Revista Iberoamericana de Evaluación Educativa, 2020, 13(2), 107-128.
https://doi.org/10.15366/riee2020.13.2.006

El Examen de Ingreso a la Universidad Nacional Autónoma de México: Evidencias de Validez de una Prueba de Alto Impacto y Gran Escala

The Admission Exam to the National Autonomous University of Mexico: Validity Evidence of a Large Scale High-Stakes Test

Melchor Sánchez Mendiola *
Manuel García Minjares
Adrián Martínez González
Enrique Buzo Casanova

Universidad Nacional Autónoma de México, México.

Introducción**.** Los exámenes de admisión a la educación superior son evaluaciones sumativas de alto impacto para los aspirantes, por lo que requieren evidencia de validez para que las inferencias que se hagan de los resultados sean apropiadas. La Universidad Nacional Autónoma de México (UNAM) es la institución de educación superior más solicitada del país, anualmente ingresan menos del 10% de los aspirantes por examen de selección. Métodos**.** Se realizó un análisis de las fuentes de evidencia de validez del examen, con el modelo conceptual de Messick, Kane y los Estándares de la AERA-APA-NCME, con la información generada de la aplicación de febrero 2019 a 148.407 sustentantes. Resultados**:** Se identificaron evidencias de validez de contenido, proceso de respuesta, estructura interna, relación con otras variables y consecuencias del examen. Los resultados revelan que el examen de ingreso tiene suficiente evidencia de validez para afirmar que es sólido como herramienta de medición del conocimiento. Discusión**.** Por su relevancia social, es fundamental que las instituciones que usan este tipo de instrumentos documenten sus evidencias de validez. 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 dinámicas.

Palabras clave: Condiciones de admisión; Evaluación sumativa; Prueba de respuesta múltiple; Selección de estudiantes; Validez.

Introduction**.** Higher education institutions' admission exams are summative high-stakes tests that have important consequences for applicants, so they require validity evidence to assure that appropriate inferences are made with the results. The National Autonomous University of Mexico (UNAM) is the most sought-after higher education institution in the country, annually less than 10% of applicants that take the test are admitted. Methods**.** Analysis of the sources of the test validity evidence was performed using Messick and Kane conceptual frameworks, as well as the AERA-APA-NCME Standards, with the information generated from the February 2019 admission test in 148.407 applicants. Results**:** Test validity evidence was identified from content, response process, internal structure, relationship with other variables and consequences. Results suggest that the test has enough validity evidence, to state that the instrument is robust as a technical tool for knowledge assessment and as a source of information for high-stakes decisions. Discussion**.** It is crucial that institutions that use these tools document their validity evidence, since they have great social relevance. It is necessary to perform periodic longitudinal studies about the test use and its implications, since social and educational conditions in the context of the applicant population are dynamic.

Keywords: Higher education admission; Summative assessment; Multiple choice test; Student selection; Validity.

*Contacto: melchorsm@unam.mx
issn: 1989-0397
www.rinace.net/riee/
https://revistas.uam.es/riee

Recibido: 4 de mayo de 2020
1ª Evaluación: 30 de junio de 2020
2ª Evaluación: 13 de julio de 2020
Aceptado: 21 de julio de 2020

1. Introducción

El presente trabajo ofrece una descripción del examen de admisión a las licenciaturas de la Universidad Nacional Autónoma de México (UNAM), la universidad más grande del país (Ordorika, Rodríguez y Montes de Oca, 2013) y la que recibe mayor cantidad de solicitudes de ingreso a nivel nacional (ANUIES, 2019). En concordancia con la importancia del proceso de selección para la institución y para la educación superior nacional, es menester someter a escrutinio los diversos elementos que constituyen el componente central del proceso, el examen escrito de conocimientos. A través del lente más importante en evaluación educativa, el de la validez (AERA, 2014; Kane, 2016; Shepard, 2016), en este trabajo se describen diversos atributos del instrumento, su diseño y los resultados de su aplicación, para identificar fortalezas y áreas de oportunidad en el instrumento mismo y su rol en el proceso de ingreso.

2. Fundamentación teórica

El marco teórico que sustenta este trabajo guarda relación con la naturaleza de los procesos de admisión a la educación superior, así como con la validez de las evaluaciones sumativas realizadas para este fin.

2.1. Procesos de admisión en educación superior a nivel internacional y nacional

Las universidades que cuentan con procesos de selección, mediante los cuales analizan diversos elementos de los aspirantes para elegir periódicamente a las nuevas cohortes que ingresarán a sus espacios educativos, se enfrentan a un reto complejo y difícil. Los mecanismos específicos que utilizan las instituciones de educación superior para operacionalizar el proceso de selección varían entre países y universidades, dependiendo de varios factores: normatividad local, regional y nacional, carácter público o privado de la universidad, tamaño de la institución, población a quienes está dirigida, entre otros atributos (Manzi et al., 2010; Patterson et al., 2018; Trost, 1993).

Estos procesos frecuentemente incluyen un examen sumativo de alto impacto, dirigido principalmente a evaluar el conocimiento sobre las áreas relevantes a la carrera que se pretende ingresar. 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; Trost, 1993; Zwick, 2006). Si bien la selección para ingresar a los planteles que ofertan educación superior tiene múltiples aristas sociales, éticas, económicas, humanas, políticas y afectivas, la mayoría de las instituciones educativas privilegian aspectos primordialmente académicos para elegir a sus estudiantes, por razones de índole práctico, tradición, factibilidad, y la evidencia publicada que establece una correlación importante entre el desempeño académico antes de ingresar con el desempeño durante la licenciatura (Frey y Detterman, 2003; Juarros, 2006; Manzi et al., 2010; Patterson et al., 2018; Sigal y Dávila, 2004).

En un mundo ideal, el proceso de admisión a la educación superior incorporaría todos los elementos disponibles, realizaría procesos libres de sesgos, evaluaciones que exploraran las dimensiones más importantes de cada uno de los aspirantes, con jueces imparciales que balancearan la información obtenida de manera integral, equitativa y ponderada, para así seleccionar a los "mejores" candidatos. Desafortunadamente este proceso ideal no ocurre en el mundo real. No hay cabida en las instituciones de educación superior para toda la

población, por lo que las universidades y los gobiernos se ven obligados a diseñar complejos esquemas de selección que sean aceptados por la sociedad, la institución y los aspirantes, lo que hace inevitable que muchos aspirantes queden fuera de su primera elección (OCDE, 2018).

En algunos países existen exámenes nacionales estandarizados que se utilizan en el proceso de selección de las universidades, como el ACT y el SAT en los Estados Unidos, que proveen un elemento común de decisión a los organismos responsables del ingreso a las universidades (Frey y Detterman, 2003). En el caso de los Estados Unidos, el *Educational Testing Service* es una organización grande sin fines de lucro que se encarga de desarrollar exámenes estandarizados, con expertos que generan investigación original sobre el desarrollo, aplicación e interpretación de exámenes (Bennett, 2005). Estas organizaciones publican investigación original sobre sus instrumentos de medición en la literatura internacional, lo que tiene varios efectos: contribuye al conocimiento en evaluación educativa, legitima el uso de los instrumentos ante la comunidad académica y la sociedad, y genera una plataforma de evidencia de validez y confiabilidad de sus exámenes a partir de la cual pueden mejorarlos.

En el caso de nuestro país contamos con organizaciones similares (como el CENEVAL), que, si bien tienen expertos en evaluación educativa, se dedican principalmente a proveer un servicio y en menor medida a publicar en la literatura académica con arbitraje por pares trabajos de investigación sobre el tema, específicamente de la evidencia de validez de sus instrumentos (Gago, 2000). Las organizaciones de este tipo en Latinoamérica publican una abundante cantidad de manuales, reportes e informes externos, aunque las publicaciones sobre las evidencias de validez de sus instrumentos son pocas y se infiere que la validez de los mismos está documentada a través de mecanismos internos de control de calidad y los reportes técnicos correspondientes.

