0.9995	1.0025	0.9992
8. Matemáticas	-0.00000000739	0.001	1.0030	1.0030	1.0510	1.0570
9. Física	0.00000000173	-0.00000000626	0.9990	1.0010	1.0030	1.0040
10. Química	-0.001	0.0000000167	0.9990	0.9960	1.0030	1.0180
11. Biología	0.00000000334	0.00000000727	1.0010	1.0010	1.0040	1.0020

Obsérvese que la dificultad D de cada tema es prácticamente 0 en todas las áreas (el valor más grande corresponde con una diferencia de1 al millar, lo cual no representa diferencias significativas), tanto para hombres como para mujeres (figura 2). Esto garantiza que el instrumento está bien calibrado y que tiene un origen común para todos los sustentantes.

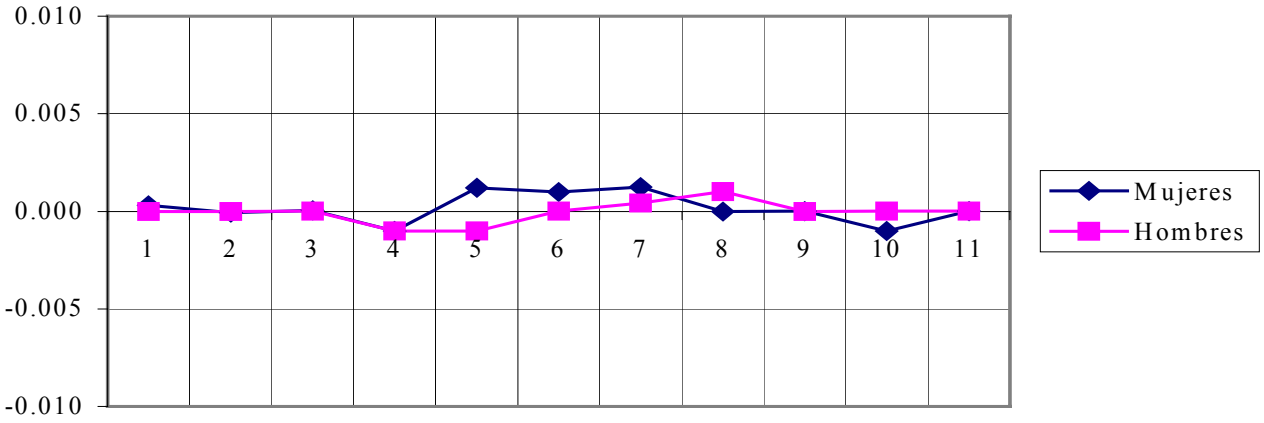

Figura 2. Valores de Dificultad D de la prueba

ISSN 0188 3364

Por su parte, los valores de INFIT y de OUTFIT, como era deseable, están muy cercanos a 1. La mayor discrepancia es de un poco más del 5% de OUTFIT en el área 8 (Matemáticas) tanto para hombres como para mujeres (figura 3). La única área donde difieren las respuestas de hombres y mujeres es en el área 10 (Química), con una diferencia no significativa, menor de 2% en OUTFIT.

