El examen de ingreso a las licenciaturas de la UNAM: evidencias de validez de sus resultados

Melchor Sánchez Mendiola
Universidad Nacional Autónoma de México

Adrián Martínez González
Universidad Nacional Autónoma de México

Enrique Buzo Casanova
Universidad Nacional Autónoma de México

Área temática: A.12) Evaluación educativa.
Línea temática: 10. Evaluación a gran escala.
Tipo de ponencia: Reportes finales de investigación.

Resumen: Los exámenes de admisión a la educación superior son evaluaciones sumativas de alto impacto para los estudiantes y la sociedad, por lo que requieren abundante evidencia de validez para que las inferencias que se hagan de los resultados y el uso institucional de esta información sean apropiados. La Universidad Nacional Autónoma de México es la institución de educación superior más solicitada del país, anualmente aplica exámenes de ingreso para seleccionar a los estudiantes de sus licenciaturas. El instrumento es un examen estandarizado a gran escala, con papel y lápiz, compuesto de reactivos en formato de selección de respuesta. Se realizó un análisis de las fuentes de evidencia de validez del examen, usando el modelo conceptual de Messick, Kane y los Estándares de la AERA-APA-NCME, con la información generada en la aplicación de la prueba de febrero, 2019. En esta fecha 148,407 sustentantes respondieron la prueba para ingresar a las diversas licenciaturas que ofrece la Universidad. Se identificaron múltiples evidencias de validez de cinco fuentes: de contenido, de proceso de respuesta, de estructura interna, de relación con otras variables y de consecuencias del examen. La cantidad y calidad de estos datos revelan que el examen de ingreso de la UNAM cuenta con la suficiente evidencia de validez y confiabilidad para afirmar que el instrumento es sólido como herramienta de medición, y que evalúa el conocimiento de forma metodológicamente apropiada. El examen per se no resuelve la compleja red de necesidades de admisión a la educación superior, por lo que no es apropiado asignarle propiedades que van más allá de las limitaciones de un examen escrito a gran escala. Es necesario realizar investigaciones periódicas longitudinales sobre el uso del examen, ya que las condiciones sociales y educativas del contexto de la población de aspirantes son variables y dinámicas.

Palabras claves: Admisión a la universidad; evaluación sumativa; evaluación a gran escala; selección de estudiantes; validez.

Introducción

Las instituciones de educación superior cuentan con procesos de selección, mediante los cuales eligen cada año a las nuevas generaciones de alumnos que ingresarán a sus aulas. Los mecanismos que utilizan las universidades para operacionalizar el proceso de selección varían entre instituciones, dependiendo de factores como la normatividad nacional y local, el tamaño de la universidad, la población a quien está dirigida, su naturaleza pública o privada, entre otras características (Patterson et al, 2018; Trost, 1993). Estos procesos generalmente incluyen por lo menos un examen sumativo de alto impacto, dirigido principalmente a evaluar el conocimiento sobre las áreas del mismo relevantes a la carrera a que se pretende ingresar, y en ocasiones se suplementa con otros elementos como entrevistas, pruebas psicológicas, antecedentes académicos, actividades extracurriculares, exámenes por instancias externas a la universidad, entre otros (Patterson et al, 2018). Si bien la selección para ingresar a los planteles que ofertan educación superior tiene múltiples aristas sociológicas, éticas, económicas, humanas y políticas, la mayoría de las instituciones educativas privilegian los aspectos primordialmente académicos para elegir a sus estudiantes, por razones de índole práctico, tradición, factibilidad, y la gran cantidad de trabajos de investigación publicados que establece una correlación importante entre el desempeño académico antes de ingresar a la universidad con el desempeño durante la licenciatura (Juarros, 2006; Patterson et al, 2018).

Como consecuencia de la naturaleza de los exámenes estandarizados de alto impacto con carácter sumativo, la información obtenida de su aplicación y el análisis de sus resultados se mantienen generalmente en secreto como información reservada, y casi nunca se divulgan en la literatura académica con arbitraje. Las razones de esto son múltiples, desafortunadamente el resultado es una marcada ausencia de publicaciones que muestren con claridad y sustento metodológico el rigor académico de la elaboración del instrumento, así como del análisis de sus resultados (Sánchez Mendiola, 2016). Al revisar la literatura latinoamericana del tema, que es más relevante al contexto nacional de México, encontramos algunos estudios la mayoría de los cuales no presentan las evidencias de validez del uso del instrumento, probablemente por el supuesto de que al ser exámenes sumativos importantes, su proceso de elaboración y control de calidad es de alto nivel. En los tiempos actuales consideramos que es pertinente que la comunidad académica y la sociedad conozcan los atributos principales del instrumento, los fundamentos conceptuales y metodológicos que sustentan su elaboración, análisis y control de calidad, para identificar áreas de oportunidad de mejora que podrían atenderse. Esto es fundamental, ya que de los aspirantes que presentan el examen de admisión a la Universidad Nacional Autónoma de México (UNAM), ingresan menos del 10%, lo que lo convierte en uno de los exámenes más selectivos del país (Guzmán Gómez, 2011).