2.2. Naturaleza sumativa de los procesos de admisión y sus instrumentos

Los exámenes de admisión a las universidades se consideran evaluaciones sumativas de alto impacto o de altas consecuencias, ya que tienen potencial de generar efectos importantes en las personas que los toman (Cizek, 2001; Lane et al., 2016; Sánchez-Mendiola y Delgado-Maldonado, 2017). Estos efectos son económicos, sociales, educativos, e incluso en la salud física y mental de los sustentantes y sus familiares. También generan consecuencias inesperadas, positivas y negativas, lo que nos obliga a analizarlos con rigor. Por su naturaleza sumativa, la información obtenida de su aplicación y el análisis de sus resultados se mantienen en secreto como información reservada, pocas veces se divulgan en la literatura académica, y los reportes que reciben los sustentantes son escuetos y en ocasiones solo se les informa si acreditan o no el examen. El resultado es una ausencia de publicaciones que muestren con claridad y sustento metodológico el rigor académico de la elaboración de los instrumentos, así como del análisis de sus resultados, por lo que persiste la controversia sobre su utilidad real y sus implicaciones educativas (Martínez-Rizo, 2001; Sánchez-Mendiola y Delgado-Maldonado, 2017; Zwick, 2006).

2.3. La Universidad Nacional Autónoma de México: la encrucijada del proceso de ingreso

Al revisar la literatura latinoamericana sobre exámenes de admisión a la universidad, encontramos pocos estudios sobre las características psicométricas y evidencias de validez del uso de instrumentos de admisión (Backhoff, Tirado y Larrazolo, 2001; Buendía y

Rivera, 2010). Hasta la fecha no se ha publicado un análisis del examen de ingreso a las licenciaturas de la UNAM, a pesar de que es uno de los exámenes sumativos de alto impacto más importantes del país (Sánchez-Mendiola, 2017). Consideramos que es pertinente que la comunidad académica conozca 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. 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 y Serrano, 2011).

2.4. Evolución del concepto de validez y su valor como lente de análisis

Para que los resultados de los procesos de evaluación tengan un robusto sustento y se utilicen de forma apropiada, es indispensable abordarlos desde la lente de la validez. Validez de un proceso de evaluación es el grado con el que mide lo que se supone que mide, tradicionalmente se clasificaba como las tres C: de contenido, de criterio y de constructo (Buntis, Buntis y Eggert, 2017; Sánchez-Mendiola, 2015; Young et al., 2018). En las últimas décadas el concepto ha evolucionado a una definición más amplia (AERA, 2014; Gregory, 2016; Kane, 2016; Kane y Bridgeman, 2017; Shepard, 2016). Actualmente se considera que se trata de un juicio valorativo holístico, en el que toda la validez es validez de constructo que se alimenta de diferentes fuentes. El concepto intenta responder a la pregunta: ¿qué inferencias pueden hacerse sobre la persona basándose en los resultados del examen? (Kane, 2016; Mendoza, 2015). No es el instrumento 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. En otras palabras, la validez no es una propiedad intrínseca del examen, sino del significado de los resultados en el entorno educativo específico y las inferencias que pueden hacerse de los mismos, por lo que el término "el instrumento es válido" es incorrecto (Kane, 2016; Sánchez, 2015).

Este modelo se ha construido a partir de las aportaciones metodológicas y conceptuales de varios autores como Messick y Kane, y si bien no está exento de controversia, es el actualmente aceptado por las principales organizaciones de evaluación educativa del mundo (AERA, 2014; Kane, 2016; Shepard, 2016).

3. Objetivo del estudio

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 mostrar aspectos técnico-metodológicos que puedan ser de utilidad para el desarrollo y análisis de este tipo de evaluaciones. Las expectativas del estudio fueron que, al seguir el proceso de elaboración del examen, pudieran obtenerse evidencias de validez que ofrecieran un panorama amplio de sus resultados.

4. Material y método

Los diferentes elementos metodológicos que se utilizaron en el estudio se describen a continuación.

4.1 Contexto

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 (OCDE, 2018). En el caso de la UNAM se trata de un examen escrito de conocimientos. Desde 1997, la Dirección General de Evaluación Educativa de la UNAM formalizó la planeación general, definición del contenido y especificaciones de la prueba de admisión, atendiendo al Reglamento General de Inscripciones: *"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"* (UNAM, 1997). La dependencia 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 (Graue, 2018). La Dirección General de Administración Escolar expide las convocatorias para Concurso de Selección en febrero y junio de cada año, para aspirantes a ingresar al nivel Licenciatura en el Sistema Escolarizado y en el Sistema Universidad Abierta y Educación a Distancia (SUAyED) -modalidades Abierta y a Distancia-. Esta dependencia se encarga de la administración y logística del examen en las diversas sedes en que se aplica (DGAE, 2020).

4.2 Diseño de investigación y marco conceptual

Utilizamos el modelo de "la brújula de la investigación en educación" de Ringsted, Hodges y Scherpbier (2011), basado en los estudios de clasificación de investigación educativa de Cook, Bordage y Schmidt (2008). El centro de este modelo es un marco conceptual teórico, que en este estudio es el modelo de validez de Messick y Kane (Kane, 2016), adoptado por la *American Educational Research Association*, *American Psychological Association* y el *National Council of Measurement in Education* (AERA, 2014). En el modelo de la "brújula" de Ringsted hay cuatro categorías de estudios en educación, nuestro estudio encaja en la categoría de estudios exploratorios, enfocados en identificar y explicar fenómenos y sus relaciones, dentro del subtipo de estudios psicométricos que pretenden establecer evidencia de validez y confiabilidad de instrumentos de medición educativa (Ringsted, Hodges y Scherpbier, 2011).

4.3 Metodología de elaboración del instrumento

El marco conceptual de exámenes objetivos que utilizamos 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 utilizados en el mundo, se integra de 12 componentes y se apoya en los Estándares para Pruebas Educativas y Psicológicas de la AERA-APA-NCME (Lane et al., 2016):

- Componente 1. Plan general y global del examen
- Componente 2. Definición del dominio y declaraciones que se harán sobre 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 del examen de la UNAM, el Componente 9 (establecimiento de punto de pase) no se realiza ya que la interpretación del examen es de índole normativa, no criterial, está sujeto principalmente al límite de espacios en la universidad en virtud de la excesiva demanda de aspirantes.

El examen es un instrumento de evaluación del conocimiento compuesto de reactivos de selección de respuesta con cuatro opciones, una de ellas correcta. La secuencia de desarrollo del instrumento se muestra en la figura 1 (AERA, 2014; Haladyna, Downing y Rodriguez, 2002; Lane et al., 2016).

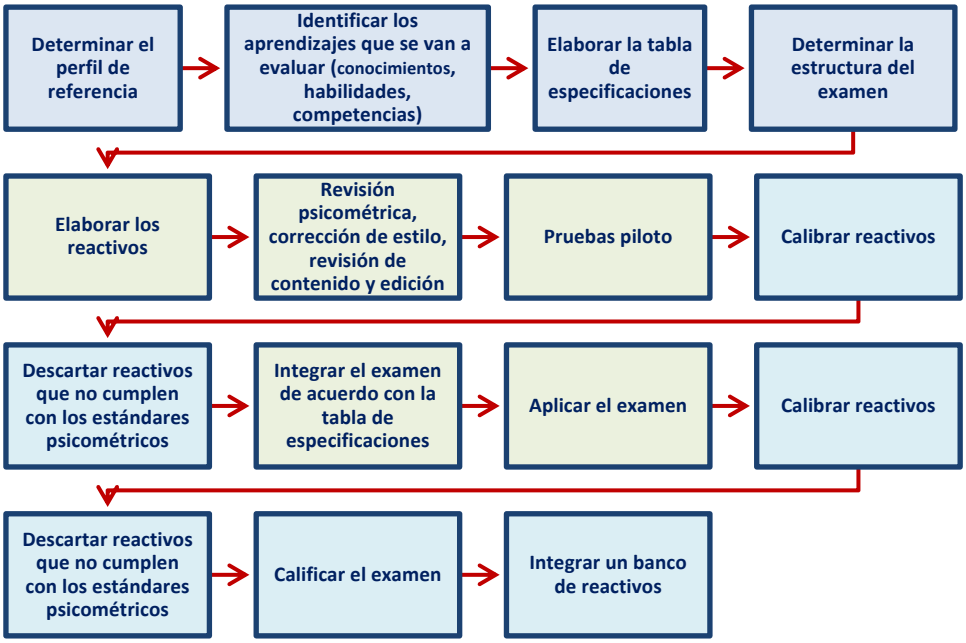

Figura 1. Metodología del diseño de exámenes de ingreso de la Universidad Nacional Autónoma de México
Fuente: Dirección de Evaluación Educativa de la CODEIC. UNAM.