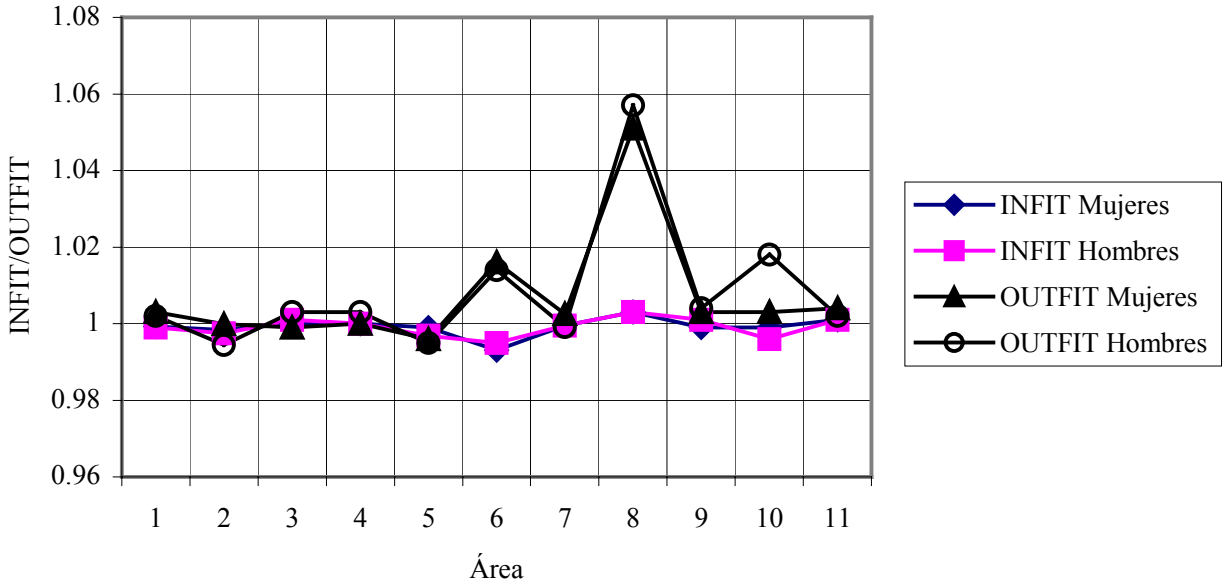

Figura 3. Valores de ajuste al modelo de Rasch

Dado que el intervalo de aceptación para INFIT y OUTFIT es de 0.8 a 1.2, se concluye que la prueba tiene un buen ajuste al modelo en todas las áreas, sin diferencia entre los géneros.

Puede concluirse que el instrumento calibrado por medio del modelo de Rasch ajusta perfectamente en todos los casos y, en consecuencia, no hay diferencias en los resultados de la prueba para los hombres y para las mujeres. El argumento de que la prueba tiene un sesgo en la forma de responder por parte de hombres y mujeres no tiene sustento, dado que el modelo de Rasch muestra que las calibraciones en ambos casos son prácticamente 0 sin distinción de género.

Medición de los sujetos

Una vez calibrado el instrumento, y que se ha mostrado que no tiene un desbalance o sesgo en función del género, se procedió a 'calificar' a los sustentantes. Se emplearon los resultados proporcionados por BIGSTEPS para construir la Tabla 3, donde se tienen las medidas de dominio 'B' (en lógitos) de los hombres y de las mujeres en cada área, así como el ajuste INFIT y OUTFIT en cada caso.

Tabla 3.
Medidas de dominio de los sujetos y ajuste al modelo de Rasch

Área	Medida B Mujeres	Medida B Hombres	INFIT Mujeres	INFIT Hombres	OUTFIT Mujeres	OUTFIT Hombres
1. Global	-0.0569	-0.0154	1.0004	1.0004	1.0027	1.0026
2. Habilidad verbal	0.2263	0.2221	1.0009	1.0033	1.0000	0.9949
3. Español	-0.0720	-0.2842	1.0011	1.0008	0.9984	1.0026
4. Historia	-0.2364	-0.1399	1.0013	1.0003	0.9999	1.0026
5. Geografía	0.1327	0.2537	1.0009	1.0020	0.9946	0.9959
6. Civismo	0.2068	0.3014	0.9994	0.9993	1.0157	1.0135
7. Habilidad matemática	-0.1262	0.0521	1.0001	0.9992	1.0022	0.9990
8. Matemáticas	-0.6698	-0.6092	0.9926	0.9922	1.0486	1.0536
9. Física	-0.1776	-0.1955	0.9993	0.9985	1.0050	1.0041
10. Química	-0.0926	-0.0639	0.9992	0.9979	1.0126	1.0184
11. Biología	-0.1140	-0.1038	0.9987	0.9993	1.0034	1.0027

Los valores de INFIT y OUTFIT cercanos a 1 obtenidos tanto para hombres como para mujeres, muestran que la forma de responder de los sustentantes se ajusta perfectamente al modelo de Rasch, con un comportamiento similar al obtenido en la sección anterior para la prueba misma.

