UNIVERSIDAD NACIONAL AUTÓNOMA DE MÉXICO

PROGRAMA DE MAESTRÍA Y DOCTORADO EN CIENCIAS MÉDICAS, ODONTOLÓGICAS Y DE LA SALUD

CAMPO DEL CONOCIMIENTO:

Ciencias Sociomédicas y Humanidades en Salud

CAMPO DISCIPLINARIO:

Educación en Ciencias de la Salud

TÍTULO:

Una aproximación metodológica para evaluar la validez del proceso de admisión en las escuelas de medicina: aplicación en la Universidad Autónoma de San Luis Potosí.

Modalidad tesis doctoral con producción científica que para optar por el grado de

DOCTORA EN CIENCIAS

PRESENTA:

M. en C. Blanca Ariadna Carrillo Avalos

TUTOR PRINCIPAL:

Dr. Melchor Sánchez Mendiola

COMITÉ TUTOR:

Dr. Iwin Leenen

Dr. Juan Andrés Trejo Mejía

Ciudad de México, Febrero de 2024

PROTESTA UNIVERSITARIA DE INTEGRIDAD Y
HONESTIDAD ACADÉMICA Y PROFESIONAL
(Graduación con trabajo escrito)

De conformidad con lo dispuesto en los artículos 87, fracción V, del Estatuto General, 68, primer párrafo, del Reglamento General de Estudios Universitarios y 26, fracción I y 35 del Reglamento General de Exámenes, me comprometo en todo tiempo a honrar a la Institución y a cumplir con los principios establecidos en el Código de Ética de la Universidad Nacional Autónoma de México, especialmente con los de integridad y honestidad académica.

De acuerdo con lo anterior, manifiesto que el trabajo escrito titulado "**Una aproximación metodológica para evaluar la validez del proceso de admisión en las escuelas de medicina: aplicación en la Universidad Autónoma de San Luis Potosí.**" que presenté para obtener el grado de Doctora en Ciencias en Educación en Ciencias de la Salud es original, de mi autoría y lo realicé con el rigor metodológico exigido por mi programa de posgrado, citando las fuentes de ideas, textos, imágenes, gráficos u otro tipo de obras empleadas para su desarrollo.

En consecuencia, acepto que la falta de cumplimiento de las disposiciones reglamentarias y normativas de la Universidad, en particular las ya referidas en el Código de Ética, llevará a la nulidad de los actos de carácter académico administrativo del proceso de graduación.

Atentamente

Blanca Ariadna Carrillo Avalos
Número de cuenta 508212354

Vo. Bo. Tutor principal
Dr. Melchor Sánchez Mendiola

CONTENIDO

ÍNDICE DE FIGURAS

ÍNDICE DE TABLAS

DEDICATORIA

Para Marco Banda

Esta tesis es para Marco Banda: eres mi inspiración, estímulo, apoyo y mucho más. Gracias por estar a mi lado y, con tu ejemplo, guiarme para ser mejor.

A Ana y a Emma

Por el ánimo que me dan, sus lindas palabras, su cariño y el tiempo que tomé de ustedes para este proyecto.

A José Luis y Azucena

Por estar incondicionalmente con nosotros, apoyándonos, enseñándonos y alentándonos. Gracias, sin su apoyo no lo hubiera logrado.

Sam, gracias por estar siempre pendiente.

AGRADECIMIENTOS

Al Dr. Melchor Sánchez Mendiola

Gracias por recibirme como su alumna y darme la oportunidad de aprender y publicar junto a usted: ha sido un gran honor. Gracias por ser un gran ejemplo de maestro e investigador, por su paciencia y por abrir mi visión ante el amplio campo de la Educación en Ciencias de la Salud.

Al Dr. Iwin Leenen

Gracias por su enorme paciencia, las discusiones en torno a los temas de validez, el tiempo que dedicó a explicarme el análisis estadístico y a revisar y mejorar los manuscritos que publicamos juntos.

Al Dr. Andrés Trejo

Gracias por sus observaciones a las presentaciones de la tesis y a los manuscritos que publicamos juntos, y por todo su apoyo.

A la Dra. Fortoul

Gracias por su franqueza y por sus enseñanzas, aprendí muchísimo de usted. Es una gran maestra tanto en el campo de la Educación en Ciencias de la Salud como en Histología, ambas maravillosas ciencias.

Al Dr. Adrián Martínez

Gracias por aceptar fungir como tutor provisional para poder ingresar a este posgrado, por sus observaciones y comentarios tanto durante las presentaciones de la tesis como durante el examen de candidatura. Su apoyo ha sido invaluable.

A la Facultad de Medicina de la UASLP

Gracias a mi alma máter por compartir los datos necesarios para la elaboración de esta tesis. Espero que este manuscrito contribuya a los esfuerzos que hacemos todos para seguir siendo de las mejores Facultades de Medicina del país.

1. RESUMEN

Introducción. Las evaluaciones de alto impacto para profesiones de la salud, como los exámenes de admisión y los exámenes profesionales, tienen consecuencias sobre sustentantes, profesores, administradores educativos y pacientes, por lo que se debe asegurar la validez de la interpretación de sus resultados. Los marcos de referencia de validez más conocidos son el de Messick y el de Kane, que en general son aplicados de manera parcial para validación en exámenes de admisión.

Objetivo. Aplicar una nueva propuesta de integración de los marcos de referencia de validez modernos para conocer el grado de validez del proceso de admisión en la facultad de Medicina de la Universidad Autónoma de San Luis Potosí.

Métodos. Se realizó un estudio observacional analítico en la Facultad de Medicina de la Universidad Autónoma de San Luis Potosí, sobre el proceso de admisión de las generaciones 2013 y 2014, que incluyó tres evaluaciones en cada cohorte: examen psicométrico (EP), examen de conocimientos (EC) y EXANI-II. Se aplicó un método de validación consistente en dos pasos generales: 1. Argumento de usos e interpretaciones y 2. Argumento de validez, conformado, a su vez, por tres etapas: I. Validez instrumental, II. Verificación de la interpretación y la decisión, y III. Utilidad de las acciones. En cada etapa se establecieron hipótesis sobre las inferencias de Kane para comprobarse por medio de las fuentes de evidencia de Messick.

Resultados. Se recabó evidencia para evaluar la validez de la interpretación de los resultados de los procesos de admisión de 2013 y de 2014 en una escuela de medicina de México. Se analizaron resultados de 1,373 aspirantes en 2013 y 1,554 de 2014, así como de los 145 alumnos admitidos en cada generación. Se identificó un factor "g" de inteligencia, que tiene repercusión en los resultados del proceso de admisión. El EC explica el 15% de la varianza de las calificaciones del primer año (2013), el EP y el EXANI-II no tienen influencia en el primer año de la carrera (2013 y 2014). Hay correlación entre algunos subcomponentes que miden un constructo semejante. Existe diferencia en los resultados de Biología (EC) entre hombres y mujeres (2014). El nivel de satisfacción de los alumnos acerca del proceso es bueno.

Discusión. Las fuentes de evidencia apoyan un grado de validez adecuado para la interpretación de los resultados de los procesos de admisión evaluados. Se requiere de un equipo

interdisciplinario para recabar y evaluar sistemáticamente todas las fuentes de evidencia. En cuanto a limitaciones del estudio se observó que los datos crudos de algunas de las fuentes de evidencia de validez para los procesos estudiados no están completos o disponibles; los datos corresponden a una sola escuela de medicina pública; y el proceso de recopilación de información y análisis fue retrospectivo. A través de los conocimientos generados en este trabajo se abren varios horizontes de aplicación, como aplicar el método a los procesos de admisión desde el momento que inicia el desarrollo de las pruebas que lo conformen, a otras evaluaciones de altas consecuencias como exámenes de titulación, así como a otros procesos similares en el área de las profesiones de la salud.

Conclusiones. Es necesario utilizar un marco de referencia claro para la validación de evaluaciones de alto impacto como los exámenes de admisión a las escuelas de medicina, así como planear prospectivamente la recolección de datos para evitar la falta de información relevante.

2. INTRODUCCIÓN

Las escuelas y facultades de medicina llevan a cabo procesos de admisión para seleccionar a los aspirantes a ingresar a la licenciatura en médico cirujano con base en las necesidades de la población a la que servirán los egresados, la política universitaria y sus capacidades de atención a los estudiantes. Los procesos de selección están diseñados para determinar cuáles son los aspirantes que tienen mayor probabilidad de tener éxito en sus estudios, pero no necesariamente predicen el desempeño académico durante los estudios. Es más, que un alumno tenga un excelente desempeño académico, no quiere decir que vaya a ser un excelente médico (Edwards et al., 2013). En general, se considera que este tipo de evaluaciones debe cumplir cuatro requisitos: determinar quiénes tienen mayor probabilidad de terminar el programa de estudios; que los aspirantes aceptados posean características deseables que no pueden ser enseñadas, mientras que carezcan de aquellas que son indeseables en un médico; que sea un método justo de selección con todos los aspirantes; y que provea de un estudiantado heterogéneo para que pueda afrontar todas las necesidades de la sociedad (Powis, 2015; Shulruf et al., 2012). A pesar de estos objetivos en común, no existe un método de selección de estudiantes que sea uniforme, internacional o estandarizado, quizá debido a que tampoco puede aislarse del contexto social, político y económico. Por lo anterior, los procesos de admisión son muy variables entre las instituciones de educación en todo el mundo en cuanto a tipo de evaluación, cantidad de pruebas, número de aspirantes a los que se aplican, y la forma en que se consideran los resultados. Estos procesos son la barrera que deben superar los aspirantes que deseen ingresar a cualquier escuela de medicina en todo el mundo, y su importancia en educación en ciencias de la salud radica en que las decisiones de hoy definirán el carácter de las profesiones de la salud por los siguientes 50 años, por lo que se incluyen en las llamadas pruebas de alto impacto o de altas consecuencias (Norman et al., 2002).

Una evaluación de alto impacto, «se indica cuando los resultados del instrumento tienen consecuencias importantes para las personas o las instituciones; por ejemplo, en los procesos de admisión o certificación», según el Instituto Nacional para la Evaluación de la Educación en México (Instituto Nacional para la Evaluación de la Educación, 2017). Por otro lado, los *Estándares para pruebas educativas y psicológicas* la definen como "Prueba usada para obtener resultados que tienen consecuencias directas y significativas para las personas, programas o instituciones que participan en la prueba” (American Educational Research Association et al., 2018).

Debido a lo anterior, es menester asegurarnos de que estas pruebas estén bien hechas, sean bien aplicadas, y que los resultados de los sustentantes se interpreten y se utilicen adecuadamente. Para evaluar la calidad de los procesos de admisión se deben considerar tres dimensiones por separado: validez, confiabilidad y justicia; aunque algunos autores incluyen la justicia y la confiabilidad como componentes de la validez (Lane et al., 2016, pp. 17 y 269).

Los *Estándares para pruebas educativas y psicológicas* (American Educational Research Association et al., 2018) indican que la validez es una consideración fundamental durante el proceso de desarrollo y evaluación de la interpretación y uso de los resultados de las pruebas, ya sean de alto o de bajo impacto. En ese documento se explica que la validez es el juicio acerca del grado en que la evidencia empírica y las razones teóricas que se presentan apoyan o refutan lo apropiado o adecuado de la interpretación, así como de los usos que se dan a los resultados de una evaluación. Los dos marcos de referencia de validez más conocidos, el de Messick y el de Kane, proveen de mecanismos para evaluar el grado de validez de la interpretación y de los usos de los resultados de las evaluaciones; sin embargo, sus aportaciones suelen utilizarse de manera parcial. Por otro lado, existe la dificultad de aplicar estos marcos de referencia a la vida cotidiana de las universidades, en donde serían de gran ayuda como parte de la evaluación de pruebas de altas consecuencias, como los procesos de admisión.

No es común encontrar fuentes de evidencia de validez publicadas con respecto a las evaluaciones en general en las facultades y escuelas de medicina, aunque es sumamente importante determinarlas en los exámenes de altas consecuencias como los exámenes y procesos de admisión a estas instituciones. Por otro lado, las publicaciones al respecto no siempre incluyen todas las fuentes de validez, pues en general se favorece únicamente a la validez predictiva, además de que la mayoría son estudios que se han realizado en países como Estados Unidos, Canadá, Reino Unido y Australia. Tampoco se ha demostrado en alguna publicación la validez de los usos e interpretaciones de los resultados de evaluaciones de esta naturaleza en nuestro país, incluyendo a la facultad de medicina de la Universidad Autónoma de San Luis Potosí. Por lo anterior, es importante determinar cómo adaptar estos marcos de referencia para llevar a cabo la validación en nuestro medio, lo que podría extrapolarse a los procesos de admisión que se realizan en el resto de las escuelas de medicina de nuestro país y latinoamérica.

Con base en lo anterior, este trabajo contiene los antecedentes del problema que se ha descrito brevemente; después, se describen los marcos de referencia que hemos mencionado, el de Messick y el de Kane, así como otros conceptos relacionados con la validez. En el marco metodológico se explican la justificación, los objetivos de este proyecto y el tipo de estudio que se realizó, así como las consideraciones éticas y financieras pertinentes. Posteriormente, en el apartado de los resultados, se elabora la propuesta del método, que incluye las etapas de investigación de validez de Russell, recientemente descritas en la literatura, y la aplicación del mismo con los datos de los procesos de admisión que se llevaron a cabo durante los años 2013 y 2014 para la selección de alumnos de la licenciatura en médico cirujano de la Facultad de Medicina de la Universidad Autónoma de San Luis Potosí.

3. ANTECEDENTES

3.1. Situación actual de los procesos de admisión a la licenciatura en médico cirujano.

3.1.1. Características de los exámenes de admisión

Las instituciones educativas en ciencias de la salud deciden de manera individual cuál es la mejor forma de seleccionar a los alumnos que recibirá entre los aspirantes que desean participar en sus programas académicos. Esto depende del contexto legal nacional, las características de ingreso y de egreso del programa y de la sociedad en la que se encuentran estas instituciones. Para guiar los procesos de selección se han elaborado informes en los que se detallan los conocimientos previos que deben tener los aspirantes para ingresar a la licenciatura en médico cirujano (Association of American Medical Colleges & Howard Hughes Medical Institute, 2009; Finnerty, 2010), mismos que son evaluados en los exámenes de admisión por medio de diferentes procesos cuyo enfoque puede ser uno de los siguientes:

1. Individual.- El ser aceptado depende del éxito académico del sustentante durante las pruebas de selección que evalúan los conocimientos previos, como el MCAT (Medical College Admission Test), el UKCAT (United Kingdom Clinical Aptitude Test) y el GAMSAT (Graduate Australian Medical School Admissions Test).
2. Basado en competencias.- Se evalúan las conductas y actitudes que se considera que podrían indicar el éxito como estudiante o profesionista a través de métodos como la mini entrevista múltiple (multiple mini-interviews – MMI) o las pruebas de juicio situacional (Situational Judgement Testing – SJT).
3. En la sociedad.- Toma en cuenta lo que espera la sociedad de los profesionistas que forman las instituciones educativas, así como las competencias individuales que deben poseer estos egresados (Patterson et al., 2018).

Con respecto de los procesos cognitivos que evalúan las pruebas de admisión, se distinguen dos tipos de inteligencia: fluida (IF) y cristalizada (IC). La primera se refiere a las habilidades inductivas y deductivas involucradas en el mantenimiento de la información, como la capacidad de razonamiento lógico y habilidades de procesamiento, y el razonamiento verbal o visual – espacial en la memoria de trabajo; permite planeación y llevar a cabo conductas dirigidas a alcanzar objetivos. Por otro lado, la IC se refiere a las habilidades de conocimiento adquiridas a lo largo del tiempo y a través de la educación, por ejemplo conocimientos de ciencias biológicas y físicas (Alfonso et al., 2005; Blair, 2006; McManus et al., 2011). Estos dos tipos de inteligencia suelen evaluarse por separado en varios de los procesos de admisión a las escuelas de medicina.