En el modelo actual de validez, existen cinco fuentes importantes de la misma: 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).

4.4 Análisis psicométrico de reactivos

Se realizó el análisis psicométrico de reactivos con los modelos de la teoría de medición clásica (TMC) y de la teoría de respuesta al ítem (TRI) de uno, dos y tres parámetros (Andrich y Marais, 2019; Raykov y Marcoulides, 2016). El análisis con la TMC se realizó con el programa Iteman versiones 3.5 y 4 para el modelo de un parámetro, y BILOG-MG 3 para los modelos de dos y tres parámetros. El análisis de Rasch (Andrich y Marais, 2019; Boone y Noltemeyer, 2017) con el modelo de un parámetro especifica que la

probabilidad de que un examinado i con habilidad θ_i genere una respuesta x_{ij} al reactivo j con una dificultad b_j de acuerdo con la siguiente fórmula:

$$P(x_{ij}|\theta_i, b_j) = \frac{\exp(\theta_i - b_j)}{1 + \exp(\theta_i - b_j)}$$

Donde:

$x_{ij} = Respuesta\ del\ i - ésimo\ examinado\ al\ j$

$- ésimo\ reactivo\ (1\ si\ es\ correcto; 0\ en\ otro\ caso).$

$\theta_i = Habilidad\ del\ i - ésimo\ examinado.$

$b_j = Dificultad\ del\ j - ésimo\ reactivo.$

En el modelo de TRI de dos parámetros se usó la siguiente fórmula:

$$P(x_{ij}|\theta_i, b_j) = \frac{\exp[Da_j(\theta_i - b_j)]}{1 + exp[Da_j(\theta_i - b_j)]}$$

Donde:

$x_{ij} = Respuesta\ del\ i - ésimo\ examinado\ al\ j$

$- ésimo\ reactivo\ (1\ si\ es\ correcto; 0\ en\ otro\ caso).$

$\theta_i = Habilidad\ del\ i - ésimo\ examinado.$

$a_j = Discriminación\ del\ j - ésimo\ reactivo.$

$b_j = Dificultad\ del\ j - ésimo\ reactivo.$

$D = Factor\ de\ escala\ para\ aproximar\ la\ función\ a\ una\ ojiva\ normal\ (1.7).$

Para el modelo de TRI de tres parámetros se utilizó la siguiente fórmula:

$$P(x_{ij}|\theta_i, b_j) = c_j + (1 - c_j)\frac{\exp[Da_j(\theta_i - b_j)]}{1 + exp[Da_j(\theta_i - b_j)]}$$

Donde:

$x_{ij} = Respuesta\ del\ i - ésimo\ examinado\ al\ j$

$- ésimo\ reactivo\ (1\ si\ es\ correcto; 0\ en\ otro\ caso).$

$\theta_i = Habilidad\ del\ i - ésimo\ examinado.$

$a_j = Discriminación\ del\ j - ésimo\ reactivo.$

$b_j = Dificultad\ del\ j - ésimo\ reactivo.$

$c_j = pseudo\ oportunidad\ (adivinación)\ del\ j - ésimo\ reactivo.$

$D = Factor\ de\ escala\ para\ aproximar\ la\ función\ a\ una\ ojiva\ normal$ (1.7).

Se realizó análisis del funcionamiento diferencial de los ítems por sexo (DIF, por sus iniciales en inglés), de acuerdo con el procedimiento de Mantel-Haenszel con la modificación del modelo de Rasch de un parámetro (García-Medina, Martínez-Rizo y Cordero Arroyo, 2016; Linacre y Wright, 1989). Se utilizaron los criterios del ETS para identificar DIF insignificante, leve-moderado y moderado-alto (Dorans y Holland, 1992)

5. Resultados

El examen se aplicó con la metodología descrita para exámenes sumativos de gran escala (Lane et al., 2016), en instrumentos impresos a contestar con lápiz y hoja de respuestas, los días 23 y 24 de febrero de 2019, en 25 sedes en el área metropolitana de la Ciudad de México. El examen fue respondido por 148.407 aspirantes a las licenciaturas de las cuatro áreas de conocimiento que oferta la UNAM: CFMI=Ciencias Físico Matemáticas e Ingenierías; CBQS=Ciencias Biológicas, Químicas y de la Salud; CS=Ciencias Sociales; HyA=Humanidades y Artes (cuadro 1).

Cuadro 1. Número y porcentaje de aspirantes por área del conocimiento

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

CFMI=Ciencias Físico Matemáticas e Ingenierías; CBQS=Ciencias Biológicas, Químicas y de la Salud; CS=Ciencias Sociales; HyA=Humanidades y Artes.
Fuente: Elaboración propia con información de la Dirección General de Administración Escolar (DGAE) de la UNAM.

En el cuadro 2 se muestra el número y porcentaje de aspirantes por sexo y área del conocimiento solicitada.

Cuadro 2. Número y porcentaje de aspirantes por sexo y área de conocimiento

	Área				Total
	CFMI	CBQS	CS	HyA	
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%

CFMI=Ciencias Físico Matemáticas e Ingenierías; CBQS=Ciencias Biológicas, Químicas y de la Salud; CS=Ciencias Sociales; HyA=Humanidades y Artes.
Fuente: Elaboración propia con información de la Dirección General de Administración Escolar (DGAE) de la UNAM.

Se analizó la información del examen que atendiera a cada una de las fuentes de evidencia de validez descritas en la sección de Método, que a continuación se describen:

- *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, posteriormente 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 3, integrándose con 120 reactivos.

Cuadro 3. Número de reactivos del examen de ingreso a las licenciaturas de la UNAM, por área del conocimiento

MATERIA	CFMI	CBQS	CS	HYA
Matemáticas	26	24	24	22
Física	16	12	10	10
Química	10	13	10	10
Biología	10	13	10	10
Historia universal	10	10	14	10
Historia 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

CFMI=Ciencias Físico Matemáticas e Ingenierías; CBQS=Ciencias Biológicas, Químicas y de la Salud; CS=Ciencias Sociales; HyA=Humanidades y Artes.
Fuente: Elaboración propia con información de la Dirección de Evaluación Educativa de la CODEIC UNAM.

- *Procesos de respuesta.* Este apartado se refiere a 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 habilidad en el uso de computadoras. 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.

  Desarrollamos una plataforma informática para el desarrollo y validación del examen, que tiene más de una década de perfeccionamiento e integración, el "Sistema Integral de Gestión de Exámenes" (SIGE UNAM, marca registrada), lo que proporciona un elemento más de validez al desarrollo del examen, la validación de los reactivos, y la integración del banco de reactivos con características apropiadas. 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.

- *Estructura interna.* Se refiere a las características estadísticas del examen, como estadísticas descriptivas y análisis de reactivos, el funcionamiento de los distractores, la confiabilidad del examen, entre otros (AERA, 2014). En la figura 2

se muestra la distribución de aciertos por área del conocimiento, en la que se observa claramente una tendencia de agrupamiento de los aspirantes hacia la izquierda en el área de menor cantidad de aciertos. Los patrones de distribución de aciertos en las versiones y ordenamientos del examen son prácticamente idénticas (datos no mostrados).