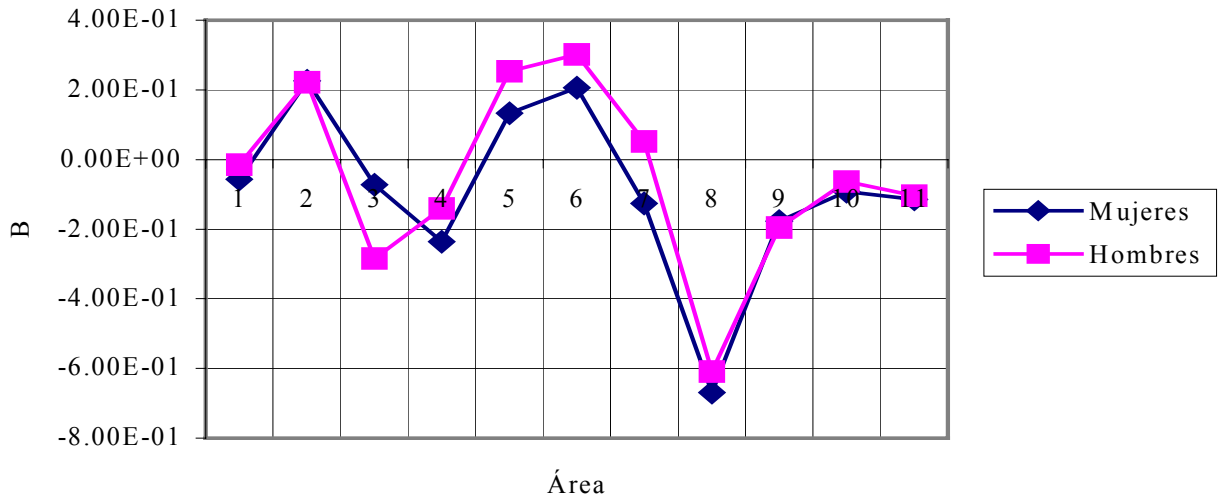

Figura 4. Medida de dominio de hombres y mujeres en la prueba

En lo referente al valor B, mientras más se aproxime a 0, quiere decir que el instrumento explora en los niveles precisos de dominio de los sustentantes. La figura 4 muestra el dominio B de hombres y mujeres; las diferencias de medida indican diferencias de dominio real entre las personas, atendiendo a que no hay relación alguna con diferencias en la forma de responder a la prueba, misma que no mostró discrepancias entre los géneros. Puede afirmarse por lo tanto que el dominio de los sustentantes en las áreas es muy parecido entre sexos, aunque se aprecian diferencias en algunas áreas.

Tabla 4.
Diferencia de dominio entre géneros

Área	B Mujeres	B Hombres	Diferencia Hombres-Mujeres
1. Global	-0.0569	-0.0154	0.042
2. Habilidad verbal	0.2263	0.2221	-0.004
3. Español	-0.0720	-0.2842	-0.212
4. Historia	-0.2364	-0.1399	0.097
5. Geografía	0.1327	0.2537	0.121
6. Civismo	0.2068	0.3014	0.095
7. Habilidad matemática	-0.1262	0.0521	0.178
8. Matemáticas	-0.6698	-0.6092	0.061
9. Física	-0.1776	-0.1955	-0.018
10. Química	-0.0926	-0.0639	0.029
11. Biología	-0.1140	-0.1038	0.010

De acuerdo con los datos obtenidos, se tiene una diferencia máxima de 0.21 lógitos en el área 3.- Español a favor de las mujeres y un máximo de 0.18 en el área 7. Habilidad Matemática a favor de los hombres. En el área 2. Habilidad Verbal, no se encuentran diferencias relevantes (figura 5).