3.1.2. MÉTODOS DE SELECCIÓN PARA ADMISIÓN EN LAS ESCUELAS DE MEDICINA.

Los mejores métodos de selección toman en cuenta la filosofía de la universidad o institución educativa, el contexto de la sociedad a la que proporcionarán sus servicios los profesionales en salud, y el sistema de salud en el que se insertarán estos egresados (Patterson et al., 2018). Con base en lo anterior, para elegir el mejor método se deben identificar en primer lugar cuáles son las características y habilidades que se esperan del estudiante que ingresa a la licenciatura en medicina, del médico interno, y del médico graduado que transiten por la institución de la que se trate (Patterson, Prescott-Clements, et al., 2016). Aclarados estos objetivos, será más sencillo determinar qué se necesita para ingresar a la licenciatura en médico cirujano y cómo se debe evaluar. Para elegir a los que considera a los mejores aspirantes cada institución educativa utiliza uno o varios métodos entre los que se encuentran los siguientes (Kelly et al., 2018; Lin et al., 2022; Patterson et al., 2018; Patterson, Prescott-Clements, et al., 2016; Patterson, Zibarras, et al., 2016):

1. **Desempeño académico.**

Se refiere a las calificaciones obtenidas en el nivel académico previo. Es cuestionable qué tanto se puede comparar entre los resultados de cada aspirante debido a las diferencias entre los sistemas educativos y la selectividad social. Aunado a esto, el poder de discriminación va disminuyendo, pues cada vez más estudiantes obtienen calificaciones más altas en el bachillerato. A pesar de lo anterior, la evidencia internacional parece indicar que los candidatos que han sido admitidos en las escuelas de medicina con base en este método suelen tener menor deserción que otros.

2. **Pruebas de aptitud.**

Se consideran el método más útil por parte de los seleccionadores, aunque la evidencia con respecto de su validez predictiva no es clara, además de que algunos aspirantes y alumnos opinan que son pruebas poco transparentes, no justas, difíciles e incluso dudan de su relevancia. Por otro lado, existe evidencia que sugiere que los estudiantes seleccionados por medio de este método pueden ser más capaces y más motivados para estudiar medicina que quienes no llevaron a cabo este proceso.

3. **Declaraciones personales y currículum vitae.**

Este método podría ser útil para los sustentantes para que se den cuenta de las características de la carrera a la que desean ingresar y así tomar una decisión informada. No obstante, los análisis internacionales han demostrado poca validez predictiva, además de que es un método altamente susceptible de plagio y se podría alterar el documento por medio de asesorías externas. También se debe tomar en cuenta la presencia de sustentantes que "estiran la verdad", viéndolo como algo necesario para "entrar en el juego".

4. **Cartas de recomendación y de referencia.**

Aunque a los candidatos les parece un método recomendable, no existen pruebas de que en efecto sea útil para la selección. En ocasiones podrían ser útiles cuando se describen actitudes y comportamientos específicos del sustentante, y más aún si provienen de médicos conocidos por el examinador.

5. **Pruebas de juicio situacional.**

En el caso de grandes cantidades de sustentantes son útiles y, aunque al principio se perciban costosas, a la larga no lo son. Los sustentantes las perciben como pruebas con mayor validez que las pruebas de aptitud académica, además de que por medio de un buen diseño se puede superar el reto de los candidatos que han recibido entrenamiento previo. También ha demostrado ser un método confiable y válido para la selección de estudiantes de medicina.

6. **Pruebas de personalidad**

No se ha encontrado evidencia clara acerca de cuáles son las características de personalidad que se relacionen con mejores resultados durante la carrera; sin embargo, parecen ser útiles al momento de hacer preguntas más dirigidas durante las entrevistas. Las características de personalidad que podrían predecir el desempeño académico son minuciosidad, extroversión, amabilidad, franqueza y neuroticismo. El grado de validez de los estudios de personalidad va en aumento conforme el estudiante avanza en la carrera y se vuelve más clínico.

7. **Entrevistas y MMI.**

Las entrevistas no estructuradas no son tan útiles y carecen de validez predictiva, a menos que se unan con las pruebas de personalidad. Las MMI son las entrevistas más estructuradas que hay, pues cuentan con seis o más estaciones estandarizadas de entrevistas, lo que incrementa su confiabilidad aunque se puede presentar sesgo del evaluador. También son métodos onerosos por el diseño y la implementación, pero ofrecen validez predictiva.

8. **Centros de selección que utilizan muestras de trabajo y simulación**

Los centros de selección con estaciones múltiples son caros por el diseño y la implementación, pues requieren de varias simulaciones. Las evidencias muestran su utilidad en la selección de posgrado, mas no hay publicaciones al respecto de la selección en pregrado.

3.1.3. Los procesos de admisión para la licenciatura en medicina en el mundo y sus análisis de validez.

Estados Unidos y Canadá

En Estados Unidos y en Canadá se utiliza el MCAT. Anteriormente, esta prueba evaluaba razonamiento verbal, ciencias biológicas, ciencias físicas y escritura por medio de una evaluación de opción múltiple. Julian (2005), investigó la relación entre los GPAs y los resultados del MCAT con las calificaciones en la escuela de medicina, los resultados de los Step 1, 2 y 3 del USMLE (United States Medical Licensing Examinations) y las distinciones o dificultades académicas, encontrando que ambas pruebas de admisión son buenos predictores de desempeño académico; sin embargo, el MCAT ha demostrado ser mejor predictor, específicamente la parte de razonamiento verbal. Se ha observado que en general estos predictores eran efectivos al inicio de la carrera y su efecto se va desvaneciendo conforme pasa el tiempo; sin embargo, el segmento de razonamiento verbal no solo no disminuye, sino que aumenta con el tiempo (Violato & Donnon, 2005).

A partir de 2015, se implementó una nueva versión del MCAT que evalúa la IC y la IF a través de cuatro secciones: fundamentos biológicos y bioquímicos de los sistemas vivientes, fundamentos químicos y físicos de los sistemas vivientes, fundamentos psicológicos, sociales y biológicos de la conducta, y habilidades de razonamiento y análisis crítico. Las primeras tres secciones prueban diez conceptos fundamentales y cuatro habilidades de investigación científica y razonamiento, mientras que la cuarta sección comprueba qué tan bien comprenden, analizan y evalúan lo que leen los aspirantes, las inferencias que sacan a partir del texto, y cómo aplican estos argumentos a nuevas situaciones. Las calificaciones se reportan en un rango de 118 a 132 para cada sección, con un promedio de 125; en conjunto el rango es de 472 a 528, con un promedio de 500 (Association of American Medical Colleges, 2020).

En varias universidades se combina el resultado de esta prueba con el de una entrevista y el promedio preuniversitario GPA (Grade Point Average). Este último suele estandarizarse debido a que puede variar el rango entre bachilleratos de Estados Unidos y de Canadá (Association of American Medical Colleges, 2020, 2023).

EUROPA

Reino Unido.

Ferguson et al. (2002; McManus et al., 2011), realizaron una revisión sistemática de la literatura en la que examinaron los datos acerca de la validez predictiva de los criterios de admisión utilizados en el Reino Unido: factores cognitivos (habilidad académica previa), factores no cognitivos (personalidad, estilos de aprendizaje, entrevistas, referencias, declaraciones personales) y factores demográficos (sexo y edad). En su estudio encontraron que la habilidad cognitiva era un predictor moderado de éxito y que los estudios acerca de los estilos de aprendizaje pueden ser útiles en el futuro.

El BMAT (Biomedical Admissions Test) era el examen que se aplicaba hasta 2006 en Reino Unido y otros países de la Commonwealth con dos secciones: la primera, de aptitud y habilidad, evaluaba resolución de problemas, comprensión de argumentos, e interpretación de datos y gráficas. La segunda parte evaluaba el conocimiento científico y su aplicación, verificando los conocimientos sobre biología, química, física y matemáticas. McManus et al. (2011), analizaron el valor predictivo de ambas partes con el desempeño académico en el primero y el segundo año de medicina, y encontraron que la IC evaluada por medio de los conocimientos sobre biología, química, física y matemáticas tiene un índice de correlación mayor que la IF en ambos años (primer año: IC r=036, IF=0.19, p<0.0001; segundo año: IC r=0.23, IF=0.15, p=0.017).

Actualmente las universidades en el Reino Unido requieren altas calificaciones en el nivel previo de estudios y una entrevista en la universidad de elección para iniciar el proceso de admisión. Estas pruebas se evalúan junto con los resultados del UKCAT, examen que realiza a nivel nacional desde 2006 y que valora el razonamiento verbal, cuantitativo y abstracto, así como la decisión por análisis. Las universidades revisan toda esta información y deciden quiénes pueden ingresar a la carrera. Poco después de implementar el UKCAT se analizó su valor predictivo sobre el desempeño académico del primer año en medicina, encontrando que no existe correlación (Lynch et al., 2009).

<u>Alemania</u>.

En Alemania se consideran que las habilidades cognitivas poseen una capacidad de predicción muy limitada, pues se da por sentado la buena calidad de la educación pre-universitaria de los aspirantes, así que la variación en el éxito académico debe estar determinado por otros factores, tales como la motivación, consciencia y estabilidad emocional. Las escuelas de medicina pueden seleccionar al 60% de sus estudiantes al tomar en cuenta los logros en el bachillerato y exámenes optativos, como el Test für Medizinische Studiengänge, semejante al UKCAT, y el HAM-Nat, que evalúa los conocimientos en física, química y biología. El 40% restante se admite de manera proporcional entre quienes poseen calificaciones excelentes en el bachillerato, están en lista de espera, estudiantes internacionales y otros (por ejemplo, desfavorecidos). Además de los exámenes mencionados, también puede aplicarse una MMI llamada HAM-Int como prueba de competencia social. Meyer et al. (2019), realizaron un análisis de regresión para conocer la relación predictiva del proceso de admisión sobre el desempeño académico de estos estudiantes, y observaron que los alumnos con mejor desempeño eran quienes ingresaron por medio del examen de admisión o los de la proporción de calificaciones excelentes en el bachillerato.

Países Bajos.

Entre los procesos que se aplican está la lotería, en que la posibilidad de ser admitido es mayor de manera proporcional al GPA, MMIs o exámenes cognitivos. Schreurs et al. (2018), hicieron un análisis de costo-beneficio de la lotería comparada con el proceso de admisión para una universidad de Maastricht. El proceso se llevó a cabo en dos pasos: elaborar un portafolio con la información del GPA, los logros extracurriculares, una declaración personal y una autoevaluación en cuanto a su capacidad para el aprendizaje basado en problemas. Quienes obtuvieron las mejores calificaciones continuaron al segundo paso, con pruebas de juicio de situación y un examen escrito. Al final, se seleccionó a los alumnos con el mejor desempeño en estas pruebas con base en el número de lugares disponibles. Se encontró que es mejor llevar a cabo un proceso de admisión bien estructurado y rinde más, económicamente hablando, que la lotería, por los alumnos que abandonan la carrera (las universidades reciben dinero por cada alumno inscrito) o tienen que repetir materias (hay que pagar a los profesores para impartirlas de nuevo, así como mantener la infraestructura para dar estas clases extra).

Asia

Arabia Saudita.

Se aplican tres exámenes en tres momentos diferentes: el primero es un examen nacional que aplica la Secretaría de Educación al terminar el primer semestre del último año de preparatoria para quienes están llevando química, bioquímica, matemáticas, biología, árabe, inglés y religión; esta evaluación cuenta 35% de la calificación final. Posteriormente, presentan el Examen Nacional de Desempeño de Arabia Saudita al finalizar la preparatoria, que evalúa inglés, biología, química, física y matemáticas, proveyendo el 35% de la calificación final. Además de estas pruebas, el Examen Nacional de Aptitud de Arabia Saudita aporta el 30% de la calificación final, y posee una sección de lingüística y una de matemáticas. Al Alwan et al. (2013), evaluaron la validez predictiva de estas pruebas sobre el desempeño académico de sus alumnos, y encontraron una correlación positiva moderada ($r=0.65$ a 0.66, $p<0.05$) de las tres pruebas con los resultados de hasta tres años en la carrera de Medicina.

Jordania.

Tienen varias vías de entrada a la escuela de medicina: el Proceso Nacional Abierto de Admisión incluye un examen nacional de competencia; la proporción para estudiantes desfavorecidos, principalmente hijos de personal de servicio en la milicia o de empleados del Ministerio de educación; los hijos de empleados universitarios; el proceso paralelo, para sustentantes con calificaciones bajas que deben pagar colegiaturas más altas; la proporción para estudiantes internacionales, quienes también deben pagar altas cuotas de colegiatura; y otros, como estudiantes internacionales que ingresaron por medio de convenios o estudiantes de estrato socioeconómico bajo y de áreas remotas del país. Tamimi et al. (2023), compararon las diferentes vías de ingreso y el desempeño académico de los estudiantes, y observaron que quienes tenían más probabilidades de graduarse a tiempo y con mejores calificaciones eran los admitidos a través de la competencia abierta y los estudiantes desfavorecidos.

De los siguientes siete países solo se cuenta con información acerca del proceso de admisión, pero no de estudios de validación (Soemantri et al., 2020):

Indonesia.

El Ministerio de Educación y Cultura regula el proceso de admisión a través de dos procesos posibles: en uno, los aspirantes son invitados a ingresar a las escuelas de medicina públicas con base en sus calificaciones de bachillerato; en el otro, los aspirantes presentan un examen nacional de admisión que mide la aptitud académica. Además, las mejores universidades públicas también pueden aplicar sus propios métodos de admisión, como MMI y pruebas de juicio de situaciones. Por otro lado, las universidades privadas toman en cuenta el promedio del bachillerato y llevan a cabo exámenes de aptitud académica y entrevistas.

Japón.

El Ministerio de Educación, Cultura, Deportes, Ciencia y Tecnología regula el proceso de admisión a las 82 escuelas de medicina en Japón. El Centro Nacional de Evaluación aplica una prueba que evalúa inglés, japonés, historia, geografía, matemáticas, física, química y biología. Con base en este resultado, cada universidad hace sus propias evaluaciones, como entrevistas. El proceso en general está en evaluación para incluir aspectos no cognitivos y para atender los reportes de manipulación de los resultados para admitir a más hombres que a mujeres.

Malasia.

El Consejo Médico Malayo determina los requisitos de admisión que incluyen un diploma de educación básica en ciencia y matemáticas con calificación de B en casi todas las materias, y un diploma de bachillerato con calificación de BBB en biología, química y físico-matemáticas, además de un GPA de 3.5/4 y haber aprobado inglés, matemáticas y otra asignatura. No se explica claramente si existen evaluaciones de aspectos no cognitivos aunque algunas escuelas utilizan las MMI.

Filipinas.

Quienes deseen ingresar a una escuela de medicina en Filipinas deben poseer el diploma de bachillerato y presentar el Examen Nacional de Admisión a Medicina, que incluye aspectos cognitivos y no cognitivos; el aspirante debe alcanzar al menos la percentila 40 en esta evaluación que consta de dos partes. La primera evalúa razonamiento verbal, razonamiento inductivo, habilidades cuantitativas, y habilidades de agudeza de percepción. La segunda evalúa biología, física, ciencias sociales y química. Suelen utilizarse las MMI para valorar los aspectos no cognitivos como la motivación, integridad, conciencia social y la tolerancia al estrés.

Singapur.

La Universidad Nacional de Singapur lleva a cabo evaluaciones de aspectos cognitivos y no cognitivos, tales como MMI, pruebas de juicio de situaciones, declaraciones personales, logros curriculares y extracurriculares, el Examen de Admisión Biomédica, criterios de logros no académicos y reportes de árbitros. Las otras dos escuelas de medicina también incluyen las calificaciones de bachillerato.

Sri Lanka.

En Sri Lanka se busca la admisión con base en los principios de mérito y equidad, por lo que todos los aspirantes presentan el mismo examen que evalúa tres asignaturas de ciencias. Los aspirantes que hayan obtenido las calificaciones más altas en biología pueden ocupar el 40% de los lugares disponibles. El 55% de los lugares disponibles se distribuyen entre los distritos administrativos del país de forma proporcional al promedio de población en cada uno. El resto, 5%, se distribuyen en 16 distritos con "desventaja educativa".

Taiwan.

Desde 1994 se utiliza el Examen de Habilidad Académica General (General Scholastic Ability Test – GSAT), y solo quienes alcanzan la percentila 93 pueden obtener una entrevista de selección. Sin embargo, en cada escuela aplican diferentes métodos que pueden incluir cartas de recomendación del bachillerato o entrevistas de tipo MMI.

Oceanía

En 2013 se realizó un análisis acerca del valor predictivo del Undergraduate Medical and Health Sciences Admissions Test (UMAT) y promedio del bachillerato con o sin entrevista (ya fuera no estructurada o MMI) en Australia en tres instituciones, en el que se llevaron a cabo correlaciones con r de Pearson, así como regresión múltiple. Se encontró que en una de las instituciones cada uno de los componentes del proceso de admisión contribuye a explicar la varianza de los resultados de manera independiente de los otros componentes, por lo que consideran que aumenta la validez predictiva de su proceso. Estos investigadores solo se enfocan en el análisis de la validez predictiva, sin considerar realmente los marcos de referencia de validez en toda su extensión y sin llevar a cabo un análisis sistemático (Edwards et al., 2013). Por otro lado, Shulruf et al. (2018), evaluaron la eficacia de los métodos de selección: el UMAT, las calificaciones del bachillerato y la entrevista de selección sobre el desempeño académico durante la carrera y encontraron que los dos últimos tenían mayor poder predictivo que el UMAT.