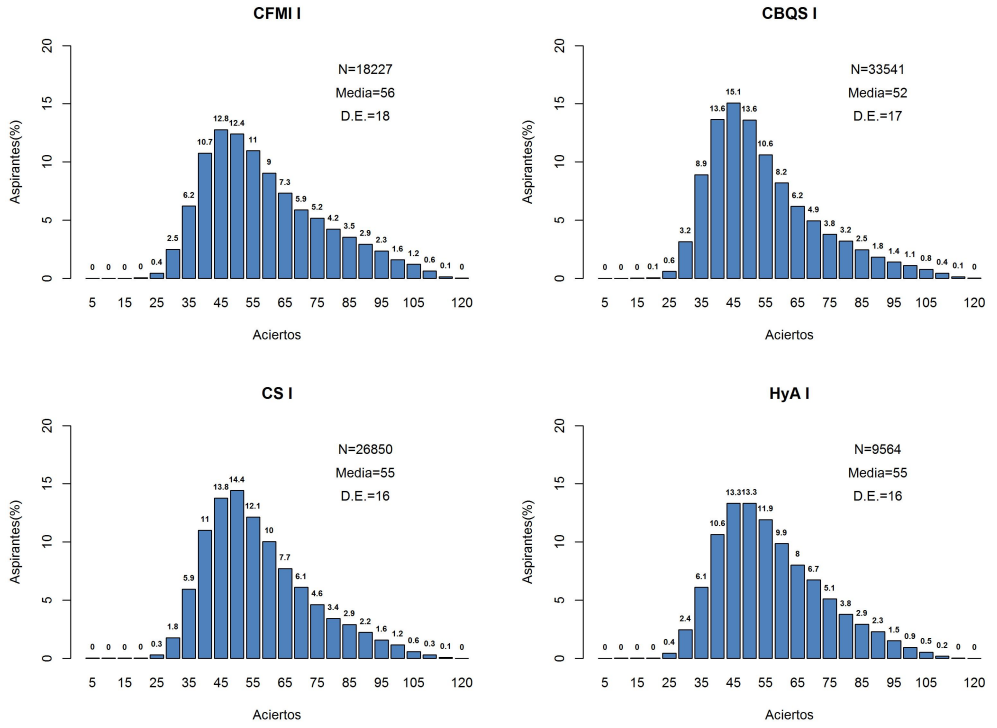

Figura 2. Distribución de aciertos en el examen de admisión a la licenciatura de la UNAM, por área del conocimiento

CFMI=Ciencias Físico Matemáticas e Ingenierías; CBQS=Ciencias Biológicas, Químicas y de la Salud; CS=Ciencias Sociales; HyA=Humanidades y Artes.

Fuente: Elaboración propia.

En el cuadro 4 se presentan los resultados del análisis psicométrico con la TMC, incluyendo índice de dificultad promedio, índice de discriminación promedio con coeficiente de punto biserial, error estándar de medición y confiabilidad determinada con el alfa de Cronbach.

Cuadro 4. Análisis psicométrico con la Teoría de Medición Clásica del examen de ingreso a la licenciatura de la UNAM, por campo de conocimiento y versión del examen

	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 de aciertos	55,4	55,7	52,5	52,9	54,8	54,6	54,8	55,1
Desviación 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	5,00	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
CPB 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

N=148,407. CFMI=Ciencias Físico Matemáticas e Ingenierías; CBQS=Ciencias Biológicas, Químicas y de la Salud; CS=Ciencias Sociales; HyA=Humanidades y Artes; EEM=error estándar

de medición; CPB=correlación punto biserial.
Fuente: Elaboración propia.

En el cuadro 5 se presentan los resultados del análisis psicométrico con la TRI.

Cuadro 5. Análisis psicométrico con la Teoría de Respuesta al Ítem de uno, dos y tres parámetros, del examen de ingreso a la licenciatura de la UNAM, por campo de conocimiento

		Modelos de TRI											
		Un parámetro		**Dos parámetros**				**Tres parámetros**					
		b		**a**		**b**		**a**		**b**		**c**	
Área	**r**	$\bar{x}$	**D.E.**	$\bar{x}$	**D.E.**	$\bar{x}$	**D.E.**	$\bar{x}$	**D.E.**	$\bar{x}$	**D.E.**	$\bar{x}$	**D.E.**
CFMI	204	0,277	1,053	0,788	0,381	0,540	1,968	1,409	0,842	1,009	1,745	0,189	0,089
CBQS	201	0,455	1,012	0,710	0,313	0,616	1,429	1,458	0,777	1,073	0,960	0,200	0,116
CS	204	0,253	1,155	0,703	0,336	0,303	1,963	1,305	0,776	0,830	1,229	0,197	0,116
HyA	200	0,311	1,195	0,715	0,346	0,790	3,947	1,136	0,604	1,021	1,799	0,181	0,094
Total	809	0,324	1,109	0,729	0,347	0,561	2,517	1,327	0,765	0,983	1,479	0,192	0,105

N=148,407. r=reactivos únicos por área; TRI=Teoría de Respuesta al Ítem; D.E.=Desviación estándar; CFMI=Ciencias Físico Matemáticas e Ingenierías; CBQS=Ciencias Biológicas Químicas y de la Salud; CS=Ciencias Sociales; HyA=Humanidades y Artes.
Fuente: Elaboración propia.

Todos los reactivos que se utilizan en el examen cubren criterios psicométricos predefinidos, obtenidos con el análisis psicométrico de TMC y TRI. Todos provienen de un banco de reactivos extenso, y han sido analizados en estudios piloto con estudiantes para documentar su comportamiento psicométrico previo. En la figura 3 podemos ver los resultados del mapa de Wright, comparando la dificultad de los reactivos con la habilidad de los aspirantes, en el área de las Ciencias Físico Matemáticas e Ingenierías, con el modelo de Rasch de la TRI (Andrich y Marais, 2019; Boone y Noltemeyer, 2017). 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 el rango de niveles de 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 mayor concentración de estudiantes hacia el extremo de menor habilidad.

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 y Serrano, 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 el contraste en el nivel de dificultad de un reactivo entre los grupos de interés no supera los 0,43 lógitos representa un DIF sin importancia; si es mayor a 0,64 lógitos representa DIF moderado-alto; si se encuentra entre estos valores se trata de un DIF leve-moderado (Holland y Weiner 1993). Encontramos muy pocos reactivos con DIF leve-moderado, los cuales fueron valorados por un cuerpo colegiado para analizar la lógica del reactivo y el resultado de aprendizaje explorado, y así determinar su potencial efecto en los resultados. En la figura 4 podemos ver un ejemplo de los reactivos del área de Matemáticas, en los que no se encontraron reactivos con DIF en cuanto a sexo.

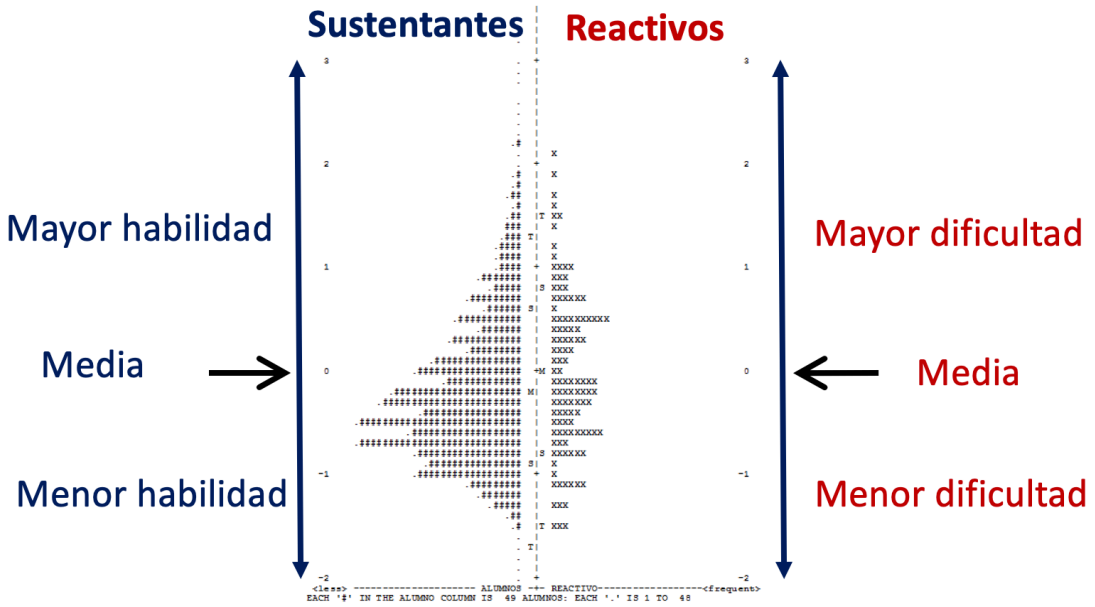

Figura 3. Mapa de dificultad de los reactivos y habilidad de los estudiantes, con el Modelo de Rasch
N=148.407
Fuente: Elaboración propia.