En este trabajo se analizó el proceso de elaboración, análisis y control de calidad del examen de ingreso a las licenciaturas de la UNAM, en su versión de febrero 2019, para generar información que pudiera servir como plataforma basal de seguimiento de este importante examen, así como mostrar aspectos técnico-metodológicos que sean de utilidad a las universidades de México y Latinoamérica para el desarrollo y

análisis de este tipo de evaluaciones sumativas. La dependencia universitaria actualmente a cargo de la elaboración del examen es la Coordinación de Desarrollo Educativo e Innovación Curricular (CODEIC) de la UNAM, a través de la Dirección de Evaluación Educativa (www.codeic.unam.mx). Las expectativas del estudio fueron que, al seguir un riguroso proceso de elaboración del examen, pudieran obtenerse evidencias de validez que ofrecieran un panorama amplio de la solidez de sus resultados.

Desarrollo

Los procesos de selección de aspirantes a la educación superior tienen diversos componentes y fases, que reflejan las prioridades y realidades de cada institución educativa en su contexto local y nacional (Patterson et al, 2018; Trost, 1993). Existen múltiples métodos e instrumentos de selección, con variados niveles de respaldo investigacional. En el caso de la UNAM se trata de un examen de conocimientos, que por su naturaleza sumativa de alto impacto, requiere ser explorado con el marco conceptual inherente a estas pruebas, como es el modelo del proceso de desarrollo y validación de exámenes de Haladyna y Downing (Lane et al, 2016). Este marco de desarrollo de exámenes objetivos es uno de los más utlizados en el mundo, se integra de 12 componentes (Cuadro 1), y se apoya en los Estándares para Pruebas Educativas y Psicológicas de la *American Educational Research Association, American Psychological Association y el National Council of Measurement in Education* (AERA, 2014).

Cuadro 1: Modelo del desarrollo de exámenes objetivos utilizado por la Dirección de Evaluación Educativa de la Universidad Nacional Autónoma de México (adaptado de Lane et al, 2016).

Componente 1. Plan general y global del examen

Componente 2. Definición del dominio y declaraciones que se harán de los resultados

Componente 3. Especificaciones del examen

Componente 4. Desarrollo de los ítems

Componente 5. Diseño y montaje del examen

Componente 6. Producción del examen

Componente 7. Aplicación del examen

Componente 8. Calificación del examen

Componente 9. Establecimiento de punto de pase

Componente 10. Reporte de resultados del examen

Componente 11. Seguridad del examen y banco de reactivos

Componente 12. Reporte técnico de la prueba

En el caso que nos ocupa, el Componente 9 (establecimiento de punto de pase) no se realiza ya que la interpretación del examen en nuestro contexto es de índole normativa, no criterial, debido principalmente al límite de espacios en la universidad. Esta característica institucional nos lleva a aceptar a los que obtengan las puntuaciones más altas en el examen, sin definir un estándar criterial o absoluto específico.

Desde 1997, la entonces Dirección General de Evaluación Educativa de la UNAM formalizó la planeación general, la definición del contenido y las especificaciones de la prueba de admisión a la Universidad, atendiendo al Reglamento General de Inscripciones de la Universidad: "Para ingresar a la Universidad es indispensable ser aceptado mediante concurso de selección, que comprenderá una **prueba escrita** y que deberá realizarse dentro de los periodos que al efecto se señalen" (https://www.dgae-siae.unam.mx/acerca/normatividad.html#leg-3). Para ello se desarrolló una metodología para diseñar el examen, atendiendo las buenas prácticas de elaboración de exámenes objetivos (AERA, 2014; Haladyna et al, 2002), se muestra en la Figura 1.

Figura 1: Metodología del diseño de los exámenes de ingreso de la Universidad Nacional Autónoma de México (Dirección de Evaluación Educativa, CODEIC, UNAM).