En 2020, en Australia y Nueva Zelanda se cambió el examen de admisión del UMAT al University Clinical Aptitude Test (UCAT), basado en el UKCAT. El UCAT consiste en una evaluación de cuatro secciones con preguntas de opción múltiple (POM) en donde se evalúan habilidades cognitivas: razonamiento verbal, toma de decisiones, razonamiento cuantitativo y razonamiento abstracto. La quinta sección es una prueba de juicio de situación (Griffin et al., 2021).

3.1.4. El examen de admisión para la licenciatura en medicina en México y sus análisis de validez.

El Consejo Mexicano para la Acreditación de la Educación Médica (COMAEM) lleva a cabo la acreditación de los programas de formación médica que se imparten en las facultades y escuelas de medicina de México. En este proceso se siguen lineamientos específicos para verificar que cuentan con las características necesarias para ofrecer un programa académico de calidad para quienes cursen los estudios de Medicina General. Uno de los rubros que evalúan es la existencia de un sistema de admisión establecido, con manuales de operación, con base en normativa y que tenga relación con el programa académico correspondiente; sin embargo, como parte del proceso de acreditación no pueden imponer un método de admisión, sino solo hacer sugerencias al respecto (Consejo Mexicano para la Acreditación de la Educación Médica (COMAEM), 2018).

En el sitio web de este consejo se encuentra un listado de todas las facultades y escuelas que ofrecen esta licenciatura, explicando cuáles son públicas, cuáles son privadas y cuál es su estado de acreditación. En noviembre de 2023, se encontraron 162 escuelas de medicina en total, de las cuales, 58.64% están acreditadas (Consejo Mexicano para la Acreditación de la Educación Médica (COMAEM), 2023).

Las escuelas y facultades de medicina mexicanas utilizan uno o varios de los siguientes métodos para sus procesos de admisión:

1. Examen de conocimientos elaborado por la misma universidad.
2. Examen Nacional de Ingreso a la Educación Superior (EXANI-II) del CENEVAL.
3. Examen de habilidades y conocimientos básicos - EXHCOBA.
4. Examen psicométrico.
5. Entrevista.
6. Curso propedéutico.
7. Promedio del bachillerato.
8. Ninguno.

El Centro Nacional de Evaluación para la Educación Superior, A.C. (CENEVAL) es el órgano que se encarga de desarrollar el Examen Nacional de Ingreso a la Educación Superior (EXANI-II), el método que más se utiliza en nuestro país. Este examen estaba integrado por dos pruebas: EXANI-II de selección, que evaluaba razonamiento verbal, razonamiento matemático, español y matemáticas, y EXANI-II de diagnóstico opcional, sobre áreas temáticas relacionadas con la carrera a la que se desea ingresar. Los resultados se reportaban por medio del "índice Ceneval", con una escala de 700 a 130 puntos, con promedio de 1,000 (Centro Nacional de Evaluación para la Educación Superior, 2013a).

Martínez Villarreal (2013), analizó el valor predictivo del EXANI-II del CENEVAL como único parámetro de admisión para ingresar a la Universidad Autónoma de Nuevo León sobre el desempeño de los alumnos de medicina en los primeros dos años en siete generaciones, y encontró una correlación positiva entre las calificaciones de los tres primeros años de la carrera y los resultados del EXANI-II ($r=0.43$, $R^2=0.18$, no se reporta p). Por otro lado, en la Universidad del Mayab, en Yucatán, el proceso de admisión contempla el promedio de bachillerato, las calificaciones de un curso propedéutico, el resultado de la Prueba de Aptitudes Académicas del College Board y el resultado del EXANI-II. García Domínguez (2016), analizó la correlación de estas evaluaciones con el desempeño académico de los alumnos en medicina de dos generaciones durante el primer semestre y durante los tres primeros semestres, y encontró que el valor que mejor correlaciona es el promedio del curso propedéutico (2012: $r= 0.63$, $R^2=0.4$, $p=0.0$; 2013: $r= 0.60$, $R^2=0.36$, $p=0.0$).

A partir de 2021 se aplica una versión nueva de esta prueba que evalúa comprensión lectora, redacción indirecta y pensamiento matemático, inglés como lengua extranjera en un nivel B1, y habilidades socioemocionales. También valora los conocimientos específicos por carrera, que en el caso de medicina incluyen salud pública y medicina comunitaria, anatomía y fisiología, biología celular y microbiología, y bioquímica y biología molecular (Centro Nacional de Evaluación para la Educación Superior, 2023).

La Facultad de Medicina de la Universidad Nacional Autónoma de México (UNAM) es la escuela de medicina más grande en nuestro país. En 2018, se presentaron 11,198 aspirantes para ingresar, quienes habían cursado sus estudios de bachillerato en preparatorias que no pertenecen a la UNAM, y fueron seleccionados 162 alumnos. Estos aspirantes presentaron un examen de 120 reactivos de opción múltiple, con una mínima de 108 aciertos. En este examen se evalúan conocimientos en Matemáticas, Física, Química, Biología, Geografía, Historia Universal, Historia de México y Literatura (DGAE UNAM, 2018). El resto de los alumnos admitidos a la licenciatura ingresa por medio de pase reglamentado debido a que realizaron los estudios de bachillerato en Colegios de Ciencias y Humanidades o la Escuela Nacional Preparatoria y poseen promedio mínimo de 9.00 (DGAE UNAM, 2017). A todos los alumnos de nuevo ingreso a la UNAM se les aplica un examen diagnóstico para conocer su nivel de conocimientos basal, mismo que evalúa los temas de Matemáticas, Física, Química, Biología, Geografía, Historia Universal, Historia de México y Literatura. En un análisis de seis ciclos escolares (2006 a 2011), y al comparar los resultados entre el origen del alumno (Escuela Nacional Preparatoria, Colegio de Ciencias y Humanidades y bachillerato no UNAM), se encontró que existe correlación positiva entre los resultados de este examen y el desempeño académico durante el primer año de la licenciatura en médico cirujano. La correlación más fuerte fue entre las calificaciones del examen diagnóstico con las calificaciones de primer año de los alumnos provenientes de bachilleratos no UNAM (promedio de correlaciones de las asignaturas con $r=0.69$, no se reporta p) (Muñoz-Comonfort et al., 2014).

En la Facultad de Medicina de la Universidad Autónoma de San Luis Potosí (FMUASLP) se ha analizado el valor predictivo del examen de admisión únicamente sobre las ciencias morfológicas con base en los resultados de dos generaciones de estudiantes, por medio de una prueba de rangos de Spearman, cuyo resultado en ambas generaciones fue una relación moderada entre el lugar de ingreso y el lugar al primer año en ciencias morfológicas (2013: $r=0.37$, $p<0.001$; 2014: $r=0.38$, $p<0.001$) (Carrillo Avalos et al., 2016). Por su parte, Espinoza del Río (2017), analizó como parte de su tesis de maestría si existía relación entre el examen de admisión y el desempeño durante el primer año de la licenciatura en Médico Cirujano en la FMUASLP, al tomar el examen de admisión como un todo con sus tres componentes a la vez. Encontró una correlación entre la calificación del examen de admisión y el promedio del primer semestre de -0.22 ($p=0.013$), -3507 ($p<0.0001$) con el promedio del segundo semestre, y de -0.3162 ($p=0.0004$) con el promedio del primer año de la carrera.

3.3. Contexto de la UASLP

3.3.1. Historia del proceso de admisión para la licenciatura en Médico Cirujano en la UASLP

El Dr. José Miguel Torre López fue director de la FMUASLP de 1959 a 1967, y durante este tiempo ocurrieron varios eventos de importancia para nuestra institución. Uno de ellos fue que al inicio de su gestión fundó el Boletín Informativo de la FMUASLP, una edición trimestral que sigue editándose hasta la fecha. Otro fue que se inauguró el nuevo edificio de la Escuela de Medicina (Alcocer Andalón, 1976), el 2 de diciembre de 1963. Otro más fue el inicio de los exámenes de selección de alumnos para ingresar a la licenciatura en medicina.

¿Cómo se decidía el ingreso de los estudiantes antes de estas pruebas?

El 29 de enero de 1877 se publicó en el periódico oficial el aviso de inicio de la cátedra de primer año de medicina en el Instituto Científico y Literario de la ciudad de San Luis Potosí. Atendiendo a este aviso, se inscribieron cuatro alumnos para iniciar clases el 15 de febrero de 1877; tres terminaron el primer año y solo dos se graduaron (Alcocer Andalón, 1976). De esta manera, quien tuviera los medios y cumpliera los requisitos, solo tenía que acudir a inscribirse sin mayor trámite.

La fundación del Boletín Informativo de la Escuela de Medicina nos permite estar enterados de cómo se han ido desarrollando diversos eventos y situaciones en nuestra institución, pues de ello dan fe los principales involucrados de manera directa. En estas publicaciones empezaron a consignarse, entre otros temas, los esfuerzos realizados hacia la mejora de la selección de alumnos a lo largo del tiempo. El Dr. Torre explicó en el boletín del 16 de febrero de 1960 la necesidad de llevar a cabo un proceso de selección para el ingreso a la entonces Escuela de Medicina. Además, también habló acerca de las razones para limitar el número de estudiantes, entre las que mencionó el número de camas del Hospital Universitario en donde se llevaban a cabo las prácticas clínicas, el número de cadáveres para la enseñanza de Anatomía, y el espacio de laboratorios en la Escuela de Medicina, así como la eficiencia terminal de generaciones anteriores, la que era de 22%. De esta manera se había destinado la mayor parte del presupuesto de la escuela en alumnos que no terminaron la carrera. Con base en estas y otras razones, el Consejo Universitario estableció el número de alumnos de nuevo ingreso en 75 por año (Torre López, 1960b).

Así, se inició la elaboración del instructivo para admisión de alumnos, mismo que fue publicado en julio de 1960 para la selección de los alumnos que iniciarían el primer año de la carrera el siguiente ciclo escolar. Los requisitos de admisión incluían presentar el certificado de calificaciones de secundaria (aportando un valor de 20 puntos), el certificado de calificaciones de bachillerato (40 puntos), así como presentar dos pruebas escritas sobre "conocimientos indispensables para cursar los estudios de Medicina", someterse a un examen psicométrico y sostener una entrevista personal; estos tres últimos elementos conformaban el examen de admisión y aportaban los restantes 40 puntos para un máximo de 100 (Torre López, 1960a). A pesar de las instrucciones, al final se decidió tomar en cuenta solamente tres factores para determinar la selección de los nuevos estudiantes: el promedio del bachillerato, el examen de conocimientos (de matemáticas y biología) junto con la prueba psicométrica, y la entrevista personal. Al año siguiente se decidió que los elementos a considerar para el ingreso de los nuevos alumnos serían el promedio de la secundaria (20 puntos), el promedio de las calificaciones del bachillerato (40 puntos), y examen de conocimientos más examen psicométrico y entrevista personal (40 puntos). Además, se determinó que los alumnos seleccionados serían quienes obtuvieran un mínimo de 60 puntos, por eso en ese año fueron seleccionados solo 64 estudiantes. Por lo anterior, el director consideraba que más que pruebas de admisión o selección, lo que se aplicaba eran pruebas de eliminación de candidatos (Torre López, 1962). Además de conformarse como el primer examen de admisión para ingresar a cualquier escuela o facultad de la UASLP, también fue el primero de este tipo que se llevó a cabo en una escuela de medicina en el país (Leiva Garza, 2003), sentando un precedente importante para todas las escuelas y facultades de medicina mexicanas.

Cuando el Dr. Torre publicó los resultados del proceso de selección de 1962, también hizo un ejercicio de análisis predictivo comparando con las calificaciones al final del primer año. Concluyó que la prueba escrita era el componente que mayor grado de predictibilidad parecía tener con respecto a la trayectoria académica, mientras que los que tenían menor grado eran el promedio del bachillerato, la entrevista personal y el examen psicométrico (Torre López, 1963b).

En 1963 se decidió eliminar las entrevistas personales debido a que ya contaban con la prueba psicométrica y a que la escuela no poseía el número suficiente de profesionales para llevar a cabo las entrevistas a todos los candidatos en tiempo y forma. Por otro lado, también elaboraron el examen de conocimientos con base en el programa académico vigente del bachillerato de la UASLP, decidiendo evaluar los conocimientos de matemáticas, física, química, biología y un idioma extranjero (inglés o francés). De esta manera, ese año el proceso de selección incluyó el promedio de secundaria, el promedio del bachillerato, prueba de conocimientos y examen psicométrico (Torre López, 1963a).

En 1964, al igual que otros años, el Dr. Torre presentó los resultados del proceso de selección de 1963 junto con la siguiente Tabla (Torre López, 1964):

Año	Número de solicitudes	Número de alumnos que presentaron las pruebas	Número de aceptados	Número de aprobados en todas las materias	Aprobaron alguna o algunas materias	No aprobaron ninguna materia	Abandonaron sus estudios
1960	123	88	77	20 (25%)	33	6	18
1961	109	77	64	24 (57%)	14	10	16
1962	99	86	70	18 (25%)	12	12	6
1963	137	123	66	54 (50%)	8	8	16

Tabla 1. Resultados del proceso de selección de 1960 a 1963.

El Dr. Torre explicó en 1966 que la ponderación de los componentes del proceso de selección era la siguiente: promedio de secundaria 20 puntos, promedio del bachillerato 40 puntos y 40 puntos para la prueba de conocimientos. Se admitía a los alumnos que obtuvieran 60 puntos o más. También comentó que en el caso de que los resultados de la prueba psicométrica pusieran en duda el ingreso de un aspirante, un psicólogo llevaba a cabo una exploración personal (Torre López, 1966).

El Dr. Torre dejó de ser director de la escuela en 1967, y los informes y análisis sobre el proceso de admisión dejaron de publicarse hasta 1970, cuando se llevó a cabo una nueva revisión de este proceso. Se propusieron tres fases para el proceso de selección: I prueba de inteligencia, II capacidad de aprendizaje, y III conocimientos básicos elementales e instrumentales. La fase II sería evaluada por medio de una prueba de conocimientos de acuerdo con los programas académicos de las escuelas preparatorias (200 preguntas), y una evaluación de la capacidad de aprendizaje (80 preguntas), que valoraba la comprensión, abstracción, generalización, atención, memorización deducción, inducción y procesos analógicos para medir la capacidad para el manejo del lenguaje escrito y la capacidad para obtener relaciones numéricas o lenguaje matemático. Los conocimientos básicos eran explorados por medio de 120 preguntas repartidas equitativamente entre los temas de biología, matemáticas, física, química, etimologías e inglés. Con estas características, el nuevo proceso de admisión se llevó a cabo para elegir a 77 alumnos de entre los 315 aspirantes que se presentaron ese año. El estudio que realizaron para determinar la selección de alumnos incluyó el análisis de correlación entre la prueba de conocimientos y la prueba de capacidad de aprendizaje, así como la prueba de Donnaiewsky. De esta manera se eliminaron los promedios de la secundaria y la preparatoria como elementos de selección (Garrocho Sandoval & Torre López, 1970).

En 1975 se conformó la Comisión de Admisión de Alumnos con los doctores Fernando Ávila, Federico Dies, Carlos Garrocho, José J. Macías, Benjamín Moncada y Julio Sepúlveda. También se estableció el reglamento de dicha comisión cuya última versión fue aprobada el 26 de enero de 2009 (Comisión de Admisión de la Facultad de Medicina de la UASLP, 2009), en el que se estableció desde el inicio la obligación de rendir un informe anual.

En el informe de la comisión de admisión del 25 de abril de 1977 se emitió la recomendación de admitir cada año a 100-110 alumnos, debido a que entre 1967 y 1971 el número de nuevos estudiantes aumentó de 75 a 93, en 1972 se admitieron 117, en 1973 fueron 147, al año siguiente fueron 153 y, finalmente, en 1975 aumentó a 181; de estos últimos solo 44.8% consiguió pasar al segundo año de la carrera (Comisión de Admisión de la Escuela de Medicina de la UASLP, 1977b; Zazueta Quirarte, 2003). A pesar de esta recomendación, en 1976 se recibió a 133 estudiantes seleccionados a partir de 618 aspirantes; de estos alumnos 66.1% aprobaron las asignaturas de Ciencias Morfológicas. La comisión realizó un análisis de correlación en ambas generaciones entre los resultados de las pruebas de conocimientos y capacidad de aprendizaje, y los resultados académicos del primer año tomando como medida las calificaciones de las Ciencias Morfológicas, concluyendo que en realidad “no se selecciona a candidatos buenos, sino que se asignan las plazas a los mejores, que resultan ser, en realidad, los menos malos.” (Comisión de Admisión de la Escuela de Medicina de la UASLP, 1977a).