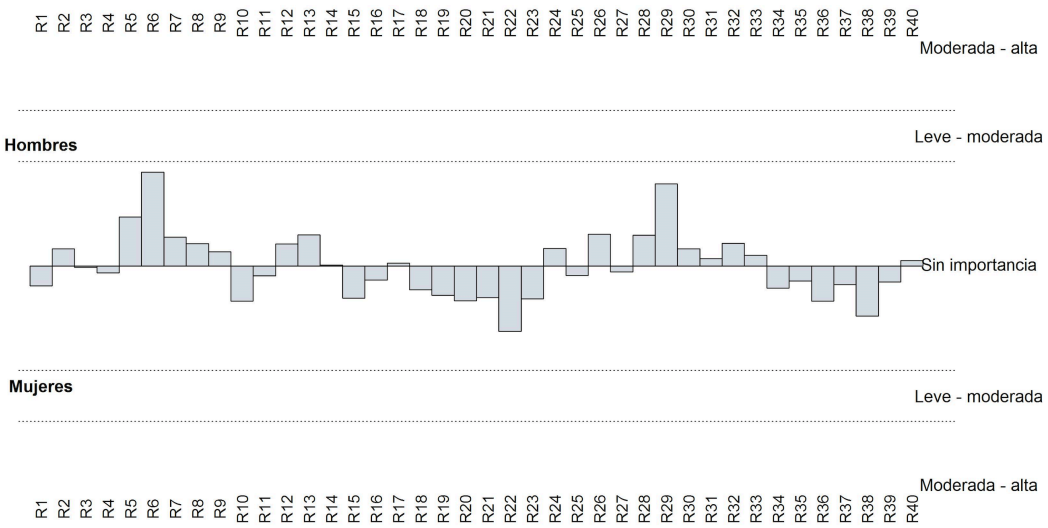

Figura 4. Funcionamiento diferencial de los ítems (DIF) de los reactivos de Matemáticas del examen de admisión a la licenciatura de la UNAM de febrero de 2019 según el sexo de los aspirantes
Fuente: Elaboración propia.

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

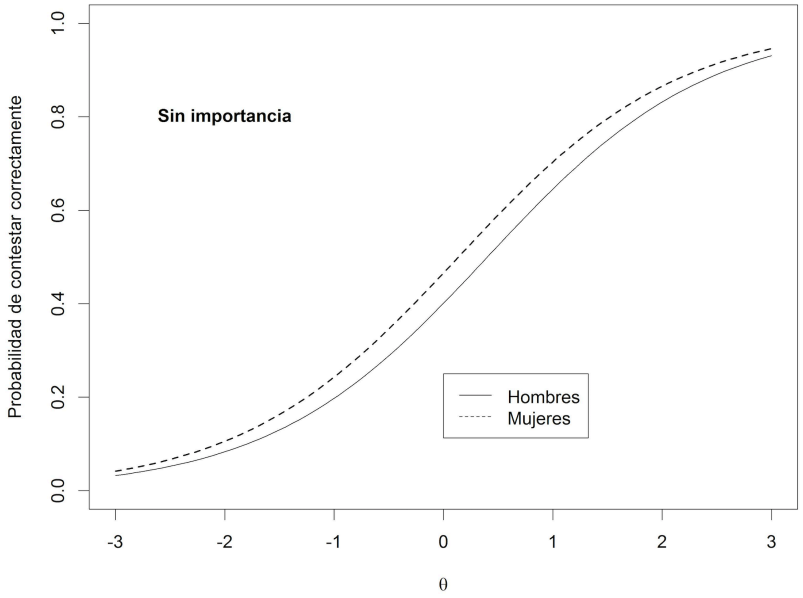

Figura 5. Visualización de un reactivo de matemáticas del examen de admisión a la licenciatura de la UNAM con un funcionamiento diferencial (DIF) sin importancia por sexo

Fuente: Elaboración propia.

En la figura 6 se observa 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 cancela con los reactivos con DIF leve para mujeres. Es importante destacar que estas cifras son solamente elementos estadísticos, que por sí solos no documentan sesgo a favor o en contra de una población, deben evaluarse cualitativamente por un grupo de expertos en contenido y evaluación, para analizar su potencial efecto en los resultados.

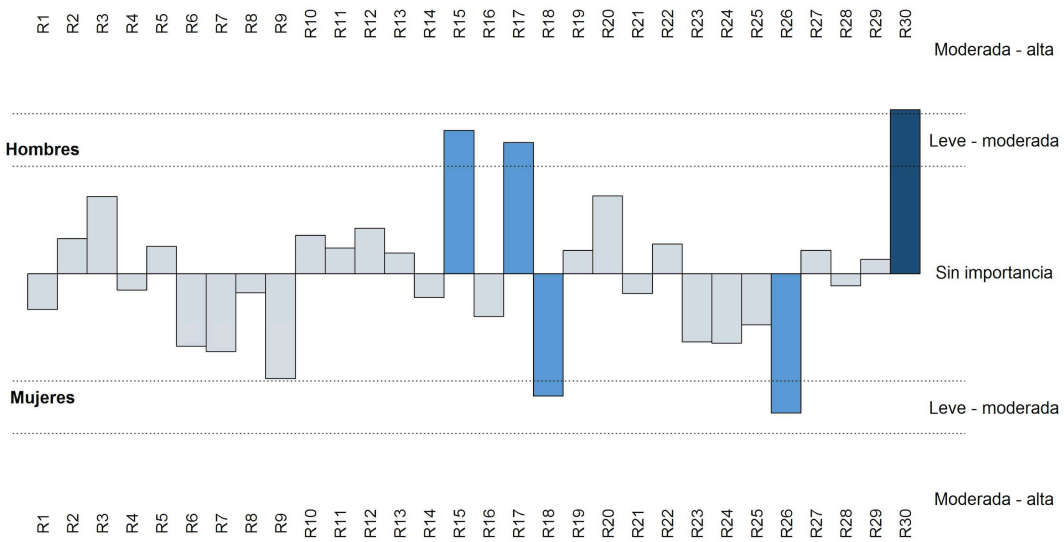

Figura 6. DIF por sexo de los reactivos de Español, del área de Ciencias Biológicas, Químicas y de la Salud, en el examen de admisión a la licenciatura de la UNAM, febrero de 2019

Fuente: Elaboración propia.

- *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. Actualmente estamos explorando la correlación del examen de admisión con el desempeño escolar a lo largo de las carreras y la eficiencia terminal. Por otra parte, encontramos una importante relación entre el desempeño en el examen diagnóstico al ingreso de la UNAM y el éxito en sus trayectorias académicas (Martínez-González et al., 2018). Al existir correlación entre el examen de ingreso y el examen diagnóstico, y entre el examen diagnóstico y el éxito académico, pudiera existir correlación entre el examen de ingreso y el desempeño a lo largo de la carrera, hipótesis que debe probarse.
- *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, por lo que 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 del proceso de selección.

6. Discusión y conclusiones

El análisis del examen de ingreso a las licenciaturas de la UNAM ofrece un panorama de información con datos de diversas fuentes de evidencia de validez, que proveen un retrato evaluativo del instrumento. Validez en evaluación educativa implica una aproximación científica a la interpretación de los resultados de los exámenes, es decir, probar hipótesis sobre los conceptos evaluados en el examen (AERA, 2014; Kane, 2016; Shepard, 2016). La información proporcionada por el uso de un instrumento de evaluación no es válida o inválida *per se*, sino que forma parte de un espectro de datos e información que deben utilizarse de manera sensata, integral y contextualizada, en el sentido de que las puntuaciones obtenidas por los sustentantes en un examen proveen más o menos evidencia para apoyar o rechazar una interpretación específica (por ejemplo aprobar o no un curso, admitir o no a un estudiante en la universidad) (Buntis, Buntis y Eggert, 2017; Kane, 2016; Mislevy, 2016; Young et al., 2016). Las organizaciones que elaboran e implementan los exámenes sumativos de alto impacto (entidades gubernamentales, instituciones educativas) son los principales responsables de 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; Gregory, 2016; Sireci, 2016). Las dependencias universitarias que elaboran exámenes tienen la obligación ética de documentar qué tan defendible es la interpretación de los resultados.