El concepto más más importante para que los resultados de los procesos de evaluación tengan sustento sólido y uso apropiado, es el de validez. La validez de un proceso de evaluación es el grado con el que mide lo que se supone que mide, actualmente se le concibe como un concepto unitario y holístico, en el

que toda la validez es validez de constructo (AERA, 2014; Downing, 2003; Kane, 2013). Para efectos de este trabajo, utilizamos el concepto moderno de validez, como un juicio valorativo holístico e integrador que requiere múltiples fuentes de evidencia para la interpretación del constructo evaluado, ya que intenta responder a la pregunta *"¿qué inferencias pueden hacerse sobre la persona basándose en los resultados del examen?"* (Downing, 2003; Mendoza Ramos, 2015). No es el examen el que es válido per se, ya que la validez de un examen es específica para un propósito y se refiere más bien a lo apropiado de la interpretación de los resultados. Las cinco fuentes importantes de validez en evaluación del aprendizaje son: contenido, procesos de respuesta, estructura interna (que incluye la confiabilidad y el comportamiento estadístico de los reactivos), relación con otras variables y consecuencias (AERA, 2014; Downing, 2003).

Se analizó toda la información del examen que atendiera a cada una de estas fuentes, a continuación se describen de forma resumida:

1) Contenido. El contenido del examen se fundamentó en los planes de estudio de la educación media superior. Se estableció un perfil de referencia por cuerpos colegiados universitarios. Comisiones de profesores del bachillerato de la UNAM, expertos en contenido, revisaron los temarios y determinaron los temas y niveles cognitivos a evaluar. Se elaboró una tabla de especificaciones con los resultados de aprendizaje esperados y se ponderaron las áreas del conocimiento a explorar. Los académicos elaboradores de reactivos fueron entrenados para elaborar preguntas de opción múltiple de características técnicas apropiadas. La estructura del examen se muestra en el Cuadro 2, integrándose con 120 reactivos.

Cuadro 2: Número de reactivos del examen de ingreso a las licenciaturas de la UNAM, por área del conocimiento (CFMI=Ciencias Físico Matemáticas e Ingenierías; CBQS=Ciencias Biológicas, Químicas y de la Salud; CS=Ciencias Sociales; HyA=Humanidades y Artes).

Materia	CFMI	CBQS	CS	H y A
Matemáticas	26	24	24	22
Física	16	12	10	10
Química	10	13	10	10
Biología	10	13	10	10
H. universal	10	10	14	10
H. de México	10	10	14	10
Literatura	10	10	10	10
Geografía	10	10	10	10
Español	18	18	18	18
Filosofía	-	-	-	10
Total	**120**	**120**	**120**	**120**

2) **Procesos de respuesta.** Evidencia de integridad de los datos de manera que las fuentes de error que se pueden asociar con la administración del examen han sido controladas en la medida de lo posible. Uno de ellos es la familiaridad del estudiante con el formato de preguntas de opción múltiple, lo cual se cumple en la actualidad. Al ser con lápiz y papel no introduce la variable de la habilidad en el uso de computadoras. Los reactivos tienen cuatro opciones de respuesta, una de las cuales es la correcta. Cada reactivo es revisado por personal técnico para verificar congruencia, relación con el resultado de aprendizaje y estructura gramatical. Se efectúa la validación de la clave de respuestas, así como el control de calidad del reporte de resultados.

Hemos desarrollado una plataforma informática para el desarrollo y validación del examen, que tiene más de una década de perfeccionamiento e integración, denominada "Sistema Integral de Gestión de Exámenes" (SIGE), lo que proporciona un elemento más al desarrollo del examen, la validación de los reactivos, y la integración de un banco de reactivos con características apropiadas a la tarea. En el SIGE se capturan y validan los reactivos por tres expertos en contenido, y transitan por el proceso de corrección de estilo, inclusión de los resultados de aprendizaje, entre otros aspectos técnicos. En el sistema se captura el historial del desempeño psicométrico del reactivo.

3) **Estructura interna.** Se refiere a las características estadísticas del examen, como el análisis de reactivos, el funcionamiento de los distractores, la confiabilidad del examen, entre otros (De Champlain, 2010). En el caso del examen de la UNAM se realiza el análisis psicométrico con la Teoría Clásica de los Tests (TCT) con el software Iteman versión 4, y con el modelo de la Teoría de Respuesta al Ítem (TRI) de tres parámetros, con el programa BILOG. En las Tablas 1 a 3 se presentan algunos resultados del examen.