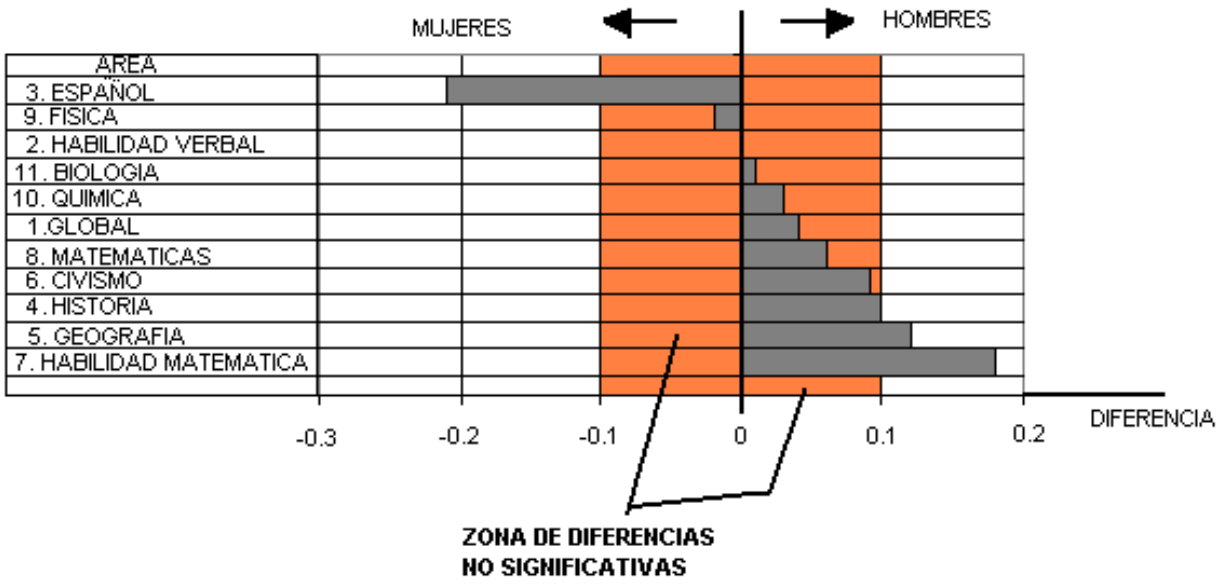

Figura 5. Diferencias de dominio en lógitos entre géneros

Como se mencionó en la sección 4.1, se considera que las diferencias entre uno y otro sexo son significativas cuando su valor relativo en lógitos es superior al 0.10, esto solamente tiene lugar en las áreas de Geografía y de Habilidad Matemática a favor de los hombres y de Español a favor de las mujeres. Curiosamente, la diferencia es más marcada en el caso de Español. Las diferencias observadas en las demás áreas no son significativas.

Conclusión

Se demuestra que el EXANI I es un instrumento bien calibrado, centrado, que no presenta diferencias significativas en la forma de responder entre hombres y mujeres. La prueba presenta un buen ajuste con el modelo de Rasch, garantizando que el instrumento se ha medido o calibrado independientemente de los sujetos que contestaron la prueba. La calibración garantiza que el instrumento está centrado para ambos sexos y que no muestra sesgos hacia uno u otro, por lo que las diferencias que pudieran observarse en el desempeño de los sustentantes por sexo no están relacionadas con defectos en la prueba.

Una vez realizada la calibración del instrumento y verificada su independencia, se calcularon los valores de medida de dominio de los sustentantes (hombres y mujeres) para cada una de las áreas, lo cual condujo a identificar que sí hay diferencias de desempeño entre sexos. En el caso particular de este estudio se observaron diferencias a favor de los hombres en 7 de las áreas y a favor de las mujeres en 2 de las áreas, quedando un área en la cual la diferencia fue prácticamente nula. De estas diferencias, solamente tres son significativas: español

a favor de las mujeres, geografía y habilidad matemática a favor de los hombres.

Las diferencias observadas en cada una de las áreas pueden ser distintas a las reportadas aquí para otras aplicaciones del EXANI I. Lo importante de este estudio es resaltar que las diferencias que se encuentran en el desempeño de hombres y mujeres en el EXANI I son atribuibles a las propias diferencias de dominio que tiene cada género y no a sesgos del instrumento.

La calibración de la prueba para hombres y mujeres usando la técnica de Rasch, brinda un procedimiento que responde a la necesidad de no utilizar a los mismos sujetos como referencia para la calibración del instrumento.

Referencias

Draba, E. (1977). "The identification and interpretation of item bias". Educational Statistics Laboratory, Memorandum 25. EUA, Chicago: Mesa press, pp. 11.

Linacre, M. & Wright B.D. (1994). "A user's guide to BIGSTEPS". MESA Press, Chicago, EUA, pp. 96

Lord, M. & Novick M.R. (1968) "Statistical theories of mental scores". EUA, Massachusetts Addison-Wesley Publishing Co. Reading, 568 pp.

Tristán, .A. "EXPLORER". Manual de usuario. Ingeniería Computarizada Integral. México, 1995. Pp. 30.

Tristán, .A. (1998). "Análisis de Rasch para todos". International Engineering and Statistics. Pp. 141.

Wright, D. & Stone M.H. (1988). "Identification of item bias using Rasch measurement". Research memorandum N. 55, Mesa Press, pp. 11.

Wright, D., Mead R. & Draba R. (1976). "Detecting and correcting test item bias with a logistic response model". Research Memorandum N. 22. Statistical Laboratory Department of Education. EUA: The University of Chicago, 24 pp.

Wright, D. & Douglas G.A. (1977) "Best procedures for sample-free item analysis". Applied Psychological Measurement, Vol. 1, N.2, Primavera 1977, pp. 281-295

Wright, D. & Stone M.H. (1979). "Best test design". EUA, Chicago: MESA Press, 222 pp. (También en versión en español: "Diseño de mejores pruebas", Ceneval, 1998).