Los resultados descritos en este trabajo proporcionan evidencias de validez para el uso de este examen como instrumento de evaluación del conocimiento. Los datos obtenidos y sus magnitudes son compatibles con los recomendados por organizaciones nacionales e

internacionales para exámenes sumativos de alto impacto a gran escala, desde la estrategia de elaboración del instrumento, hasta las cifras de análisis psicométrico y confiabilidad (AERA, 2014; Young et al., 2016). Es importante enfatizar que el examen de ingreso a la Universidad es solamente un instrumento de medición del conocimiento, nada más, pero tampoco nada menos. El uso de los resultados que se obtienen con los instrumentos de evaluación y las inferencias que se hacen de los mismos es un tema extraordinariamente complejo.

En las últimas décadas las principales organizaciones de evaluación educativa del mundo han hecho énfasis en la necesidad de que se incluyan elementos que propicien justicia y equidad en el proceso, 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 utilizan instrumentos uniformes que se aplican en contextos altamente controlados, con la intención de que cada estudiante se enfrente al mismo reto en igualdad de condiciones, por definición tratan a todos los estudiantes de la misma manera. Por ejemplo, la riqueza de la heterogeneidad de los seres humanos es poco susceptible de "medirse" con instrumentos estandarizados, ya que estos no capturan fácilmente los matices de la individualidad de las personas. Además, los resultados individuales en un examen sumativo, en un momento específico en el tiempo, no necesariamente reflejan la realidad holística y longitudinal de la persona, ni de forma absoluta su nivel preciso de conocimiento y aplicación del mismo en solución de problemas en la vida real. Esta tensión entre las diversas perspectivas filosóficas de lo que debe ser la evaluación educativa continúa sin resolverse, lo que motiva discusiones intensas en contextos académicos (Martínez-Rizo, 2001; Sánchez-Mendiola y Delgado-Maldonado, 2017).

El hecho es que el ingreso de aspirantes a organizaciones que tienen recursos y cupo limitados, como las universidades públicas, obliga a tomar decisiones difíciles que no dejan totalmente satisfecha a la población, principalmente a aquellos que no son admitidos. La comprensión cabal del concepto moderno de validez por la comunidad académica 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.

Es indispensable explorar diversos mecanismos tanto tradicionales como innovadores, que puedan combinarse para generar procesos de admisión socialmente aceptables, metodológicamente correctos y éticamente justificables, tarea compleja en la época actual, especialmente si los recursos son limitados. A nivel internacional el inexorable proceso de expansión de la demanda por la educación superior se ha asociado con respuestas diferenciadas en regiones, países y universidades (Sigal y Dávila, 2004; Trost, 1993). En algunos casos se realizan estrategias nacionales rigurosas con un mecanismo centralizado que usa diferentes elementos evaluativos como en Chile, en otros como Argentina se ha usado el acceso irrestricto, y en otros países como Colombia, Brasil y Panamá se emplean diferentes esquemas de exámenes de admisión, pruebas estandarizadas y una variedad de metodologías en sus procesos (Sigal y Dávila, 2004; Trost, 1993). En México cada institución educativa define sus mecanismos de ingreso, así como los criterios de selección y los niveles de exigencia en los diferentes componentes del proceso de acuerdo a su normatividad y criterios técnicos. Varias universidades mexicanas utilizan instrumentos estandarizados desarrollados por organizaciones dedicadas a evaluación educativa, como el Centro Nacional de Evaluación para la Educación Superior (CENEVAL, 2020), además

de de evaluaciones psicológicas, el promedio del bachillerato, entrevistas personales, ensayos, entre otros. Cada institución lo hace de acuerdo a sus recursos, normatividad, posibilidades y tamaño de la demanda. Es importante hacer notar que algunas instituciones que utilizan varios elementos e instrumentos de evaluación, generalmente no hacen público el proceso en detalle, como la ponderación de cada componente y cómo integran un resultado final que les permita tomar una decisión, por lo que el proceso se convierte en una especie de "caja negra" que puede dar lugar a desconfianza en el mismo.

El caso de la UNAM es único en el país, ya que es la institución de educación media superior y superior de México más reconocida a nivel nacional e internacional, y al ser pública y para fines prácticos gratuita, es solicitada por un número creciente de aspirantes. Actualmente la UNAM tiene 360.883 estudiantes (DGPL-UNAM, 2020) de los que 217.808 (60,3%) son de licenciatura. De acuerdo con la DGAE (2019), en el ciclo académico 2018-2019 participaron en el concurso de selección a las licenciaturas de la UNAM 261.157 aspirantes, de los cuales ingresaron 24.007 (9,2%). La UNAM tiene además la particularidad de que más de la mitad de sus estudiantes de ingreso a la licenciatura (aproximadamente el 55%) provienen de los bachilleratos de la misma institución, por el mecanismo de pase reglamentado. De cualquier manera, es necesario identificar las fuentes de evidencia de validez del examen de ingreso a las licenciaturas, ya de ello depende la decisión de ingreso por concurso de selección. El examen es el principal elemento decisorio, no se toma en cuenta el promedio en el bachillerato ni otros tipos de instrumentos externos, estudios psicológicos o entrevista personal, en virtud de la gran cantidad de aspirantes, la heterogeneidad de los promedios en las escuelas del sistema educativo nacional y las dificultades éticas y logísticas de utilizar otros elementos de forma equitativa y válida, como para incluirlos de manera ponderada en el puntaje de ingreso.

El examen tiene números satisfactorios desde el punto de vista psicométrico, incluyendo grado de dificultad, discriminación, confiabilidad, entre otros. Son pocos los estudios en nuestro país que documenten este tipo de cifras para establecer comparaciones (Backhoff, Tirado y Larrazolo, 2001), aunque en el diálogo con colegas de otras universidades refieren que sus exámenes tienen evidencia de validez, de acuerdo con sus reportes internos y mecanismos de control de calidad. En el contexto moderno de la educación superior en el que la demanda sobrepasa con mucho a la oferta, es importante que los exámenes estandarizados de alto impacto dejen de ser percibidos como un instrumento punitivo o de control, y que se haga conciencia de que es una herramienta académica que debe elaborarse con profesionalismo y atención a los detalles técnicos (AERA, 2014). Por otra parte, es necesario repensar el ingreso a la universidad como un sistema del que un examen escrito de conocimientos es un elemento del proceso, y reflexionar sobre el peso que se les da a los atributos exclusivamente académicos de los aspirantes (AERA, 2014; Juarros, 2006; Patterson et al., 2018).

El estudio tiene algunas limitaciones, ya que analiza solo una institución y los resultados pueden no ser aplicables en su totalidad a las demás universidades en México y Latinoamérica, por diferencias en tamaño, recursos y normatividad interna. A pesar de ello, creemos que al tratarse de la institución de educación superior más grande del país y una de las más grandes del mundo, con una gran cantidad de aspirantes y una baja tasa de aceptación, la descripción de la metodología y resultados encontrados es útil para la comunidad académica y tomadores de decisiones educativos, ya que presenta datos concretos y métodos rigurosos de medición educativa. Además, comienza a romper el paradigma tradicional de exceso de secrecía en los exámenes de gran escala y alto impacto,

para informar a la comunidad académica sobre estos temas. Generalmente la sociedad, los aspirantes y una parte del gremio docente desconocen los aspectos técnicos sofisticados de la medición educativa, mismos que dan fortaleza a los resultados del examen y el uso que se hace de ellos.

Es importante reconocer las limitaciones de los exámenes estandarizados de alto impacto, y la percepción social que se tiene de ellos en la actualidad. Este estudio se limitó a identificar las fuentes de evidencia de validez disponibles, pero no todas las que potencialmente existen, principalmente las relacionadas con consecuencias, costos sociales, psicológicos, económicos y de efectos a largo plazo en los aspirantes. Tampoco hemos analizado el seguimiento a largo plazo de los estudiantes, para identificar la correlación entre el examen de selección, la graduación y el éxito profesional de los graduados. Creemos que los espacios de oportunidad para ampliar nuestro conocimiento de los exámenes de alto impacto, y del proceso de selección en las universidades, es de gran importancia para las instituciones educativas y la sociedad en general.