Tabla 1: Número y porcentaje de aspirantes por área del conocimiento (CFMI=Ciencias Físico Matemáticas e Ingenierías; CBQS=Ciencias Biológicas, Químicas y de la Salud; CS=Ciencias Sociales; HyA=Humanidades y Artes)

Área	N	%
CFMI	30,561	20.6
CBQS	56,474	38.1
CS	45,229	30.5
HyA	16,143	10.9
Total	148,407	100.0

Tabla 2: Número y porcentaje de aspirantes por sexo y área de conocimiento (CFMI=Ciencias Físico Matemáticas e Ingenierías; CBQS=Ciencias Biológicas, Químicas y de la Salud; CS=Ciencias Sociales; HyA=Humanidades y Artes) (n=148,407).

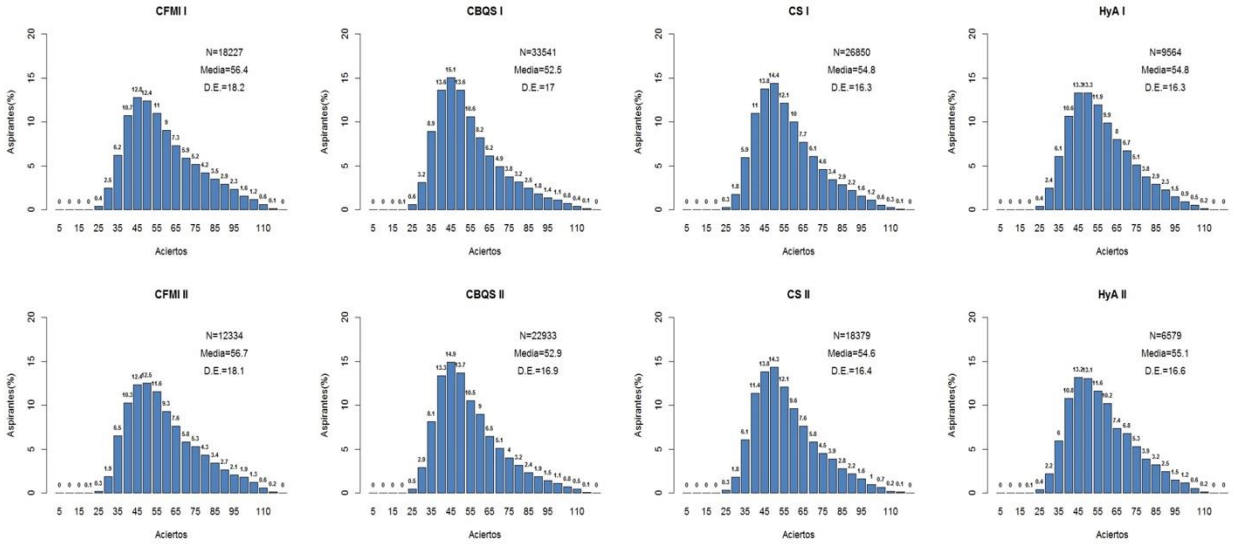
	Área				
	CFMI	CBQS	CS	HyA	Total
Hombres	21,486	18,082	20,723	5,638	65,929
%	32.6%	27.4%	31.4%	8.6%	100.0%
% en el área	70.3%	32.0%	45.8%	34.9%	**44.4%**
Mujeres	9,075	38,392	24,506	10,505	82,478
%	11.0%	46.5%	29.7%	12.7%	100.0%
% en el área	29.7%	68.0%	54.2%	65.1%	**55.6%**
Total	30,561	56,474	45,229	16,143	148,407
%	20.6%	38.1%	30.5%	10.9%	100.0%

Tabla 3: Análisis psicométrico con la Teoría Clásica de los Test del examen de ingreso a la licenciatura de la UNAM, por campo de conocimiento y versión. (EEM=error estándar de medición) (n=148,407).

	CFMI		CBQS		CS		HyA	
	I	II	I	II	I	II	I	II
N	18,227	12,334	33,541	22,933	26,850	18,379	9,564	6,579
Promedio aciertos	55.4	55.7	52.5	52.9	54.8	54.6	54.8	55.1
Desv. estándar	18.2	18.1	17.0	16.9	16.3	16.4	16.3	16.6
Mediana	52	52	49	49	52	51	52	52
EEM	4.92	4.92	4.98	5.0	4.97	4.98	4.99	4.97
Dificultad media	0.466	0.468	0.437	0.441	0.456	0.455	0.456	0.460
Correlación punto biserial media	0.298	0.296	0.273	0.271	0.260	0.263	0.258	0.264
Alfa de Cronbach	**0.927**	**0.926**	**0.914**	**0.913**	**0.907**	**0.908**	**0.906**	**0.910**