Para el ingreso de la generación 1978-1979 se desarrolló un instrumento de evaluación de aptitudes que permitía valorar la rapidez y retención en la lectura, la comprensión de vocabulario y la capacidad de pensar en forma lógica; este instrumento no fue tomado en cuenta para la admisión de los estudiantes. Sin embargo, la puntuación de la evaluación permitió categorizar a los alumnos en cinco rangos del I al V, tomando en cuenta la desviación estándar obtenida en esa ocasión. Luego, analizaron la probabilidad de admisión de cada alumno con base en el rango de clasificación y el resultado del examen de admisión que en ese momento aportaba 80% de la calificación final, mientras que la evaluación psicométrica aportaba 20%. Encontraron que la probabilidad de admisión iba disminuyendo conforme aumentaba el rango; es decir, a mayor rango, menor probabilidad de ser admitido (por ejemplo, era de 0% en el rango V) (Guerrero M, 1979).

En 1983 la escuela cambió de nombre a Facultad de Medicina y, después del establecimiento de la cantidad de aspirantes seleccionados cada año en 1976 (132), los informes anuales de la comisión de admisión se referían únicamente a la cantidad de aspirantes y sus orígenes. El EXANI-II se empezó a aplicar desde 1996 para seleccionar a los nuevos estudiantes de la FMUASLP, conformándose la siguiente ponderación de los elementos de admisión: examen de conocimientos 45%, examen psicométrico 15%, y EXANI-II 40%. Esta nueva ponderación se aplicó en todas las escuelas y facultades de la Universidad desde 2002 y no ha cambiado desde entonces (García Bonilla & Noyola Bernal, 2001). La última modificación al proceso de admisión se llevó a cabo en 2013, ya que a partir de ese año se han admitido 145 alumnos cada año a nuestra Facultad (Figura 1).

Cabe mencionar que este proceso de admisión a la carrera de médico cirujano es el mismo que se lleva a cabo para seleccionar a los alumnos de la licenciatura en ciencias ambientales y de la salud, misma que inició en 2009. También es importante señalar que los métodos de admisión a una escuela o facultad de medicina conforman un rubro a evaluar durante la acreditación de las mismas por parte del Consejo Mexicano para la Acreditación de la Educación Médica. Este Consejo otorgó la primera acreditación a nuestra Facultad el 17 de mayo de 2002, y se ha seguido obteniendo en cada ocasión, la última vigente hasta 2023. Este es uno de los motivos por lo que es conveniente revisar este tema y mantenerlo actualizado.

Existen algunos análisis al respecto de la capacidad de predicción de este proceso hacia el desempeño académico. Uno de ellos fue publicado en 2001 y escrito por el Dr. Llamas y el Dr. Escalante (Llamas & Escalante, 2001); en este se realizó un análisis de correlación entre los componentes del proceso de admisión de 1999 y las asignaturas de Ciencias Morfológicas. Observaron que las mejores correlaciones se encontraban entre histología e inglés ($r=0.70$, $p<0.01$), embriología y el EXANI-II ($r=0.59$, $p<0.01$), y anatomía e inglés ($r=0.49$, $p<0.01$). Otro análisis interesante fue el desarrollado por el Dr. Espinoza (Espinoza del Río, 2017), en el que, entre otros hallazgos, encontró que para la generación que ingresó en 2014 hacer cursos de preparación no pareció mejorar el rendimiento académico durante el primer año de la carrera; sin embargo, estos cursos, junto al esfuerzo académico durante el tiempo previo a la licenciatura, parecieron mejorar la posibilidad de ingreso.

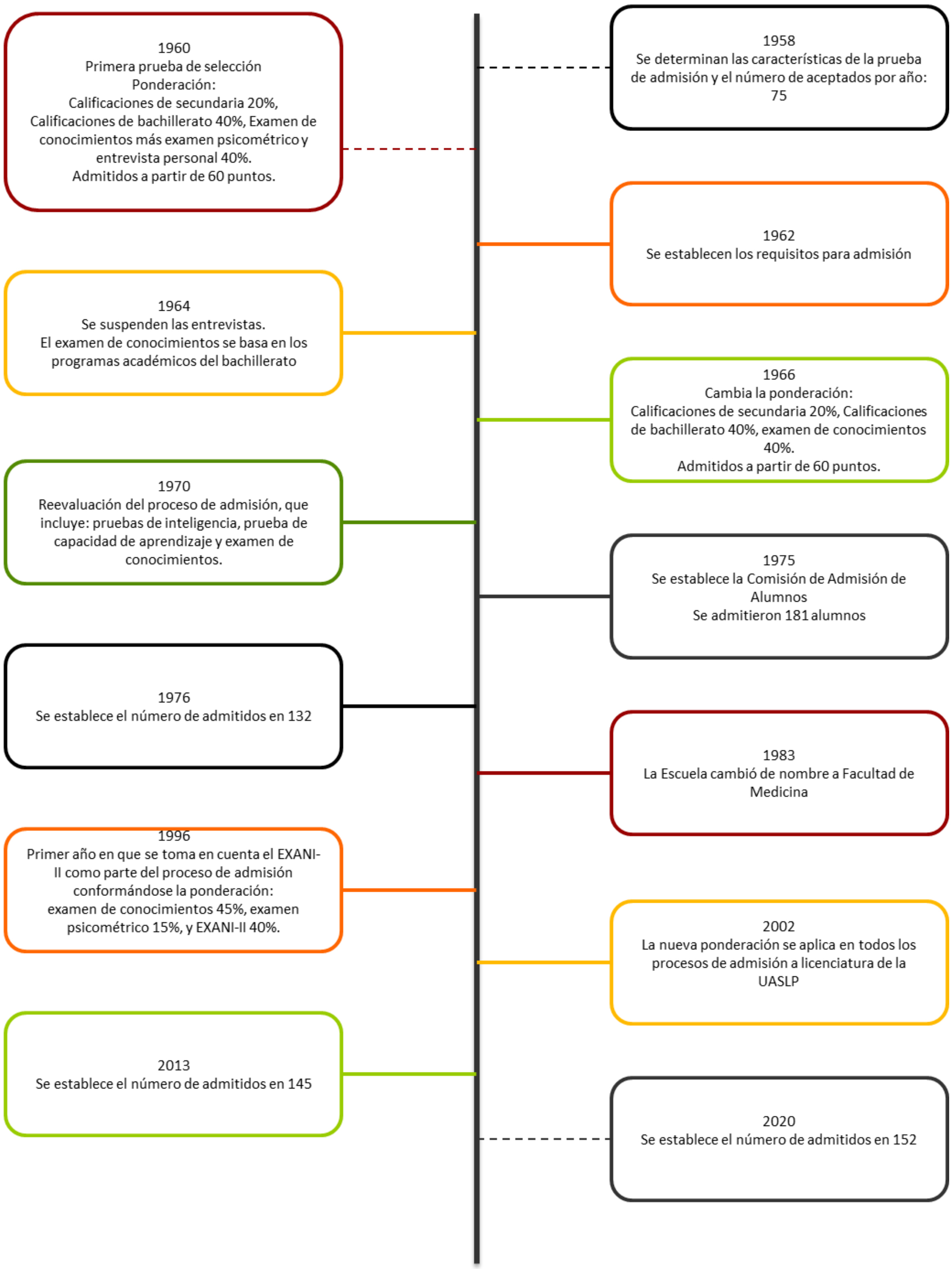


Figura 1. Línea del tiempo del proceso de admisión en la Facultad de Medicina de la UASLP.

3.3.2 Características del plan de estudios de la licenciatura en Médico Cirujano en la UASLP.

El perfil de ingreso publicado en la página web de la FMUASLP se declaró de la siguiente manera para las generaciones que ingresaron hasta 2019:

> “Características Deseables en el Estudiante
>
> ●Demanda en primer término, una tendencia humanística de ayuda al prójimo a través del tratamiento de los padecimientos del hombre, serán físicos o psíquicos, requiere además tenacidad y buenos hábitos de estudio.
> ●Responsabilidad, estabilidad emocional y capacidad para tomar decisiones y adaptarse a situaciones de urgencia. Recurrir no solo a sus conocimientos sino tener una conducta objetiva y serena, ante situaciones problemáticas en el trato con sano y enfermo.
> ●Iniciativa sana y capacidad de persuasión para obtener la confianza del paciente, lo que le permitirá poder guiarlo durante su tratamiento y recuperación.
> ●Carácter, que se reflejará en su respeto a la verdad, a la honradez, a la justicia, a la dignidad humana, al secreto profesional y a los más elevados principios de conducta.
> ●Por los horarios de clase y prácticas, todos obligatorios, debe dedicar tiempo completo a sus estudios.
> ●Es conveniente y a veces necesario, el manejo de varios idiomas, sobre todo el inglés, debido a la información que debe obtener durante su entrenamiento.” (Facultad de Medicina de la UASLP, 2015).

La licenciatura de Médico Cirujano en la FMUASLP contemplaba, hasta el ingreso en 2018, cinco años cursando asignaturas teórico-prácticas, más un año de internado y un año de servicio social (Figura 2). Durante el transcurso de cada asignatura se llevan a cabo los exámenes parciales correspondientes, así como un examen final ordinario, extraordinario o a título de suficiencia, según el promedio final y el resultado del examen final. De no aprobar esta evaluación, el alumno tiene la oportunidad de realizar un examen de regularización hasta en dos ocasiones. Si no obtiene como calificación 6.0 o más, debe ser dado de baja de la Facultad.

Como parte del proceso de titulación, los estudiantes realizan el Examen General de Egreso de la Licenciatura de Médico Cirujano (EGEL-MG) de CENEVAL, y deben aprobarlo con una calificación igual o mayor a 1000 puntos. Quienes lo logren, continuarán con el proceso al presentarse a la evaluación clínica objetiva estructurada (ECOE), la que deben aprobar con calificación mínima de 6.0 (Consejo Técnico Consultivo de la Facultad de Medicina de la UASLP, 2018).

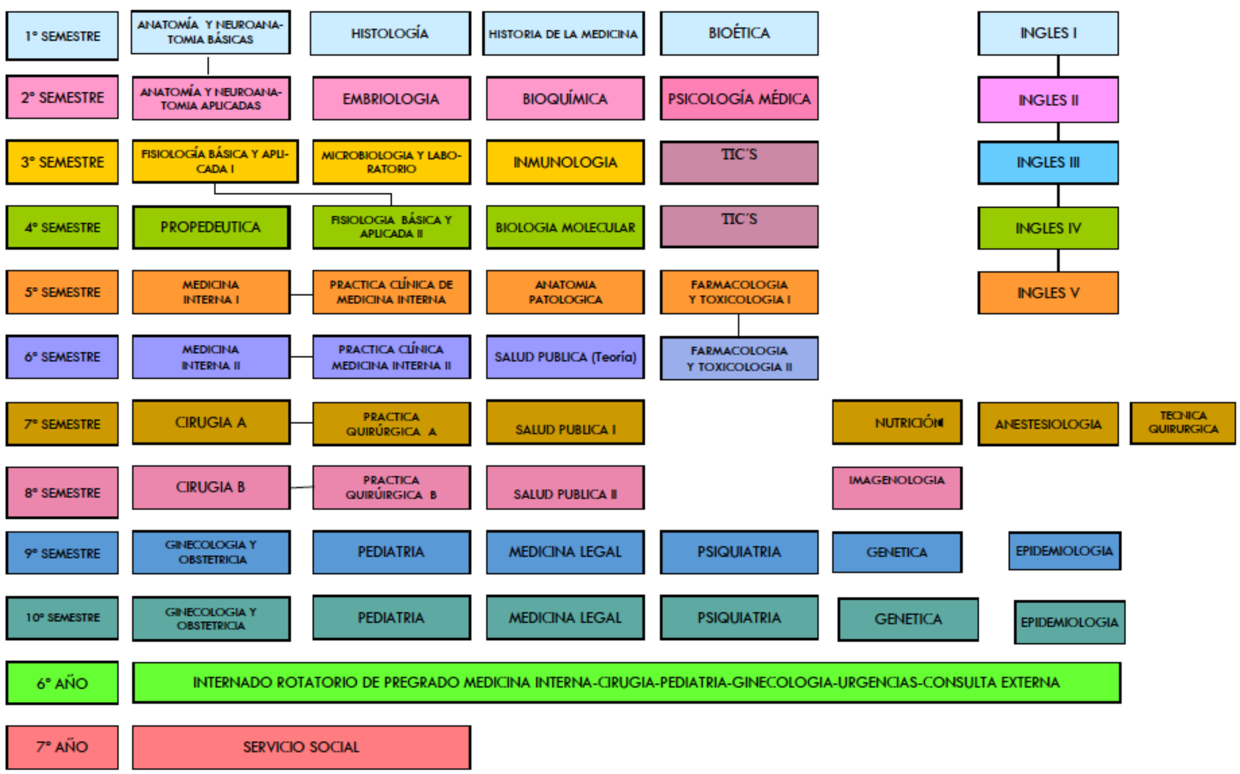


Figura 2. Mapa curricular de la licenciatura en Médico Cirujano de la Facultad de Medicina de la UASLP hasta 2018.[1]

El EGEL-MG consta de 93 reactivos de opción múltiple, y "evalúa los conocimientos y habilidades en las áreas y subáreas o de la formación del licenciado en Medicina General, acordadas por el Consejo Técnico como centrales para medir la formación profesional en este campo." (Centro

[1] Facultad de Medicina de la UASLP (2017).

Nacional de Evaluación para la Educación Superior, 2014, 2018). Por otro lado, el ECOE es una prueba estandarizada en la que los sustentantes demuestran sus habilidades prácticas (clínicas) al pasar a través de 10 estaciones con diferentes casos clínicos de diferentes especialidades, donde son valorados por profesores entrenados y quienes utilizan escalas globales para establecer las calificaciones (Cuschieri et al., 1979).

4. MARCO TEÓRICO

4.1. ANTECEDENTES DE VALIDEZ EN EVALUACIÓN

La evaluación de los exámenes se debe llevar a cabo considerando tres dimensiones por separado: validez, confiabilidad y justicia; algunos autores incluyen la justicia y la confiabilidad como componentes de la validez (Lane et al., 2016). En esta tesis se abordan de manera independiente.

La validez es el juicio acerca del grado en que la evidencia empírica y las razones teóricas que se presentan apoyan o refutan lo apropiado o adecuado de la interpretación que se da a los resultados de una evaluación. Por otro lado, la característica o concepto que se mide en una evaluación específica es un constructo, y siempre se debe especificar cuál es la interpretación que se va a dar acerca de éste con base en las puntuaciones obtenidas. Es así como las inferencias que se hacen acerca de un constructo con base en la puntuación de una evaluación son las que requieren validez, mas no la evaluación por sí misma (American Educational Research Association et al., 2018; Downing, 2003; Messick, 1989).

El marco tradicional de referencia identificaba tres tipos de validez: de contenido, de constructo y de criterio, esta última dividiéndose en validez concurrente y validez predictiva. Este esquema ya no se utiliza y actualmente existen dos marcos de referencia modernos que se toman en cuenta al evaluar la validez: el de Messick y el de Kane (Brualdi, 1999; Cronbach & Meehl, 1955).

4.2. MARCO DE REFERENCIA DE MESSICK

El marco de referencia de Messick (Messick, 1989) toma a la validez de constructo como un concepto paraguas al agrupar los tres tipos de validez tradicional debido a que considera que las evaluaciones tienen como objetivo medir constructos, es decir, las características o atributos de las personas que se miden a través del examen diseñado (American Educational Research Association et al., 2018; Cronbach & Meehl, 1955). Por ejemplo, el desempeño académico no puede ser observado directamente, por lo que se infiere a través de los resultados obtenidos en los exámenes de cada asignatura y conforma un constructo susceptible de estudio (Downing, 2003; York et al., 2015). De esta manera, cualquier estudio de validez en el marco de Messick busca aportar, de forma directa o indirecta, evidencia para el constructo que subyace la evaluación.

Por otro lado, el acercamiento a la validez siempre debe ser a través de una hipótesis o inferencia acerca de la interpretación que se pretende dar a la prueba. Después, se deben recopilar y analizar los datos, enlazarlos a un marco teórico específico, y luego determinar la validez o invalidez de la hipótesis declarada para un momento particular en el tiempo y para una población específica (Figura 3) (Downing, 2003; Messick, 1989).