Uno de los métodos más utilizados para tratar de identificar sesgos en los exámenes que pudieran poner en desventaja a algún grupo de personas, ya sea por sexo o nivel socioeconómico, es el análisis DIF (Alavi y Bordbar, 2017; Yavuz et al., 2018; Zieky,1993). En nuestra institución iniciamos en 2018 este tipo de análisis por sexo y por bachillerato de procedencia (público o privado), y no identificamos niveles de DIF importante en los reactivos del examen que pudieran comprometer sus resultados e inferencias. El análisis detallado con esta metodología será sujeto de otra publicación, ya que incorporaremos en el análisis los niveles socioeconómicos, variable por demás importante en nuestro contexto. Existen muy pocos estudios publicados sobre el uso de esta metodología en el país (García-Medina et al., 2016), por lo que consideramos debe impulsarse su uso en las universidades que realicen exámenes sumativos de alto impacto.

Los mecanismos y políticas de admisión a la educación superior son el resultado de una convergencia de factores históricos, sociales y de disponibilidad de espacio y recursos a lo largo del tiempo, y cada país, región y universidad han enfrentado el reto de acuerdo con su realidad local. No parece que a corto plazo vaya a disminuir la demanda de espacios, y el crecimiento de las universidades existentes requiere ineludiblemente de más recursos utilizados de manera eficaz para dar respuesta a las necesidades sociales. Es inevitable que los aspirantes a ingresar a la educación superior deseen hacerlo a las universidades con mayor prestigio y que no impliquen gastos directos de bolsillo para transitar en la licenciatura, por lo que las solicitudes de la mayoría de los estudiantes se concentran en unas cuantas instituciones, principalmente públicas, para seguir adelante en su trayectoria de vida. Ello implica la necesidad mecanismos de selección que impactan a gran cantidad de aspirantes.

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 la comunidad académica, las autoridades y los medios de comunicación (AERA, 2014; Martínez-Rizo, 2016). La asimetría de poder intrínseca en los procesos de evaluación conlleva una enorme responsabilidad de las autoridades institucionales, por lo que estos procesos deben realizarse atendiendo el estado del arte de la evaluación educativa, y promoviendo la profesionalización en este tema de los participantes en el proceso de admisión. Es importante explorar mecanismos alternativos e identificar elementos de decisión que pudieran incorporarse al proceso de admisión en nuestras universidades, pero también es fundamental revisar rigurosamente los mecanismos de selección existentes y

mejorar los instrumentos a la luz de publicaciones como la presente. La investigación en evaluación educativa debe convertirse en una prioridad de nuestras instituciones de educación superior (Schuwirth et al., 2010). Las universidades, espacio de trabajo de académicos de alto nivel, deben proveer servicios de evaluación educativa acordes con su prestigio institucional.

Referencias

Alavi, S. y Bordbar, S. (2017). Differential item functioning analysis of high-stakes test in terms of gender: A Rasch model approach. *Malaysian Online Journal of Educational Sciences*, *5*(1), 10-24.

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

Andrich, D. y Marais, I. (2019). A Course in Rasch Measurement Theory. En D. Andrich y I. Marais (Coords.), *Measuring in the Educational, Social and Health Sciences* (pp. 41-53). Springer.

Asociación Nacional de Universidades e Instituciones de Educación Superior ANUIES. (2019). *Anuario estadístico de la población escolar en la educación superior. Técnico Superior y Licenciatura, ciclo 2017-2018*. Recuperado de http://www.anuies.mx/informacion-y-servicios/informacion-estadistica-de-educacion-superior/anuario-estadistico-de-educacion-superior

Backhoff, E., Tirado, F. y Larrazolo, N. (2001). Ponderación diferencial de reactivos para mejorar la validez de una prueba de ingreso a la universidad. *Revista Electrónica de Investigación Educativa*, *3*(1), 1-10.

Bennett, R. E. (2005). What does it mean to be a nonprofit educational measurement organization in the 21st century?. En R. Bennett y M. von Davier (Eds.), *Advancing human assessment. The methodological, psychological and policy contributions of ETS* (pp. 1-15). Springer.

Boone, W. y Noltemeyer, A. (2017). Rasch analysis: A primer for school psychology researchers and practitioners. *Cogent Education*, *4*(14), 1-13. https://doi.org/10.1080/2331186X.2017.1416898

Buendía, M. A. y Rivera, R. (2010). Modelo de selección para el ingreso a la Educación Superior: El caso de la UACH. *Revista de la Educación Superior*, *39*(156), 55-72.

Buntis, M., Buntis, K. y Eggert, F. (2017). Clarifying the concept of validity: From measurement to everyday language. *Theory & Psychology*, *27*(5), 703-710. https://doi.org/10.1177/0959354317702256

Centro Nacional de Evaluación para la Educación Superior (CENEVAL). (2020). *EXANI-II Admisión*. CENEVAL.

Cizek, G. J. (2001). More unintended consequences of high-stakes testing. *Educational Measurement: Issues and Practice*, 20(4), 19-27. https://doi.org/10.1111/j.1745-3992.2001.tb00072.x

Cook, D. A., Bordage, G. y Schmidt, H. G. (2008). Description, justification and clarification: A framework for classifying the purposes of research in medical education. *Medical Education*, *42*(2), 128-133. https://doi.org/10.1111/j.1365-2923.2007.02974.x

Dirección General de Administración Escolar (DGAE) UNAM. (2019). *Demanda e ingreso a la licenciatura*. Recuperado de http://www.estadistica.unam.mx/series_inst/index.php

Dirección General de Administración Escolar (DGAE) UNAM. (2020). *Acerca de nosotros, quiénes somos y qué hacemos. DGAE, UNAM.CdMx.* Recuperado de https://www.dgae.unam.mx/acerca_nosotros.html.

Dirección General de Planeación (DGPL) UNAM. (2020). *Agenda Estadística 2020 UNAM.* Recuperado de: http://www.estadistica.unam.mx/agenda.php.

Dorans, N. J. y Holland, P. W. (1992). *DIF detection and description: Mantel-Haenszel and standardization. ETS Research Report Series, 1992*, 1-40. https://doi.org/10.1002/j.2333-8504.1992.tb01440.x

Frey, M. C. y Detterman, D. K. (2003). Scholastic assessment or G? The relationship between the Scholastic Assessment Test and general cognitive ability. *Psychological Science, 15*(6), 373-378. https://doi.org/10.1111/j.0956-7976.2004.00687.x

Gago, A. (2000). El CENEVAL y la evaluación externa de la educación en México. *Revista Electrónica de Investigación Educativa, 2*(2).

García-Medina, A. M., Martínez-Rizo, F. y Cordero Arroyo, G. (2016). Análisis del funcionamiento diferencial de los ítems del Excale de matemáticas para tercero de secundaria. *Revista Mexicana de Investigación Educativa, 21*(71), 1191-1220.

Graue, E. (2018). *Acuerdo que reorganiza las funciones y estructura de la Secretaría General de la Universidad Nacional Autónoma de México.* Gaceta UNAM.

Gregory, J. C. (2016). Validating test score meaning and defending test score use: different aims, different methods. *Assessment in Education: Principles, Policy & Practice, 23*(2), 212-225. https://doi.org/10.1080/0969594X.2015.1063479

Guzmán, C., y Serrano, O. (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. y Rodriguez, M. C. (2002). A review of multiple-choice item-writing guidelines for classroom assessment. *Applied Measurment in Education, 15*(3), 309-334. https://doi.org/10.1207/S15324818AME1503_5

Holland, P. y Weiner, H. (1993). *Differential Item Functioning.* Laurence Erlbaun Associates.

Juarros, M. (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. (2016). Explicating validity. *Assessment in Education: Principles, Policy & Practice, 23*(2), 198-211. https://doi.org/10.1080/0969594X.2015.1060192

Kane, M. y Bridgeman, B. (2017). Research on validity theory and practice at ETS. En R. Bennett y M. von Davier (Eds.), *Advancing Human Assessment. Methodology of Educational Measurement and Assessment* (pp. 489-551). Springer.