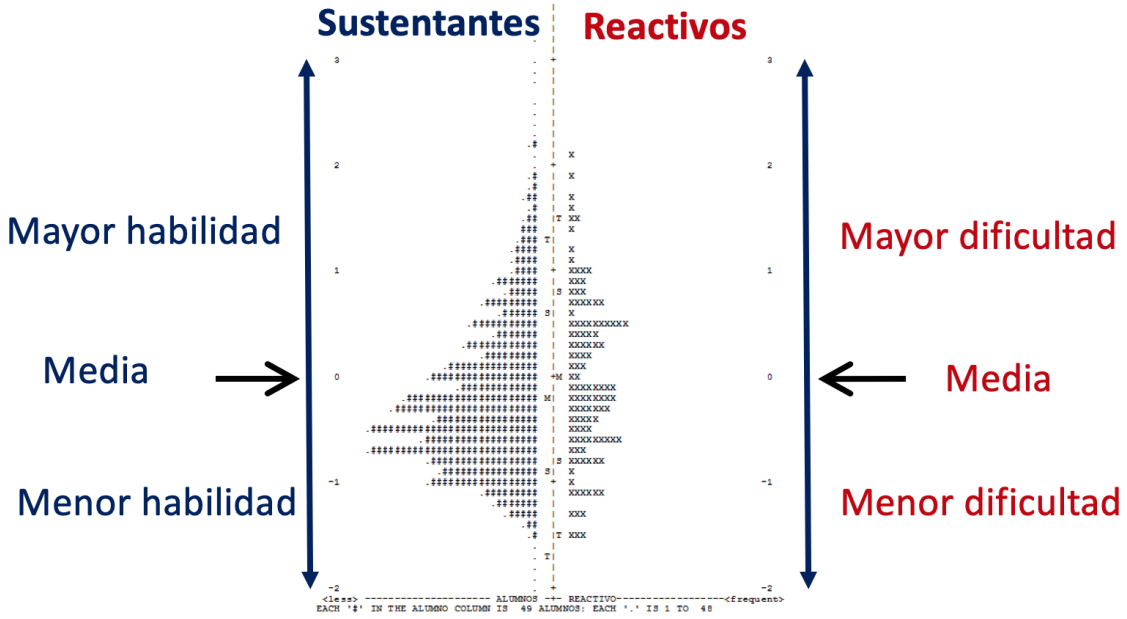
En la Figura 2 se muestra la distribución de aciertos por área del conocimiento y versión del examen.

Figura 2: Distribución de aciertos en el examen de admisión a la licenciatura de la UNAM, por área del conocimiento y versión (n=148,407).



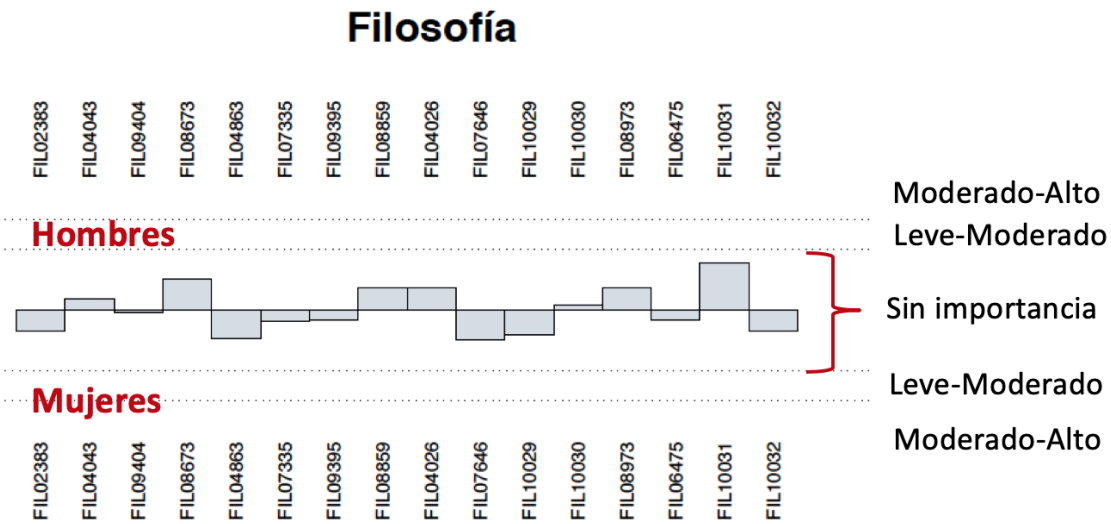
Todos los reactivos que se utilizan en el examen cubren criterios psicométricos estrictos, con análisis psicométrico de TCT, como con TRI. En la Figura 3 podemos ver los resultados del mapa de dificultad de los reactivos comparado con la habilidad de los aspirantes, en el área de las Ciencias Físico Matemáticas e Ingenierías, utilizando el modelo de Rasch de la TRI (Baker, 2001). Los resultados en las otras tres áreas del conocimiento presentaron patrones similares, lo que arroja evidencia de validez sobre lo apropiado de la dificultad del examen para la habilidad de los aspirantes. Es importante notar que todos los sustentantes están incluidos en el rango de dificultad del instrumento, y que de manera similar a la distribución estadística del número de aciertos en la Figura 2, hay más estudiantes hacia el extremo de menor habilidad.