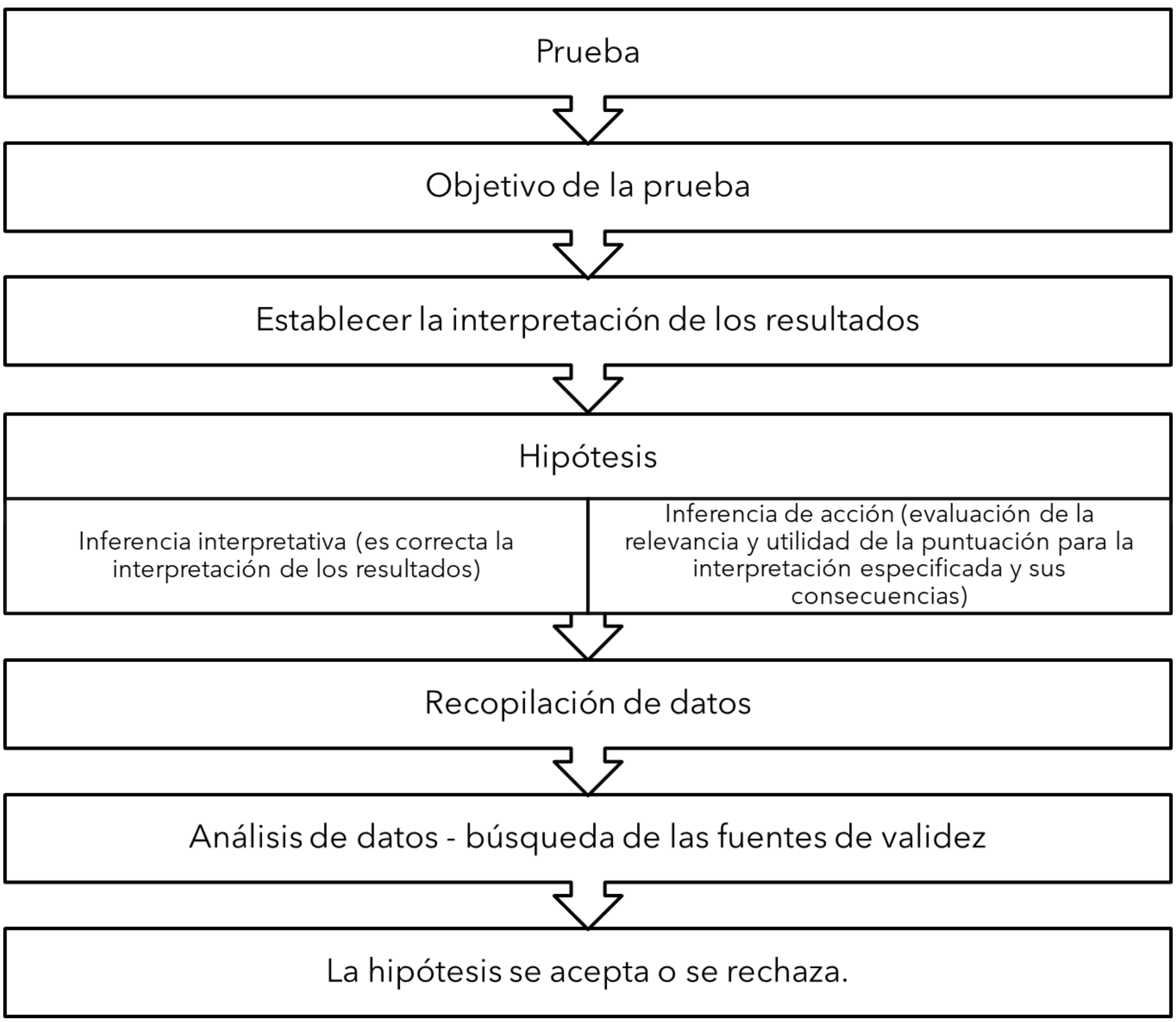


Figura 3. Resumen del marco de referencia de Messick.

La inferencia puede consistir en uno de dos tipos:

- Inferencia interpretativa.- es el grado en el que varias líneas de evidencia coinciden con la inferencia mientras que al mismo tiempo se establece que las inferencias alternas no están tan bien respaldadas por la misma evidencia.
- Inferencia de acción.- requiere de la validez de la interpretación de la puntuación y también de los resultados de acción y de las implicaciones, especialmente la evaluación acerca de la relevancia y utilidad de la puntuación para el propósito especificado y para las consecuencias que tiene esta prueba (Messick, 1989).

Con base en lo anterior, este marco de referencia se enfoca en cinco fuentes de evidencia de validez. ¿Qué tipo de evidencias y cuáles fuentes son necesarias para cada prueba? Esto depende de los objetivos de la prueba, de sus consecuencias y de que sustenten la interpretación de los resultados de la prueba, entre otros aspectos (American Educational Research Association et al., 2018; Cook & Hatala, 2016; Downing, 2003). Por ejemplo, en el caso de pruebas de altas consecuencias como un examen de admisión o de titulación, se requieren demostrar estándares de calidad mayores que los de una prueba utilizada con fines formativos.

4.2.1 Fuentes de evidencia de validez

1. Evidencia basada en el contenido de la prueba.

El contenido de la prueba se refiere a los temas que evalúa; por ejemplo, en el caso de un examen de admisión, abarcaría toda la información cuyo dominio debe demostrar un alumno antes de ingresar al nivel al que pretende. (American Educational Research Association et al., 2018). Depende de las inferencias que se hacen a partir de los resultados de la prueba sobre el constructo acerca del que el sustentante debe demostrar su aptitud. Esta fuente de evidencia posee cuatro elementos (American Educational Research Association et al., 2018; S. Sireci & Faulkner-Bond, 2014):

1. Definición del dominio. – proporciona de detalles con respecto de lo que la prueba mide, y transforma el constructo teórico en un dominio de contenido concreto. Esto se logra:

   a. Al describir detalladamente las áreas del contenido y las habilidades cognitivas para cuya medición se ha diseñado el instrumento.
   b. Por medio de la tabla de especificaciones de la prueba, en donde se deben enlistar las subáreas y los niveles cognitivos que se miden.
   c. Al mostrar los estándares específicos de contenido, objetivos curriculares, o habilidades contenidas dentro de los diferentes niveles cognitivos.

   Las fuentes de evidencia para este elemento serán las evaluaciones que realice un panel de expertos. Además, en el caso de que no estén representados aspectos importantes del constructo en la tabla de especificaciones de la prueba, debe aclararse exactamente cuáles son y por qué.

2. Representación del dominio. – es el grado en que una prueba representa y mide el dominio definido en las especificaciones. La opinión de expertos externos e independientes provee de la evidencia necesaria al calificar cada ítem para determinar si en efecto representa completa y suficientemente al dominio, si concuerda con el estándar de contenido o un elemento de la tabla de especificaciones de la prueba, con el nivel cognitivo que se pretende alcanzar durante las clases, etc.; esto es lo que se conoce como alineación (Bhola et al., 2003).
3. Relevancia del dominio. – es el grado en que cada ítem en una prueba es relevante para el dominio que se evalúa con base en la tabla de especificaciones de la prueba. Permite saber si todos los aspectos importantes del dominio son medidos por el instrumento y si se está evaluando contenido trivial o irrelevante.
4. Desarrollo apropiado del instrumento. – son los procesos utilizados al construir la prueba para asegurarse de que el contenido representa fielmente al constructo que se pretende medir y que además no mide contenido irrelevante. Se deben demostrar procedimientos de calidad fuertes y explicar las razones para utilizar formatos específicos de los ítems que conforman el instrumento. Para lo anterior existen estas fuentes:
   a. Revisión de los ítems por expertos que aseguren su exactitud técnica.
   b. Revisión de los ítems por expertos de medición para determinar qué tan bien se adhieren a los estándares de principios de escritura de ítems de calidad.
   c. Revisión de sensibilidad para evitar varianza irrelevante al constructo (VIC).

   d. Piloteo de los ítems con análisis estadístico para seleccionar los ítems más apropiados para uso operativo.
   e. Análisis de FDI.

2. Evidencia basada en los procesos de respuesta.

En los *Estándares para pruebas educativas y psicológicas* (American Educational Research Association et al., 2018) esta evidencia se refiere a que se puede comprobar la relación entre el constructo que se pretende medir y los procesos cognitivos que intervienen en la resolución de la tarea o los ítems de la prueba. En el caso de pruebas de preguntas abiertas y POM, esta evidencia puede obtenerse por medio de dos tipos de fuentes (Padilla & Benítez, 2014):

1. Fuentes directas
   a. Entrevistas cognitivas. – se usan para examinar la comprensión de los términos clave, así como entender el razonamiento utilizado para llegar a la respuesta correcta y así evitar falsos positivos (Embretson, 1998).
   b. Entrevistas. – con protocolos de pensamiento en voz alta.
   c. Grupos focales.
2. Fuentes indirectas
   a. Tiempo de respuesta.
   b. Seguimiento de movimientos oculares.

Por otro lado, Downing (2003) las define como las fuentes de error que pueden derivar de la administración del examen han sido consideradas y se han tomado medidas al respecto. Para sustentarlas se debe verificar que los estudiantes estén familiarizados con el formato del examen, por ejemplo, que sepan llenar adecuadamente las hojas de respuesta. Otro caso interesante es cuando se obtiene una puntuación final a partir de puntuaciones parciales de dos o más exámenes independientes: hay que asegurarse de que los métodos utilizados para combinar estas puntuaciones son útiles y apropiados, además de que los sustentantes reciban un reporte comprensible y adecuado de estos resultados. Además, se deben explicar las razones de la combinación de puntuaciones, por qué se eligió la escala de calificación utilizada y el significado de cada examen que aporta a la calificación final. El material de práctica y los instructivos entregados previamente al examen también aportan evidencia de validez en este rubro, así como el documentar todos los procedimientos de control de calidad, por ejemplo, el de las máquinas lectoras de calificaciones, y el de las pruebas piloto para comprobar la exactitud de la clave del examen.

3. Evidencia basada en la estructura interna.

La estructura interna es "el grado en que las relaciones entre los ítems de la prueba y los componentes de la prueba se ajustan al constructo" que se mide (American Educational Research Association et al., 2018). Se relaciona con las características psicométricas de las preguntas del examen o de los indicadores de desempeño, las características de la escala, y el modelo psicométrico que se utilizó para establecer la escala y calificar el examen (Downing, 2003).

Esta fuente presenta tres aspectos (Rios & Wells, 2014):

- Dimensionalidad. – informa acerca de las relaciones entre los ítems y así da soporte a las puntuaciones para formular la inferencia de puntuaciones. La forma de demostrarla es a través de un análisis factorial confirmatorio (AFC) y un análisis basado en la teoría de respuesta al ítem (TRI), las que permiten estudiar las relaciones entre las respuestas a los ítems y el constructo medido (Leenen, 2014).
- Invarianza de la medida. – demuestra que las puntuaciones serán las mismas cuando se comparen entre grupos con diferentes características como raza, edad o sexo, de acuerdo con la taxonomía de equivalencia. En este caso, el nivel que debe estudiarse es acerca de invarianza débil y fuerte a través de un análisis factorial exploratorio (AFE) y pruebas de funcionamiento diferencial del ítem (FDI) (Boer et al., 2018).
- Confiabilidad. – permite saber si en cada ocasión que se aplica el instrumento se obtendrán resultados semejantes y también apoya a la inferencia de puntuaciones. En el caso de exámenes de altas consecuencias es importante poder reproducir las puntuaciones al aplicarlos de manera repetida; de lo contrario, la interpretación de los resultados de este examen se ve comprometida debido a que esta fuente de evidencia de validez es pobre o nula. Se prueba por medio del alfa de Cronbach o Kuder-Richardson 20 o 21, según la prueba (Downing, 2004).

<u>4. Evidencia basada en las relaciones con otras variables.</u>

Es el análisis de la relación de los resultados de la prueba con los resultados de otras pruebas que midan el mismo constructo u otras variables externas a la prueba. Proporciona información acerca del grado en que estas relaciones son coherentes con el constructo en el que se basan las interpretaciones de los resultados de la prueba (American Educational Research Association et al., 2018). Se puede buscar evidencia por esta fuente con base en los tipos de relación: convergente, cuando se evalúan las relaciones entre las puntuaciones y medidas del mismo constructo; y discriminante, cuando se evalúan las relaciones entre las puntuaciones y medidas de constructos diferentes (Downing, 2003). Una manera de establecer ambas correlaciones es a través del multirasgo multimétodo de Campbello, en el que se establece una matriz de correlaciones entre las pruebas representando al menos dos características, cada una de las que es medida por lo menos por dos métodos (Campbell & Fiske, 1959). Se conocen dos tipos de diseño para conocer las relaciones entre la prueba y el criterio (American Educational Research Association et al., 2018):

a. Estudio predictivo.- evalúa el grado de la relación entre las puntuaciones de la prueba y las puntuaciones del criterio que se obtiene en un tiempo posterior. Por ejemplo, este tipo de estudios son útiles en las evaluaciones de admisión académica, para predecir el desempeño académico posterior (el criterio). También permiten analizar predicciones diferenciales por subgrupos, por ejemplo, grupos de edad, sexo, antecedentes académicos, etc.

b. Estudio concurrente.- evalúa el grado de la relación entre las puntuaciones de la prueba y las puntuaciones del criterio que se obtiene al mismo tiempo. En este tipo de estudios se evitan los cambios temporales y pueden ser útiles para buscar formas alternas de medición del constructo en cuestión.

La generalización de los resultados que aporta el estudio de esta fuente de validez depende de que las condiciones en la nueva situación sean iguales a las presentes en el análisis original. Los resúmenes estadísticos de los estudios de validación anteriores en condiciones semejantes, como en un metanálisis, pueden ser útiles para estimar las nuevas relaciones, pero dependen del tamaño de la muestra y de la cantidad de estudios realizados a lo largo del tiempo (American Educational Research Association et al., 2018; Coates, 2008).

Existen tres tipos (Campbell & Fiske, 1959):

1. Convergente-divergente. – convergente es en el que se hace un análisis de correlación entre los resultados de pruebas que miden lo mismo, mientras que el divergente es con resultados de pruebas que miden un constructo diferente.
2. Instrumento-criterio. – es la exactitud de las puntuaciones para predecir el desempeño en un criterio relevante, y puede ser predictivo o concurrente.
3. Generalización de validez. – es el grado en que la relación con otras variables puede generalizarse hacia una nueva situación o un dominio más amplio del conocimiento sin necesidad de hacer más pruebas.

Los dos primeros tipos se pueden demostrar por medio de un análisis de matriz multirasgo - multimétodo, mientras que el último requiere de un metaanálisis.

5. Evidencia para la validez y las consecuencias de la prueba.

La interpretación de los resultados de la prueba tiene diferentes grados de consecuencia sobre sustentantes, instituciones, sociedad y otros usuarios. Sobre todo, en las pruebas de alto impacto, es importante evaluar esta fuente de evidencia; por ejemplo, en el caso de las evaluaciones de admisión para una licenciatura, uno de sus objetivos es evitar la entrada de sujetos incapaces para la profesión, mientras que procura que los candidatos más idóneos logren iniciar sus estudios en la medida de lo posible. Por otra parte, esta fuente de evidencia da información para reflexionar sobre las equivocaciones de la interpretación de los resultados de la prueba con respecto a falsos positivos y falsos negativos. La evidencia de validez indicará si es factible o no alcanzar este objetivo.

Esta fuente se puede comprobar al analizar el impacto de los resultados de la prueba en los estudiantes y/o la sociedad, las consecuencias en los estudiantes o el futuro de su aprendizaje, el balance entre las consecuencias positivas y las negativas involuntarias, lo razonable acerca del punto de corte de aprobado/reprobado o admitido/no admitido, las consecuencias de aprobar o reprobar, de los falsos positivos y falsos negativos, y las consecuencias institucionales y del estudiante (American Educational Research Association et al., 2018; Downing, 2003). Este análisis puede realizarse por medio de entrevistas, encuestas y grupos focales, así como la teoría de acción para identificar los componentes críticos de los programas académicos y sus puntos de impacto (Lane, 2014).

Kane consideró que, aunque la visión de Messick acerca de la validez de constructo es importante, no es fácil de evaluar, ya que no provee de guía para iniciar el procedimiento ni es práctico (Kane, 2011). Por esto desarrolló su propio marco de referencia, que se enfoca en el proceso de recolección de evidencia de validez mediante cuatro inferencias para desarrollar un argumento de validez (Cook et al., 2016), pues el planear un examen considerando estas inferencias permite partir de la evaluación de una sola observación (inferencia de puntuación) hacia la puntuación general del examen (generalización) y de ahí a establecer las implicaciones de la puntuación en el desempeño en la vida real (extrapolación), llegando finalmente a la interpretación de esta información y a la toma de decisiones (implicaciones) (Cook et al., 2015). Una ventaja de este acercamiento a la validez es que es más fácil de llevar a cabo para quienes no poseen experiencia amplia en psicometría, además de que propone pasos muy claros (Brennan, 2013).