Lane, S., Raymond, M. R., Haladyna, T. M. y Downing, S. M. (2016). Test development process. En S. Lane, M. R. Raymond y T. M. Haladyna (Eds.), *Handbook of test development* (pp. 3-18). Routledge.

Linacre J. M. y Wright, B. D. (1989). Mantel-Haenszel DIF and PROX are equivalent! *Rasch Measurement Transactions, 3*(2), 52-53.

Manzi, J., Bosch, A., Bravo, D., del Pino, G., Donoso, G. y Pizarro, R. (2010). Validez diferencial y sesgo en la predictividad de las pruebas de admisión a las universidades chilenas. *Revista Iberoamericana de Evaluación Educativa, 3*(2), 30-48.

Martínez-González, A., Sánchez-Mendiola, M., Manzano-Patiño, A., García-Minjares, M., Herrera-Penilla, C. y Buzo-Casanova, E. (2018). Grado de conocimientos de los estudiantes

al ingreso a la licenciatura y su asociación con el desempeño escolar y la eficiencia terminal. Modelo multivariado. *Revista de la Educación Superior, 47(*188), 57-85.

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

Martínez-Rizo, F. (2016). Impacto de las pruebas en gran escala en contextos de débil tradición técnica: Experiencia de México y el Grupo Iberoamericano de PISA. *RELIEVE, 22*(1), M0. http://dx.doi.org/10.7203/relieve.22.1.8244

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

Mislevy, R. J. (2016), How developments in psychology and technology challenge validity argumentation. *Journal of Educational Measurement, 53*(3), 265-292. https://doi.org/10.1111/jedm.12117

OCDE (2018). *"How do admission systems affect enrolment in public tertiary education?" Education Indicators in Focus.* Recuperado de https://www.oecd-ilibrary.org/deliver/41bf120b-en.pdf?itemId=%2Fcontent%2Fpaper%2F41bf120b-en&mimeType=pdf https://doi.org/10.1787/41bf120b-en

Ordorika, I. Rodríguez, R. A. y Montes de Oca, M. M. (2013). Estudio Comparativo de Universidades Mexicanas. Fichas Institucionales 2007-2011. En DGEI-UNAM (Eds.), *Cuadernos de Trabajo de la Dirección General de Evaluación Institucional* (pp. 227-230). DGEI-UNAM.

Patterson, F., Roberts, C., Hanson, M. D., Hampe, W., Eva, K., Ponnamperuma, G., et al. (2018). 2018 Ottawa consensus statement: Selection and recruitment to the healthcare professions. *Medical Teaching, 40*(11), 1091-1101. https://doi.org/10.1080/0142159X.2018.1498589

Raykov, T. y Marcoulides, G. A. (2016). On the relationship between classical test theory and item response theory: From one to the other and back. *Educational and Psychological Measurement,* 76(2), 325–338. https://doi.org/10.1177/0013164415576958

Ringsted, C., Hodges, B. y Scherpbier, A. (2011). The research compass: An introduction to research in medical education: AMEE Guide nº 56. *Medical Teaching, 33*(9), 695-709. https://doi.org/10.3109/0142159X.2011.595436

Sánchez Mendiola, M., Delgado Maldonado, L., Flores Hernández, F., Leenen, I. y Martínez González, A. (2015). Evaluación del aprendizaje. En M. Sánchez Mendiola, A. Lifshitz Guinzberg, P. Vilar Puig, A. Martínez González, M. Varela Ruiz, M. y E. Graue Wiechers, (Eds.), *Educación Médica: Teoría y Práctica* (pp. 89-95). Elsevier.

Sánchez-Mendiola, M. y Delgado-Maldonado, L. (2017). Exámenes de alto impacto: Implicaciones educativas. *Investigación Educativa Médica, 6*(21), 52-62. https://doi.org/10.1016/j.riem.2016.12.001

Schuwirth, L., Colliver, J., Gruppen, L., Kreiter, C., Mennin, S., Onishi, H., et al. (2011). Research in assessment: Consensus statement and recommendations from the Ottawa 2010 Conference. *Medical Teaching, 33*(3), 224-233. https://doi.org/10.3109/0142159X.2011.551558

Shepard, L. (2016). Evaluating test validity: Reprise and progress. *Assessment in Education: Principles, Policy & Practice, 23*(2), 268-280. https://doi.org/10.1080/0969594X.2016.1141168

Sigal, V. y Dávila, M. (2004). La cuestión de la admisión a los estudios universitarios en Argentina. En O. Barsky, V. Sigal y M. Dávila (Eds.), *Los desafíos de la universidad argentina* (pp. 205-222). Siglo XXI Editores.

Sireci, S. G. (2016). On the validity of useless tests. *Assessment in Education: Principles, Policy & Practice*, *23*(2), 226-235. https://doi.org/10.1080/0969594X.2015.1072084

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.

UNAM. (1997). *Reglamento General de Inscripciones. Universidad Nacional Autónoma de México.* Recuperado de https://www.dgae-siae.unam.mx/acerca/normatividad.html#leg-3.

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*(4), 364-376. https://doi.org/10.1177/0734282911406666

Yavuz, S., Dogan, N., Hambleton, R. K. y Yurtcu, M. (2018). The comparison of differential item functioning predicted through experts and statistical techniques. *Cypriot Journal of Educational Science*, *13*(2), 375-384. https://doi.org/10.18844/cjes.v13i2.2427

Young, M., St-Onge, C., Xiao, J., Vachon Lachiver, E. y Torabi, N. (2018). Characterizing the literature on validity and assessment in medical education: a bibliometric study. *Perspectives on Medical Education*, *7*(3), 182-191. https://doi.org/10.1007/s40037-018-0433-x

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

Zwick, R. (2006). Higher Education Admissions Testing, En R. Brennan (Ed.), *Educational Measurement* (pp. 647-679). National Council on Measurement in Education Greenwood Press.

Breve Cv de los autores

Melchor Sánchez Mendiola

Médico pediatra por la Universidad del Ejército y Fuerza Aérea, México; Maestro en Educación en Profesiones de la Salud por la Universidad de Illinois en Chicago, EUA; Doctor en Ciencias de la Educación por la UNAM. Profesor de Carrera Titular C de Tiempo Completo Definitivo, División de Estudios de Posgrado, Facultad de Medicina, Universidad Nacional Autónoma de México (UNAM). Participa en proyectos de evaluación educativa y educación en profesiones de la salud. ORCID ID: https://orcid.org/0000-0002-9664-3208. Email: melchorsm@unam.mx

Manuel García Minjares

Actuario con estudios de Maestría en Estadística e Investigación de Operaciones por la Universidad Nacional Autónoma de México (UNAM). Profesor de Estadística, Probabilidad, Matemáticas y Operaciones en los sistemas escolarizados y a distancia de la Facultad de Contaduría y Administración de la UNAM. Actualmente es Jefe de la Unidad de Estadística y Análisis de Datos de la Dirección de Evaluación Educativa de la UNAM. ORCID ID: https://orcid.org/0000-0002-9535-5917. Email: mminjares@unam.mx

Adrián Martínez González

Médico Cirujano por la Universidad Nacional Autónoma de México. Doctor en Salud Pública y Medicina Preventiva por la Universidad Autónoma de Madrid. Profesor de Carrera Titular C Tiempo Completo Definitivo, Facultad de Medicina, UNAM. Miembro de la Academia Nacional de Medicina de México y del Sistema Nacional de Investigadores. Actualmente Director de Evaluación Educativa de la UNAM. Participa en proyectos de

evaluación educativa y educación en profesiones de la salud. ORCID ID: https://orcid.org/0000-0002-5021-9639. Email: adrianmartinez38@gmail.com

Enrique Buzo Casanova

Licenciatura en Psicología por la Universidad Nacional Autónoma de México (UNAM). Especialidad en psicoterapia de corte psicoanalítico por la UNAM. Profesor de Asignatura de la Facultad de Psicología de la UNAM. Miembro del Colegio Nacional de Psicólogos de México. Actualmente Subdirector de Evaluación de Bachillerato y Licenciatura, en la Dirección de Evaluación Educativa de la UNAM. Participa en proyectos de evaluación educativa y educación de bachillerato y licenciatura. ORCID ID: https://orcid.org/0000-0001-7490-7826. Email: erbuzo@unam.mx