Figura 3: Mapa de dificultad de los reactivos y habilidad de los estudiantes, con el Modelo de Rasch (n=148,407).



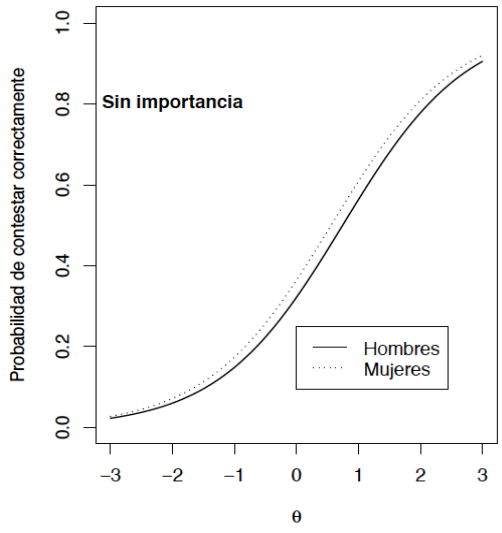
A partir de 2018, comenzamos a realizar análisis diferencial de los ítems (DIF, por sus siglas en inglés), para explorar el comportamiento del examen y los reactivos por sexo, tema que ha sido sujeto de constante debate (Guzmán, 2011). Un reactivo presenta DIF cuando los examinados de un mismo nivel de habilidad, pero provenientes de diferentes grupos, tienen una probabilidad distinta de contestarlo correctamente (Walker, 2011; Zieky, 1993) . Se empleó la técnica de TRI basada en el modelo de Rasch, donde los grupos de interés fueron las mujeres y los hombres. Si los contrastes en los niveles de dificultad de un reactivo entre los grupos de interés no supera los 0.43 lógitos, DIF = sin importancia; > 0.64 lógitos, DIF = moderado-alto; encontrarse entre esos valores DIF = leve-moderado. Hemos encontrado muy pocos reactivos con DIF leve-moderado, los cuales son valorados por un cuerpo colegiado para analizar la lógica del reactivo y el resultado de aprendizaje explorado, y determinar su potencial efecto en los resultados. En la Figura 4 podemos ver un ejemplo de los reactivos del área de filosofía, en los que no se encontraron reactivos con DIF en cuanto a sexo.

Figura 4: DIF (funcionamiento diferencial de los ítems) de acuerdo al sexo en los reactivos de filosofia, del examen de admisión a la licenciatura de la UNAM, febrero de 2019.



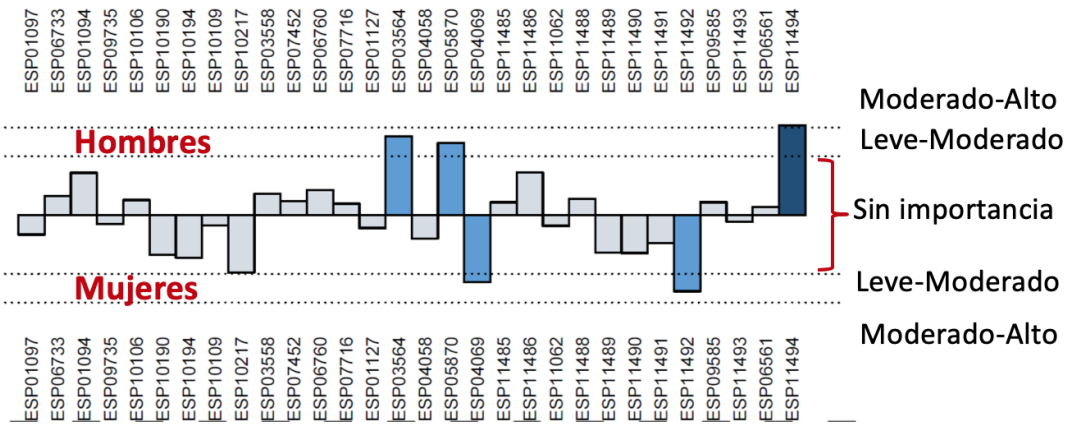
En la Figura 5 podemos observar la curva de un reactivo con el modelo de Rasch, con DIF mínimo o intrascendente.