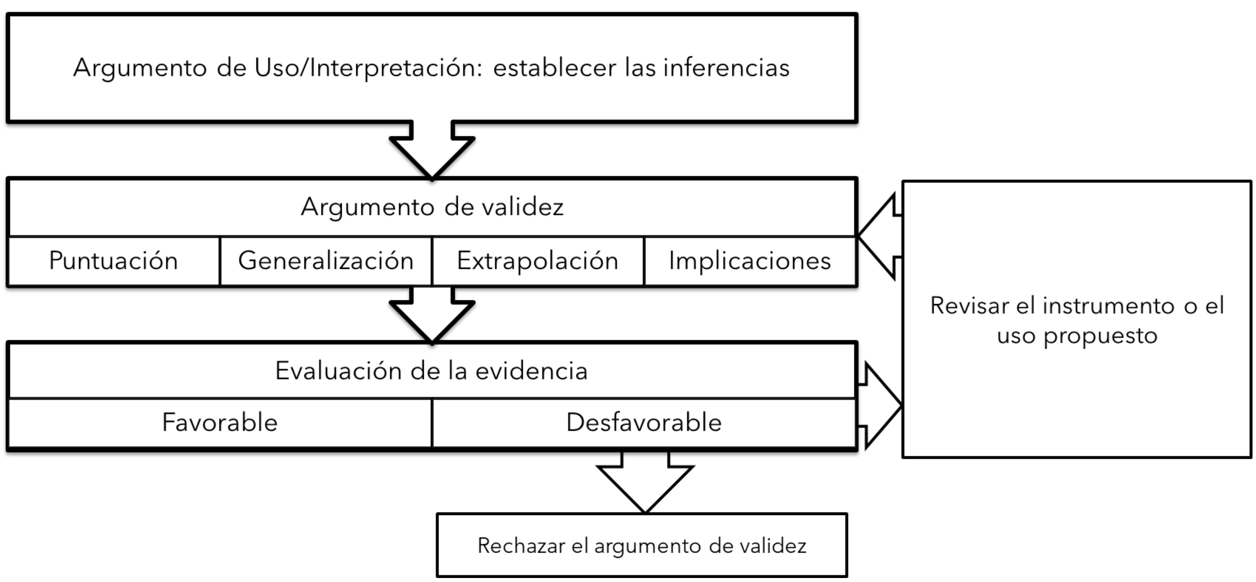


Figura 4. Marco de referencia de Kane. [2]

[2] Basado en Cook et al., 2015

En general, los pasos que propone son dos: el primero es establecer el argumento de uso o interpretación (AUI) y el segundo es desarrollar el argumento de validez; este último es facilitado al considerar los cuatro tipos de inferencias (Figura 4).

1. Establecer el argumento de uso o interpretación (AUI).

La interpretación de los resultados de la prueba consiste en las suposiciones que se hacen acerca de los sustentantes; por otro lado, el uso de las puntuaciones se refiere a las decisiones que se toman acerca de los sustentantes. Kane considera que ambos términos (interpretación y usos) incluyen todas las suposiciones que se pueden hacer al respecto de las puntuaciones de una prueba, por lo que se debe establecer la validez de la interpretación o el uso de las puntuaciones en términos de lo creíble y apropiado que tengan en un punto del tiempo. Tener claro lo que se quiere evaluar permite elaborar un plan de evaluación preciso, por lo que el AUI puede conformar una red de supuestos que vayan desde el desempeño en las pruebas hasta las conclusiones que se sacan y las decisiones que se toman con base en estas conclusiones (Kane, 1992, 2013a). Kane sugiere las siguientes inferencias que se encuentran presentes en la mayoría de los AUI, aunque también menciona que no es indispensable evaluarlas todas (Chalhoub-Deville, 2016; Kane, 2013a):

- **Inferencia de puntuación.-** Es la suposición acerca de lo apropiado de los criterios de la puntuación y las reglas para combinar las puntuaciones, mismas que conforman un estimado acerca de un atributo y son la base para la toma de decisiones.
- **Inferencia de generalización.-** Si la prueba contiene una muestra de posibles escenarios o ítems, esta inferencia supone que al sustentante va a obtener puntuaciones semejantes al presentar otra prueba con ítems diferentes extraídos del mismo universo de ítems, de manera que las puntuaciones observadas son representativas de todo el universo de puntuaciones posibles. También se conoce a esta inferencia como confiabilidad, debido a la necesidad de reproducibilidad de las puntuaciones.
- **Inferencia de extrapolación.-** Por medio de este tipo de suposiciones se podría extender la interpretación a otros dominios de desempeño y predecir cuál será el resultado del sustentante en contextos diferentes al del examen o tareas diferentes en contextos diferentes. Un ejemplo de este tipo de inferencia sería que si la puntuación observada tiene un valor particular (examen de

admisión), entonces se espera un valor específico del criterio (desempeño académico durante la carrera); esto se podría evaluar por medio de una ecuación de regresión.

- **Inferencia de implicaciones.-** Se refiere al impacto que tiene la interpretación de los resultados de la prueba en el sustentante, en su familia y en la sociedad. Kane considera que, si las consecuencias de la interpretación de los resultados de una prueba son negativas, entonces la prueba no debería utilizarse.

2. Establecer el argumento de validez.

Ya que se han establecido las inferencias concernientes a las puntuaciones de la prueba en cuestión, se deben evaluar las garantías o métodos de comprobación de estas inferencias. Por ejemplo, la garantía de una inferencia de extrapolación con interés predictivo sería una ecuación de regresión, cuyo soporte estaría conformado por un análisis empírico acerca de la relación entre la puntuación de la prueba y los resultados del criterio seleccionado. El calificador de la garantía es el término que expresa la fuerza de la relación que se está analizando, y puede expresarse de manera numérica y/o con palabras: coeficientes de correlación, por ejemplo (Kane, 2013a).

Además, es importante mencionar que los estimados de la puntuación futura de los criterios basados en regresión deben tomar en cuenta el error estándar del parámetro estimado, mientras que las inferencias de generalización podrían presentar el error estándar de la medida. El error estándar del parámetro estimado es el estimado del grado en que la puntuación predicha del criterio sea incorrecta; mientras más pequeño sea este error, es más exacta la predicción. El error estándar de la medida es qué tanto cambia la medida observada bajo diferentes circunstancias, conformándose en la incertidumbre de la generalización. Estos dos tipos de errores también fungen como calificadores cuantitativos (Fraenkel et al., 2019; Kane, 2013a).

Con estas consideraciones, el primer paso para desarrollar este argumento será realizar un análisis conceptual del AUI y verificar que sea coherente y que todas las inferencias importantes se encuentren presentes. Posteriormente se deberán evaluar las inferencias presentadas. En la Tabla 2 se resumen las inferencias que propone Kane, así como los procedimientos que se deben definir y la manera de evaluarlos.

El argumento de validez debe ser claro para poder ser reproducible por cualquier investigador, conteniendo detalles específicos y presentando información coherente, de manera que las conclusiones sean lógicas. Por lo anterior, el argumento también debe estar completo y ser verificable (Schuwirth & van der Vleuten, 2012).

Inferencia	Consiste en	Procedimientos para definir, establecer o seleccionar	Evaluación empírica de:
Puntuación	Suposición acerca de lo apropiado de los criterios de la puntuación y las reglas para combinar las puntuaciones.	● Ítems y opciones de respuesta ● Formato de la observación ● Estandarización entre formatos y ocasiones ● Rúbrica o criterio de puntuación, procedimientos de implementación, estándar de aprobado/no aprobado ● Selección y entrenamiento de los evaluadores ● Reglas para combinar los elementos relacionados con la prueba a partir de fuentes diferentes o para separar elementos no relacionados de la misma fuente ● Seguridad de los datos y control de calidad	● Desempeño de ítems y de opciones de respuesta ● Formato de observación ● Estandarización ● Rúbrica o criterio de puntuación ● Selección y entrenamiento de los evaluadores, confiabilidad y precisión de los evaluadores ● Seguridad de los datos y control de calidad
Generalización	Los ítems de la prueba conforman una muestra del universo de ítems posibles. Esta inferencia supone que se puede generalizar hacia todo el universo de ítems posibles. Se relaciona con la confiabilidad.	● Estrategia de muestreo ● Tamaño de la muestra	● Confiabilidad o generalizabilidad ● Teoría de respuesta al ítem
Extrapolación	Se podría extender la interpretación a otros dominios de desempeño y predecir cuál será el resultado del sustentante en contextos diferentes al del examen o tareas diferentes en contextos diferentes.	● Alcance de la prueba ● Autenticidad del contexto de la prueba ● Autenticidad del ítem/escenario ● Análisis que demuestren la relación entre el desempeño en la prueba y los dominios o contextos diferentes a los que se desea extrapolar.	● Análisis para definir el alcance/objetivos ● Acuerdo entre el proceso y el constructo ● Relevancia y autenticidad ● Correlación con otra medida que presente la misma relación esperada (con referencia al criterio o convergente; concurrente o predictiva) ● Discriminación ● Sensibilidad a cambiar después de la intervención ● Perfil del constructo ● Funcionamiento diferencial del ítem

Implicación	Acerca del impacto de la interpretación de los resultados de la prueba sobre el sustentante, otros interesados y la sociedad entera.	• Estándar de aprobado/no aprobado • Acciones planeadas con base en los resultados de la prueba • Consecuencias voluntarias o involuntarias de las decisiones que se toman a partir de los resultados de la prueba.	• Estándar de aprobado/no aprobado (p ej. curva ROC) • Efectividad de las acciones basadas en los resultados de la prueba • Consecuencias voluntarias o involuntarias de la prueba • Funcionamiento diferencial del ítem

Tabla 2. Las inferencias y sus fuentes de evidencia correspondientes para establecer el argumento de validez.[3]

[3] Cook et al., 2015; M. T. Kane, 2013; Schuwirth & van der Vleuten, 2012.

Las hipótesis que pueden funcionar como guía para elaborar las hipótesis propias, según los supuestos, han sido publicadas por varios autores (Clauser et al., 2008; Hatala et al., 2015), y se anotan en la Tabla 3:

Supuestos	Hipótesis asociadas a los componentes de Kane
Puntuación	Las preguntas fueron administradas bajo condiciones estandarizadas – la regla es apropiada Las puntuaciones fueron registradas de manera rigurosa – la regla se aplica como se especificó Los algoritmos de puntuación fueron aplicados correctamente – la puntuación está libre de sesgo Se implementaron los procedimientos de seguridad apropiados
Generalización	¿Cuáles son las fuentes de medición del error que contribuyen a las puntuaciones observadas en la evaluación? ¿Qué tan semejantes serían las puntuaciones entre las réplicas del procedimiento de medición? ¿En qué medida se utilizó un proceso sistemático para construir las formas de la prueba? La muestra es representativa del universo de observaciones posibles La muestra es lo suficientemente grande como para controlar para el error aleatorio La puntuación puede generalizarse de la muestra a la población específica: ítems, jueces, etc.
Extrapolación	La puntuación observada se relaciona con el constructo de la vida real de interés No hay probables errores sistemáticos que socaven la extrapolación Las puntuaciones predicen los resultados de la vida real de interés Existen aspectos artificiales de las condiciones de la prueba que impacten en las puntuaciones (VIC)
Implicaciones	La puntuación de corte fue establecida de manera razonable Las implicaciones (interpretaciones) son apropiadas Las propiedades de las puntuaciones apoyan las implicaciones (interpretaciones) asociadas

Tabla 3. Hipótesis generales que pueden desarrollarse en el AUI.

4.4 Validez de apariencia

Un factor interesante a considerar la revisión y lectura crítica de bibliografía sobre validez es la validez de apariencia (*"face validity"*). Esta es la apreciación subjetiva de validez de la interpretación de los resultados de la prueba, generalmente por medio de argumentos que se basan en el sentido común o lógica, y se considera una medición de la credibilidad. Quienes llevan a cabo estos análisis han declarado que "los juicios de validez de apariencia son percepciones y no son necesariamente correctos" (Tweed & Cookson, 2001). Los profesionales de la medición educativa no la consideran una forma académica o científica de demostrar validez, por lo que no se recomienda su uso (Downing, 2006; K. Royal, 2016).

4.5 Amenazas a la validez de constructo

El resultado del análisis de validez de los usos e interpretaciones de las puntuaciones de una prueba nos dirá el grado en que son apropiados estos usos e interpretaciones. Sin embargo, la tarea de validación no termina aquí, ya que, como Cronbach (1988) sugiere, es necesario buscar otras hipótesis que expliquen la existencia de resultados que no concuerden con la hipótesis original o que causen una disminución en el grado de validez alcanzado. Esto es importante para dar mayor fortaleza a las decisiones que se tomen con base en las puntuaciones del examen que estamos valorando y cobra mayor relevancia mientras mayor sea el escrutinio al que esté sometido dicho examen, así como el impacto de los falsos positivos y falsos negativos sobre los sustentantes, sus familias, la institución educativa y la sociedad.

Las amenazas a la validez son los factores que interfieren con la interpretación significativa de la puntuación de la evaluación, disminuyendo la evidencia de validez (Downing & Haladyna, 2004; Downing & Yudkowski, 2009). Estas amenazas pueden encontrarse en cualquier tipo de evaluación, ya sea de conocimientos teóricos o prácticos, formativa o sumativa, etcétera (Downing & Yudkowski, 2009). La mayoría de las pruebas que se aplican durante la carrera de medicina, enfermería y otras ciencias de la salud se hacen por medio de POM. Se han revisado varias de estas evaluaciones para comprobar que la calidad de tales reactivos es en general cuestionable (Downing, 2002b; Tarrant et al., 2006; Ware & Vik, 2009), ya que muchas veces se elaboran poco tiempo antes de aplicarlas y sin una opinión colegiada (Jozefowicz et al., 2002), constituyendo una amenaza a la validez de sus resultados.

Aunque se mencionan varios tipos de amenazas (por ejemplo, Crooks et al. (1996) consideran al menos 23, relacionadas con ocho inferencias), en general caen en dos clases principales (Messick, 1989): la subrepresentación del constructo (SC) y VIC), mismas que serán descritas a continuación.

4.5.1 Subrepresentación del constructo

En el caso de una prueba escrita, la SC se refiere a que, considerando un universo de ítems posibles, esta prueba contiene una muestra de ítems que puede:

- Ser insuficiente para evaluar el dominio del conocimiento correspondiente,
- Estar sesgada hacia una parte del tema a evaluar, convirtiéndose en una muestra no representativa,
- Evaluar contenido trivial o factual al nivel más bajo de la pirámide de Miller (figura 5),
- No corresponder con el dominio que se pretende evaluar, o
- Poseer baja confiabilidad (Downing, 2002b; Downing & Haladyna, 2004).

Esta es una amenaza particularmente importante para la inferencia de extrapolación, ya que la interpretación de las puntuaciones carece de significado si los resultados no son representativos de los constructos que se supone que la prueba evalúa(Hawkins et al., 2010).

Figura 5. Pirámide de Miller.[4]

A lo largo de la explicación de estas amenazas se utilizará un ejemplo de ciencias básicas: el tema de anatomía de cabeza (sin neuroanatomía) abarca 160 páginas del libro de “Anatomía con orientación clínica de Moore, 7ª edición” (Moore et al., 2013), uno de los libros más utilizados para la enseñanza de anatomía humana en nuestro país. Si fuera a aplicarse un examen del tema con POM basándose únicamente en el contenido de tal libro (Tabla 4), estas serían las amenazas a la validez con respecto a la SC:

[4] Miller (1990).