Figura 5: Visualización de DIF (funcionamiento diferencial de ítem) mínimo por sexo en un reactivo de matemáticas, examen de admisión a la licenciatura de la UNAM.



En la Figura 6 podemos visualizar el DIF por sexo de los reactivos de Español, del área de Ciencias Biológicas, Químicas y de la Salud. Muy pocos tienen DIF leve-moderado, por lo que fueron evaluados para determinar el potencial impacto en el examen. Como ocurre en exámenes de este tipo, la dirección de algunos reactivos con DIF leve para hombres se cancelan con los reactivos con DIF leve para mujeres. Es importante destacar que estas cifras son solamente un elemento estadístico, que de ninguna manera por sí solo documentan sesgo a favor de una población, sino que, como cualquier dato estadístico, deben evaluarse cualitativamente por un grupo de expertos en contenido y en evaluación, para analizar su potencial efecto en los resultados.

Figura 6: DIF por sexo de los reactivos de Español, del área de Ciencias Biológicas, Químicas y de la Salud, en el examen de admisión a la licenciatura de la UNAM, febrero de 2019.



4) **Relación con otras variables:** La relación de los resultados del examen con otras variables se refiere a la correlación estadística entre los resultados obtenidos en el mismo con otra medición de características conocidas. El examen de ingreso a la UNAM se asocia con los resultados de los exámenes de diagnóstico de conocimientos que se aplica a todos los estudiantes que ingresan a la institución ($r = 0.64$, $p<0.01$). Las correlaciones por área del conocimiento son similares.

5) **Consecuencias:** Se refiere al impacto en los estudiantes de las puntuaciones de la evaluación, de las decisiones que se toman como resultado del examen, y su efecto en la enseñanza y el aprendizaje. Por ejemplo: el método de establecimiento del punto de corte, las consecuencias para el estudiante y la sociedad, las consecuencias para los profesores y las instituciones educativas. En este apartado no contamos con fuentes de evidencia de validez propias, hay un espacio de oportunidad amplio para realizar estudios de los costos económicos y emocionales, costos sociales de falsos positivos y falsos negativos, entre otros aspectos.

Conclusiones

Validez implica una aproximación científica a la interpretación de los resultados de los exámenes, es decir, probar hipótesis sobre los conceptos evaluados en el examen. La información proporcionada por un instrumento de evaluación no es válida o inválida, sino que los resultados del examen tienen más o menos evidencia de las diferentes fuentes para apoyar o rechazar una interpretación específica (por ejemplo pasar o reprobar un curso, admitir o no a un estudiante en la universidad) (Downing, 2003; Kane, 2013). Las organizaciones que elaboran e implementan el examen (entidades gubernamentales, instituciones educativas) son los candidatos obvios para validar las afirmaciones que hacen sobre la interpretación de los resultados de un examen, ya que generalmente son quienes tienen los elementos y recursos para hacerlo (AERA, 2014). Quienes elaboramos exámenes tenemos la obligación ética y el imperativo educativo de documentar qué tan defendible es la interpretación de los resultados, en beneficio de los estudiantes y de la sociedad en general.

Los resultados descritos establecen una cantidad importante de evidencias de validez para el uso del examen como instrumento de evaluación del conocimiento, con validez y confiabilidad similares a las buenas prácticas de exámenes sumativos a gran escala, a nivel internacional. El examen de ingreso es solo un instrumento de medición del conocimiento, nada más pero tampoco nada menos. El uso de los resultados y las inferencias que se hagan de los mismos es un tema extraordinariamente complejo, con diversas aristas que deben discutirse de manera rigurosa y reflexiva. En las últimas décadas las principales organizaciones de evaluación educativa del mundo han hecho énfasis en la necesidad de justicia y equidad en el proceso educativo, incluyendo la evaluación del aprendizaje, para ser congruentes con el sentido social de la educación (AERA, 2014). Existe controversia sobre el tema, ya que los exámenes estandarizados en gran escala, que por necesidad se aplican y analizan en contextos altamente controlados para que cada estudiante se enfrente al mismo reto en igualdad de condiciones, por definición tratan a todos los estudiantes de la misma manera. Esta tensión entre lo ideal educativamente y lo real, continúa sin resolverse (Márquez Jiménez, 2014; Martínez Rizo, 2001; Sánchez Mendiola et al, 2015).

La comprensión clara del concepto moderno de validez es fundamental para entender las limitaciones de los resultados de los exámenes, ya que extrapolar conclusiones y decisiones más allá de lo académicamente sensato es inapropiado e incluso puede ser peligroso. Si un estudiante tiene un desempeño deficiente en una aplicación de un examen sumativo de alto impacto, eso no significa que sea “mala persona” o “incompetente”, calificativos que se asignan como etiquetas y que tienen un impacto emocional importante.

La responsabilidad de realizar buenos exámenes e informar a la sociedad sobre sus limitaciones recae en nuestras organizaciones y grupos de expertos, en colaboración con las autoridades y los medios de comunicación. La asimetría de poder intrínseca en los procesos de evaluación conlleva una enorme responsabilidad de las autoridades académicas e institucionales.

Referencias

American Educational Research Association, American Psychological Association, National Council on Measurement in Education, & Joint Committee on Standards for Educational and Psychological Testing. (2014). Standards for educational and psychological testing. Washington, DC: AERA.

Baker, F.B. (2001). The Basics of Item Response Theory. 2nd Ed. USA: ERIC Clearinghouse on Assessment and Evaluation. 1-896.

De Champlain, A.F. (2010). A primer on classical test theory and item response theory for assessments in medical education. Med Educ, 44(1), 109-117.

Downing, S.M. (2003). Validity: on the meaningful interpretation of assessment data. Med Educ, 37, 830-837.

Guzmán Gómez, Carlota, & Serrano Sánchez, Olga Victoria. (2011). Las puertas del ingreso a la educación superior: el caso del concurso de selección a la licenciatura de la UNAM. Revista de la Educación Superior, 40(157), 31-53.

Haladyna, T.M., Downing, S.M. & Rodriguez, M.C. (2002). A Review of Multiple-Choice Item-Writing Guidelines for Classroom Assessment. Appl Meas Educ, 15, 309-334.

Juarros, María Fernanda. (2006). ¿Educación superior como derecho o como privilegio?: Las políticas de admisión a la universidad en el contexto de los países de la región. Andamios, 3(5), 69-90.

Kane, M.T. (2013). Validating the interpretations and uses of test scores. J Educ Meas, 50(1), 1-73.

Lane, S., Raymond, M.R., Haladyna, T.M., & Downing, S.M. (2016). Test Development Process. En: Lane, S., Raymond, M.R. & Haladyna, T.M. (Eds). Handbook of Test Development. 2nd Ed. New York: Routledge.

Márquez Jiménez, A. (2014). Las pruebas estandarizadas en entredicho. Perfiles Educativos, 36(144),3-9.

Martínez Rizo, F. (2001). Evaluación educativa y pruebas estandarizadas. Elementos para enriquecer el debate. Revista de la Educación Superior, 30(120), 71-85.

Mendoza Ramos, A. (2015). La validez en los exámenes de alto impacto: Un enfoque desde la lógica argumentativa. Perfiles Educativos, 37(149), 169-186.

Patterson, F., Roberts, C., Hanson, M.D., Hampe, W., Eva, K., Ponnamperuma, G., Magzoub, M., Tekian, A. & Cleland, J. (2018). 2018 Ottawa consensus statement: Selection and recruitment to the healthcare professions. Med Teach, 40(11), 1091-1101.

Sánchez-Mendiola, M., Delgado-Maldonado, L. (2017). Exámenes de alto impacto: Implicaciones educativas. Inv Ed Med 6(21):52-62. Sánchez Mendiola, M., Delgado Maldonado, L., Flores Hernández, F., Leenen, I., Martínez González, A. (2015). Evaluación del aprendizaje. En: Sánchez Mendiola, M., Lifshitz Guinzberg, A., Vilar Puig, P., Martínez González, A., Varela Ruiz, M., Graue Wiechers, E. Eds. Educación Médica: Teoría y Práctica. Editorial Elsevier. México D.F. Cap. 14, pág. 89-95.

Trost, G. (1993). Principios y prácticas en la selección para la admisión a la Educación superior. Revista de la Educación Superior, 22(85), 1-10. Disponible en: http://publicaciones.anuies.mx/pdfs/revista/Revista85_S2A5ES.pdf

Walker, C. (2011). What's the DIF? Why differential item functioning analyses are an important part of instrument development and validation. Journal of Psychoeducational Assessment, 29, 364-376.

Zieky, M. (1993). DIF statistics in test development. En: Holland, P.W., & Wainer, H. (Eds). *Differential item functioning.* (pp. 337–347). Hillsdale, N.J.; Erlbaum.