Pregunta	Opciones de respuesta
• ¿Cuántos huesos conforman el viscerocráneo?	a. 11 b. 12 c. 13 d. 14 e. 15*
• La siguiente estructura generalmente está inervada por el nervio laríngeo interno:	a. Aritenoides oblicuo b. Cricoaritenoideo posterior c. Cricotiroideo d. Mucosa infralaríngea e. Mucosa supralaríngea*
• En la coroides no ocurre lo siguiente:	a. Contiene ramas de la arteria central de la retina* b. La lámina coroidocapilar es la más interna c. Produce el reflejo rojo del fondo de ojo d. Se encuentra entre la esclera y la retina e. Sus venas drenan en una vena vorticosa
• Una mujer joven se golpea la cabeza con el cuadro de mandos del automóvil durante una colisión frontal. A continuación, sufre un desgarro de la parte frontal del cuero cabelludo con sangrado abundante. La herida se lava con suero fisiológico y se cubre con una venda estéril. Cuando la mujer llega al hospital tiene los dos ojos morados. En la exploración posterior no se aprecia ninguna lesión ocular (Moore et al., 2002) ¿Cuál es el vaso sanguíneo que más probablemente se lesionó en este caso?	a. A auricular posterior b. A facial, porción cervical c. A mentoniana d. A supraorbitaria* e. A temporal superficial
• ¿Cuál es la acción principal del músculo recto inferior? I. Abducir el globo ocular II. Aducir el globo ocular III. Descender el globo ocular IV. Rotar lateralmente el globo ocular V. Rotar medialmente el globo ocular	a. I, II y III b. II, III y IV* c. III, IV y V d. I, III y V e. I y IV
• Which bone does NOT contribute to the orbit?	a. Frontal bone b. Maxilla c. Palate bone d. Sphenoid bone e. Temporal bone*
• ¿Cuáles son los huesos que rodean la mollera?	a. Frontales y parietales* b. Frontales y temporales c. Occipital y temporales d. Parietales y occipital e. Temporales y parietales
• ¿Cuál de los siguientes es un músculo de la cara?	a. Bíceps braquial b. Dorsal ancho c. Esternocleidomastoideo d. Frontal* e. Psoas mayor
* Respuesta correcta	

Tabla 4. Ejemplos de preguntas de un examen de anatomía de cabeza.

- Número de preguntas insuficiente. Considerando un amplio universo de ítems de anatomía de cabeza, un examen que consta de 10 ítems quizá no sea adecuado, con base en la extensión de los temas que comprende esta unidad y los objetivos de aprendizaje que se hayan establecido previamente. Downing y Haladyna (Downing & Haladyna, 2004) sugieren un mínimo de 30 preguntas en general, mientras que en el manual del NBME sugieren 100 preguntas para obtener resultados reproducibles (National Board of Medical Examiners, 2016), aunque no especifican el tipo de prueba al que van dirigidas estas recomendaciones. Para determinar la cantidad adecuada de ítems se sugiere determinar los objetivos de aprendizaje y considerar factores como el tiempo real que tienen los alumnos para contestar el examen, así como ponderar la importancia de cada uno de los temas a examinar al igual que la relevancia de la prueba en cuanto a si es una evaluación sumativa o formativa y la exactitud necesaria de las puntuaciones (American Educational Research Association et al., 2018; Moreno et al., 2004).
- Sesgo. La pequeña cantidad de ítems a su vez causaría un sesgo hacia algún área de contenido, ya que probablemente faltarían preguntas acerca de algunos temas incluidos en los objetivos de aprendizaje. Por ejemplo, que de las 10 preguntas dos fueran de cráneo, dos de ojo, una de nariz, una de oído, una de región parotídea, una de vascularización de la cabeza y dos de meninges. Una sola pregunta de oído no sería suficiente para evaluar el conocimiento que requiere un alumno durante su paso por esta asignatura, pues ¿cuál sería el tema de la pregunta? ¿Oído externo, medio o interno, exploración física, aplicación clínica, relación con el nervio facial, innervación, vascularización, drenaje venoso, drenaje linfático, etc.? Cualquiera que sea, sin duda causaría sesgo (Downing, 2002b; Downing & Haladyna, 2004).
- Concordancia del dominio. Ya que el cerebro está en la cabeza, ¿podrían hacerse preguntas sobre sus funciones y relaciones con otras estructuras neurológicas como los nervios craneales en este examen? Si entre los objetivos de aprendizaje no se establece el estudio de este tema, sería equivocado evaluarlo, ya que esto causaría que los ítems no correspondieran al dominio de contenido que se pretende evaluar (Downing & Haladyna, 2004).
- Nivel de evaluación con base en la pirámide de Miller. Se refiere a que las preguntas acerca de cabeza solo fueran acerca de hechos memorizables de dudosa utilidad (Tabla 1, pregunta

1), sin evaluar integración entre estos conocimientos y otros previamente adquiridos, o relacionándolos con su aplicación clínica, o con los contenidos de otras asignaturas para llegar a una conclusión adecuada. En una ciencia básica como anatomía, suele ser difícil elaborar ítems que vayan más allá de conocimientos factuales; sin embargo, es posible ir más allá mientras se tengan claros los objetivos de aprendizaje, así como los usos e interpretaciones de la prueba (Hadie, 2018).

- Baja confiabilidad. El haber elaborado un examen de 10 preguntas acerca de hechos memorizables de dudosa utilidad para el futuro puede producir puntuaciones con baja reproducibilidad y confiabilidad, y con un error estándar de la medida grande. Además, los intervalos de confianza de estas puntuaciones se vuelven muy amplios, causando incertidumbre acerca de cuál es la calificación real de aprobación, dependiendo de la distribución de estos datos (Downing, 2002b; Downing & Haladyna, 2004).

4.5.2 Varianza irrelevante al constructo

Este es el error sistemático causado por la medición involuntaria de constructos irrelevantes cuya medición no es el objetivo del examen, por lo que interfieren con la medición del constructo original y por lo tanto con la validez de la interpretación de la puntuación (Downing & Haladyna, 2004; Messick, 1989). Este concepto es diferente al error aleatorio, que es la diferencia entre la puntuación observada y la puntuación verdadera de cada estudiante. Es así que la puntuación observada (o puntuación total - y) se obtiene de la suma de la puntuación real (t), más el error aleatorio (er) y el error sistemático debido a VIC (es) (Haladyna & Downing, 2004):

$$y = t + er + es$$

Haladyna & Downing (2004) también mencionan que en realidad no conocemos el valor del error aleatorio, pero se espera que sea de cero. Lo que sí se puede conocer y controlar es el error sistemático o VIC, de ahí la relevancia de conocer sus posibles orígenes y cómo evitarlos.

Algunas características del examen que pueden ocasionar VIC son:

- Ítems defectuosos,
- Ítems sesgados estadísticamente,
- Nivel de lectura inapropiada para los ítems,
- Ítems muy fáciles o muy difíciles,
- Hacer trampa,
- Decisión de puntuación aprobatoria indefendible, o
- Recibir preparación para realizar el examen (Downing & Haladyna, 2004; Haladyna & Downing, 2004).

Siguiendo el mismo ejemplo de un examen de 10 preguntas acerca de anatomía de cabeza basado en el contenido del libro citado, las amenazas con respecto de la VIC pueden presentarse de la siguiente manera si algunas de las preguntas incluyeran las presentadas en la Tabla 4:

- Ítems defectuosos. Los ítems de una prueba se consideran defectuosos cuando no cumplen las normas de escritura de ítems de opción múltiple (Downing, 2005), mismas que deben conocerse para evitar que la pobre calidad de las preguntas cause mayor dificultad para contestarlas (Downing, 2002a). Por ejemplo, en la pregunta 2 de la Tabla 4 no sabemos qué significa "generalmente" ni a qué tipo de estructuras se refiere (¿músculos?). Además, la respuesta correcta es la única estructura que parece no ser un músculo (ver abajo "Recibir preparación para realizar el examen"). Otros defectos consisten en elaborar las preguntas con opciones que incluyan "todas las anteriores" o "ninguna de las anteriores" (Downing, 2005). Otra característica importante en la elaboración de ítems de opción múltiple es la cantidad de opciones, tema que ha dado para múltiples publicaciones.
- Formato no estandarizado o en negativo. Aunque algunos autores no se oponen de manera absoluta a la elaboración de este tipo de preguntas (Collins, 2006), Chiavaroli (2017) explica que deben evitarse tanto las escritas como la pregunta 3 de la Tabla 4, como aquellas que piden indicar la opción "más correcta". Esto es debido a que existe el riesgo de escribir un doble negativo (la pregunta y alguna(s) de las opciones contiene un término negativo), que el alumno no identifique la parte negativa de la pregunta, aunque la palabra "excepto" o "no" se encuentre en negritas, y que la forma de contestar no se lleve a cabo mediante el proceso de respuesta deseado, observándose identificación de la opción

correcta, afectando así al resultado de esta evidencia de validez. En el caso de POM en asignaturas de ciencias clínicas, se puede evitar la negación utilizando términos como “cuál es la contraindicación o el riesgo”.

- Formato muy complicado o extenso. La pregunta 4 de la Tabla 4 tiene una viñeta demasiado extensa y que no requería de una explicación tan amplia para la pregunta que se realiza. Un ítem como este hace que el sustentante pase más tiempo leyendo y, en lugar de evaluar el conocimiento, pone a prueba la velocidad de lectura (National Board of Medical Examiners, 2016).
- Preguntas que pueden confundir al alumno que, aunque sí conoce la respuesta, podría contestar mal. Al contestar la pregunta 5 de la Tabla 4 el alumno que quizá sí sepa las funciones del músculo referido tendrá que pasar tiempo relacionando los números romanos con la opción correcta. Además, primero debe saber los números romanos (National Board of Medical Examiners, 2016).
- Ítems sesgados estadísticamente. En el examen pueden presentarse ítems que muestren diferencias en sus características psicométricas según el grupo de alumnos que contesten con base en sus antecedentes religiosos, culturales o lingüísticos (Gómez-Benito et al., 2018). Con base en lo anterior, la pregunta 6 de la Tabla 4 resulta sumamente complicada, pues además de estar en otro idioma, su formato es negativo, de manera que los alumnos que no sepan inglés, aunque tengan el conocimiento podrían responder mal (J. W. Young, 2008), alterando también el resultado del análisis de la validez de constructo basada en la estructura interna y causando un sesgo estadístico.
- Nivel de lectura inapropiada para los ítems. La estructura de las oraciones debe ser lo suficientemente clara y evitar el uso de jerga para que la redacción de la pregunta no cause respuestas equivocadas. La pregunta 7 de la Tabla 4 se refiere a la fontanela anterior utilizando el término “mollera”, mismo que podrían desconocer algunos alumnos si no han estado en contacto con bebés o niños pequeños (Hicks, 2011).
- Ítems muy fáciles o muy difíciles. La pregunta 8 de la Tabla 4 es tan fácil, que se espera que el 100% de los alumnos la conteste correctamente, elevando artificialmente el promedio grupal e individual y, por lo tanto, causando interferencia con la validez. También se considera que no permite discriminar entre los alumnos de alto y bajo desempeño (Downing & Haladyna, 2004).

- Hacer trampa. Actualmente hay muchas formas de hacer trampa en los exámenes: desde copiar al compañero de junto hasta el uso de los smart watch. De estos hay muchos modelos y no siempre es posible distinguirlos de los no inteligentes, por lo que se aconseja prohibir el uso de cualquier tipo de reloj durante cualquier examen (Wong et al., 2017).
- Decisión de puntuación aprobatoria indefendible. En general, en los exámenes parciales de la facultad de medicina tienen una calificación mínima aprobatoria que es de 6.0. Si estos exámenes no son reproducibles o la distribución de las calificaciones es demasiado amplia (calificaciones demasiado altas y demasiado bajas), entonces no es posible justificar por qué el 6.0 es la calificación de aprobación (Norcini, 2003).
- Recibir preparación para realizar el examen. Se refiere a que los alumnos o un grupo de ellos conozcan las preguntas que se van a utilizar en el examen, lo que les daría ventaja y una puntuación artificialmente elevada. Esto puede ocurrir porque el profesor que elabora el examen presenta las mismas preguntas como parte de las clases o asesorías. También se consideran en este grupo a los estudiantes que han sido aleccionados para contestar adecuadamente los exámenes por medio de cursos de preparación, por ejemplo, y no solo conocen la estructura del examen, cómo se deben preparar, cómo son las hojas de respuesta si las hay, sino que también establecen una estrategia de uso de tiempo para contestar todos los ítems, evitan errores al leer bien las preguntas y siguen las instrucciones de manera cuidadosa (Lane et al., 2016).
- En ocasiones, los cursos de preparación consiguen que los sustentantes desarrollen *test wiseness* (TW), de manera que saben leer bien las preguntas para deducir cuál es la respuesta correcta con base en la estructura gramatical y de redacción: opciones más largas, opciones con más detalles, etcétera (Lane et al., 2016). Con base en la gran cantidad de exámenes de opción múltiple que contestan durante su vida académica, se considera que los estudiantes de medicina son *test wise* (Downing, 2005). Por otro lado, Jurado-Núñez & Leenen (2016) distinguen el TW de otros conceptos como el *educated guessing*, de manera que con el primero se pueden conseguir respuestas correctas aun sin tener conocimiento, mientras que con el segundo se los alumnos logran eliminar opciones con base en sus conocimientos (o el rasgo latente que se desea medir por medio de la evaluación), pero no consiguen identificar por completo la respuesta correcta, por lo que de todas formas

terminan adivinando. Es interesante que estos autores también consideran el TW como un factor no sistemático que contribuye al error mencionado previamente.

4.6. Confiabilidad

La importancia de la confiabilidad reside en el impacto que las evaluaciones de altas consecuencias tienen sobre la vida de los estudiantes y sus familias. Esta dimensión consiste en que los puntajes obtenidos en una evaluación sean reproducibles cada vez que se aplique y que la muestra de preguntas mida siempre el mismo contenido; de esta manera, si los sustentantes contestan diferentes versiones del mismo examen, la puntuación obtenida siempre será más o menos la misma. La consistencia interna de las pruebas se estima por medio del coeficiente de alfa de Cronbach o la fórmula de Kuder Richardson 20 (KR20). Un estimado de consistencia interna que sea alto indica que los resultados de las pruebas serán iguales o muy semejantes cada vez que se aplique al mismo alumno o a alumnos diferentes. Por otro lado, también significa que los errores aleatorios de la medida son bajos, de manera que estas pruebas han demostrado ser reproducibles (Downing, 2004; Ferguson et al., 2002).

El error estándar de la medida (EEM) es un estimado que se calcula a partir del coeficiente de confiabilidad. Este EEM indica la precisión de la medida, dada la confiabilidad de la prueba, en cada nivel de puntaje seleccionado.

El coeficiente de confiabilidad puede encontrarse entre 0 y 1.0, siendo 1.0 el valor más alto. Según el tipo de examen, se considera un nivel de confiabilidad mínimo indispensable: en el caso de exámenes de altas consecuencias, se espera que el índice sea de 0.90 o mayor; los exámenes sumativos, como los del final de curso, requieren un índice entre 0.80 y 0.89. Además, las evaluaciones formativas o de menores consecuencias pueden presentar un índice entre 0.70 y 0.79.

Al final, lo más importante de la confiabilidad no es el valor absoluto de su coeficiente, sino los falsos positivos o falsos negativos que surjan a partir de la prueba, principalmente en las de altas consecuencias en las que solo hay dos resultados: aprobado/reprobado o aceptado/rechazado. Para estimar la reproducibilidad de este índice en estas circunstancias Subkoviak (citado por Downing, 2004) desarrolló un cálculo para identificar el grado de confianza sobre los resultados. Además, la teoría de la generalizabilidad también permite calcular la precisión de la medición, lo que ayuda a evaluar la exactitud de la evaluación.

Existen varias maneras de mejorar la confiabilidad:

- Usar un número lo suficientemente grande de preguntas.
- Que las preguntas estén bien hechas: redacción clara y sin ambigüedades.
- Que el nivel de las preguntas sea medio.
- Probar las preguntas en otros estudiantes antes del examen real (Downing, 2004; Fraenkel et al., 2019).

4.7 Justicia

La justicia es importante para la validez, ya que la evaluación debe tener el mismo significado para todos los miembros de la población a la que va dirigida (American Educational Research Association et al., 2018). Existen cuatro puntos de vista de justicia: imparcialidad, diferencias en la puntuación, predicción y selección, y validez.

La imparcialidad se refiere a que todos los sustentantes de la prueba recibirán trato equitativo y la misma oportunidad de demostrar sus aptitudes sin importar la raza, el sexo, discapacidades o la etnia, cuando estas características no tengan relevancia para el constructo. Además, ya que la justicia no implica que todos obtengan la misma puntuación, las diferencias en la puntuación no indican que la prueba tenga sesgos; solo debe verificarse que cumpla con el resto de las características de justicia.

Por otro lado, según las definiciones más recientes, un examen justo en cuanto a predicción y selección requiere presentar previamente evidencia perfecta de confiabilidad y validez, lo que no se ve en la realidad. En el caso de que así fuera, los sustentantes que pueden tener éxito tienen la misma probabilidad de ser aceptados a pesar de su origen, pues una prueba es justa cuando presenta evidencia de validez para todos los grupos que la presentan (Lane et al., 2016).

5. MARCO METODOLÓGICO

5.1. Planteamiento del problema y justificación

La selección de los mejores aspirantes para ingresar a la licenciatura en Médico Cirujano es una labor que implica gran responsabilidad tanto con los estudiantes como con la sociedad en general, pues esta decisión impacta sobre quiénes se harán cargo de la salud de la población (Shulruf et al., 2012). Para determinar si los exámenes de admisión poseen las características que mencionan de justicia y discriminación, entre otras, se han realizado análisis en varios países alrededor del mundo, los que permiten conocer su valor predictivo tanto en el desempeño en ciencias básicas como en ciencias clínicas. También distinguen entre los diferentes componentes de cada evaluación, tales como exámenes de conocimientos, psicométricos, entrevistas, promedios del bachillerato y promedios de etapas universitarias previas. Además de que estos análisis se han realizado en medios diferentes al nuestro, la población que han estudiado proviene de universidades públicas y privadas, algunos de ellos en población únicamente masculina, y con otras variables que no aplican en nuestro medio.

Con base en los análisis previos sobre nuestro propio examen de admisión, se considera apropiado y conveniente estudiar el valor de cada componente de este sobre el desempeño que tendrán los alumnos seleccionados durante la carrera. Si resulta ser un buen instrumento de discriminación, garantizará tanto a la Facultad como a la población que contamos con los mejores elementos para educar y formar como médicos, permitiendo enfocarnos en otras áreas de oportunidad para mejorar los resultados. Además, también permite evaluar si las ponderaciones de cada componente del examen de admisión son las más convenientes o si vale la pena modificarlas para mejorar la discriminación entre los aspirantes.

Por otro lado, la integración de los marcos de referencia de validez más utilizados permitirá hacerlos más accesibles a los educadores en ciencias de la salud y facilitará su uso para evaluar no solo los procesos de admisión, sino también otras pruebas de altas consecuencias.

5.2 Preguntas de investigación

¿Cuáles son las aproximaciones metodológicas más efectivas para demostrar la validez de las decisiones en el proceso de admisión para la licenciatura en médico cirujano?

Al aplicar este método, ¿cuál es el grado de validez del proceso de admisión para la licenciatura en médico cirujano en la UASLP?

5.3. Objetivos

5.3.1. General

Desarrollar un método para demostrar la validez de las decisiones con base en el proceso de admisión para la licenciatura en médico cirujano.

5.3.2. Específicos

1. Elaborar una propuesta de método para determinar la validez de los procesos de admisión para las escuelas de medicina en Latinoamérica.
2. Utilizar el método para analizar el grado de validez de los procesos de admisión de 2013 y 2014 de la Facultad de Medicina de la UASLP.

5.4 Diseño de investigación y estrategias metodológicas

1. Tipo de estudio:

- Exploratorio, observacional, retrospectivo (Yuni & Urbano, 2014).

- Dentro de los estudios de exploración, este es de validez (Figuras 6 y 7) (Ringsted et al., 2011).

Figura 6. La brújula de la investigación.[5]

[5] Ringsted et al., 2011.

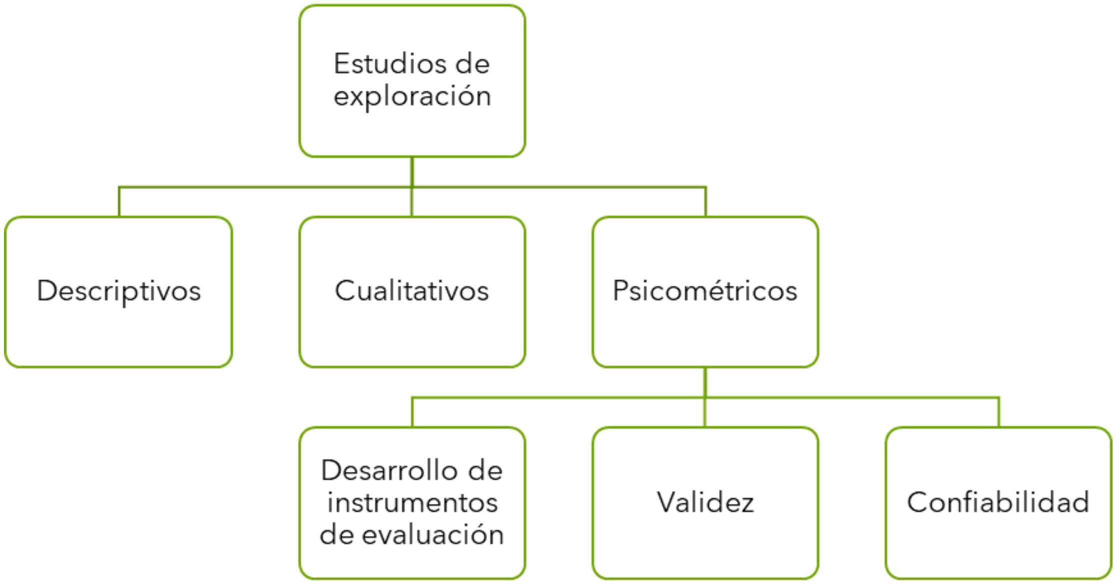


Figura 7. Estudios de exploración.[6]

2. Población:

Criterios de inclusión:

- Aspirantes que ingresaron a la licenciatura de médico cirujano en la FMUASLP que participaron en los procesos de admisión de 2013 y 2014.
- Alumnos aceptados para llevar a cabo los estudios de la licenciatura de Médico Cirujano en la FMUASLP, durante el proceso de admisión de 2013 y 2014.

Criterios de exclusión:

- Aspirantes: no haber contestado alguno de los componentes del proceso de admisión.
- Alumnos: haberse dado de baja antes de terminar el primer año de la carrera.

3. Instrumentos de recolección de información

- Pruebas de aptitud que han presentado los aspirantes a ingresar a la FMUASLP, las que en conjunto componen el examen de admisión

[6] Ringsted et al., 2011.

- Pruebas de aptitud y ejecución que se aplican para determinar las calificaciones en las asignaturas de la carrera de Médico Cirujano en la FMUASLP.
- El Examen General para el Egreso de la Licenciatura en Medicina General que se aplica como parte del proceso de titulación.
- El ECOE que se aplica como parte del proceso de titulación.

4. Revisión bibliográfica de la literatura (Manterola et al., 2023)

- Objetivo general: conocer cuáles son los mecanismos y procesos de admisión para ingresar a la licenciatura en médico cirujano que se implementan en las facultades y escuelas de medicina de México y el mundo.
- Objetivos específicos. Conocer:
    - Mecanismos de admisión.
    - Constructos evaluados y proporción de preguntas o ítems para evaluarlos.
    - Ponderaciones aplicadas.
    - Cantidad de aspirantes y cantidad de alumnos aceptados.
    - Contexto social, político y económico que atañe a cada proceso descrito.
    - Quién aplica las evaluaciones que componen los procesos de admisión.
    - Frecuencia de los procesos de admisión.
    - Análisis de validez realizados sobre los procesos de admisión.
- Bases de datos:
    - MedLine, ERIC, IRESIE, y Google Scholar.
- Criterios de inclusión de los artículos:
    - Estudios publicados entre enero de 2000 y diciembre de 2019. Este periodo de tiempo fue elegido para tener los criterios de admisión a las escuelas de medicina más recientes, sin interferencia de la pandemia causada por COVID-19.
    - Los estudios y revisiones que analizan y describen los métodos de admisión a la licenciatura de médico cirujano y semejantes en las escuelas y facultades de medicina del mundo.
    - Idioma: inglés y español.
    - País: todos.
    - Tipo de publicación: artículos originales, artículos de revisión y tesis.
- Palabras clave:

- College Admission Test: Test designed to identify students suitable for admission into a graduate or undergraduate curriculum (NCBI, 1991a).
- Schools, Medical: Educational institutions for individuals specializing in the field of medicine (NCBI, s/f).
- Validez de una prueba: grado en que una prueba, instrumento, escala de puntaje, cuestionario, etc., es un índice efectivo para lo que es utilizado o pretende medir (Education Resources Information Center, 1966).
- School Admission Criteria: Requirements for the selection of students for admission to academic institutions(NCBI, 1991b).

Se identificaron los artículos de relevancia por medio de las palabras clave y sus sinónimos, y se buscaron los términos MESH o el término más apropiado de los tesauros de cada base de datos. Se utilizó el término Booleano AND para combinar examen de admisión con el resto de los términos. Con base en lo anterior se obtuvieron 42 referencias en total.

5.5. ANÁLISIS ESTADÍSTICO

Con estos datos se realizó estadística descriptiva para el argumento de usos e interpretaciones. Se realizó análisis de regresión logística para conocer la relación de las evaluaciones de admisión con el desempeño académico, y se tomó como variable dependiente a las calificaciones del primer año de la carrera y como variable independiente a las evaluaciones del proceso de admisión.

También se desarrolló un modelo a través de AFC, para identificar las dimensiones que causan la variabilidad de las puntuaciones de los componentes del proceso de admisión. Además, se elaboró un modelo para predecir la calificación en los distintos años a partir de los resultados de los años anteriores y el examen inicial a través de análisis de senderos.

El análisis estadístico se realizó con los programas SPSS 21.0 de IBM y SAS® versión 9.4 y se consideró $p<0.05$ como significativa.

5.6 Consideraciones éticas

Este proyecto fue sometido al comité de Investigación y al comité de Ética en Investigación de la FMUASLP, y recibió los correspondientes dictámenes de aprobación (Anexos 1 y 2):

- Comité de Investigación: CI-006-2019
- Comité de Ética: CEI-2019-004

Se solicitó dispensa de consentimiento informado con base en el artículo 31 del reglamento de transparencia y acceso a la información pública de la UASLP:

> *Artículo 31. No se requerirá del consentimiento de los involucrados para proporcionar los datos personales en los siguientes casos:*
>
> *I. Los necesarios por razones estadísticas, científicas o de interés general previstos en la legislación universitaria, previo procedimiento por el cual no puedan asociarse los datos personales con el individuo a que se refieran* (UASLP, 2009).

Mientras que el comité de Investigación solicitó un reporte anual, el comité de Ética en Investigación requiere anonimizar los datos de los estudiantes, así como informar de la terminación del proyecto.

5.7 Financiamiento

Este proyecto de investigación fue financiado con recursos propios. Se recibió la prestación de año sabático y su renovación correspondiente por parte de la UASLP desde agosto de 2018 hasta julio de 2020.

6. RESULTADOS

6.1. PROPUESTA DE METODOLOGÍA PARA EVALUAR LA VALIDEZ DE LA INTERPRETACIÓN DE LOS RESULTADOS DEL PROCESO DE ADMISIÓN.

Se realizó un análisis de los marcos de referencia de validez, en el que se comparó la manera en que Kane sugiere garantizar las inferencias con las fuentes de evidencia de validez de Messick, así como la teoría subyacente a cada marco de referencia. Así, se establecieron las relaciones como se muestra en la Figura 8. Al considerar los elementos de las fuentes de evidencia, podemos emparejarlos con las inferencias y sus garantías correspondientes de manera más precisa; de esta manera una fuente de evidencia puede contribuir para más de una inferencia.

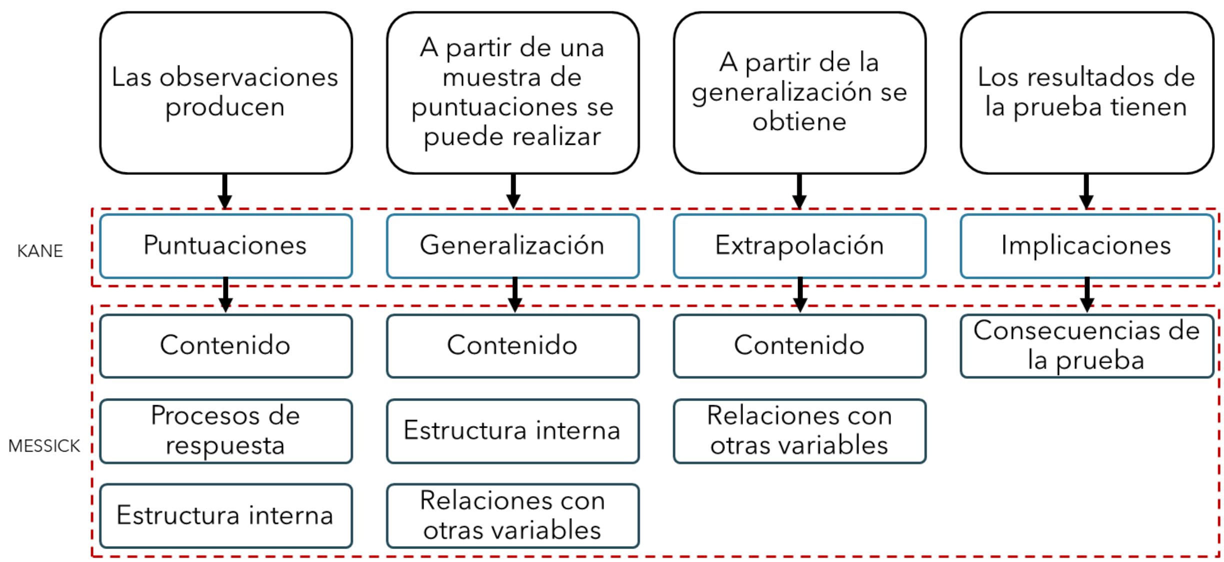


Figura 8. Las inferencias de Kane pueden probarse por medio de las fuentes de evidencia de Messick.[7]

A continuación se explica con mayor detalle cuáles son los elementos de las fuentes de evidencia de Messick que contribuyen a probar las inferencias de Kane (Tabla 5). La inferencia de puntuación depende de que se lleven a cabo procedimientos de calificación apropiados y libres de

[7] Elaboración propia.

sesgo, además de aplicar las reglas de puntuación como se precisó en la tabla de especificaciones de la prueba. Así, encontramos que puede probarse por medio de tres fuentes de evidencia: contenido, procesos de respuesta y estructura interna. La fuente de contenido contribuye con:

- Definición de dominio, demostrada a través de un análisis de la tabla de especificaciones de la prueba: que exista una descripción clara del contenido a evaluar, sus subcategorías y subclasificaciones, la proporción de las preguntas, el nivel cognitivo, y las credenciales de quienes elaboran los reactivos.
- Relevancia del dominio: análisis de las mejores prácticas para la escritura de los ítems y revisión cultural.
- Desarrollo apropiado del instrumento: un panel de expertos ha revisado los ítems y se explica cuando hay ausencia de algún aspecto del constructo.

Por otro lado, la fuente basada en procesos de respuesta prueba que las puntuaciones reflejen el proceso cognitivo correcto. Finalmente, la dimensionalidad (analizada por medio de TRI o AFC) y la invarianza de la medida (estudiada a través de AFE o FDI) aportan a la estructura interna.

		KANE			
		Puntuaciones	Generalización	Extrapolación	Implicaciones
MESSICK	Contenido	Definición del dominio Relevancia del dominio Desarrollo apropiado del instrumento	Representación del dominio	Relevancia del dominio	-
	Procesos de respuesta	Prueba del proceso cognitivo correcto	-	-	-
	Estructura interna	Dimensionalidad Invarianza de la medida	Invarianza de la medida Confiabilidad	-	-
	Relación con otras variables	-	Correlación convergente Correlación divergente	Relación prueba-criterio Generalización de la validez	-
	Consecuencias	-	-	-	Consecuencias voluntarias o involuntarias

Tabla 5. Elementos de las fuentes de evidencia de Messick y su correspondencia con las inferencias de Kane.

A partir de una muestra de puntuaciones observadas es posible generalizar, lo que se puede demostrar por medio de las fuentes de contenido, estructura interna y relación con otras variables.

La representación del dominio servirá para demostrar que el número de ítems es una muestra apropiada del universo de ítems posibles, y así contribuye a la evidencia de contenido. La estructura interna se puede probar por medio de la invarianza de la medida (utilizando teoría G, modelos lineales o no lineales) y confiabilidad (alfa de Cronbach o KR 20 o 21). La evidencia de correlación convergente o divergente será el aval de la relación con otras variables.

A partir del universo de ítems, la interpretación puede extrapolarse hacia situaciones de la vida real. La inferencia de extrapolación puede establecerse a través de las fuentes de contenido y relación con otras variables. La relevancia del dominio, un aspecto del contenido, requiere verificar si el instrumento ha evaluado correctamente el contenido descrito en la tabla de especificaciones de la prueba, y también las subcategorías y subclasificaciones del contenido. La relación de la prueba con un criterio relevante y la generalización de la validez son los aspectos de la relación con otras variables que se pueden estudiar para garantizar la extrapolación.

Finalmente, la interpretación de los resultados del instrumento tiene implicaciones, inferencia que se relaciona con la fuente de consecuencias.

PROPUESTA DE MÉTODO

La integración de la aportación de Russell (Russell, 2022) a los marcos de referencia de Kane y de Messick proporciona el orden y organización de las fuentes de evidencia cuya ausencia habían señalado varios autores (Cook & Hatala, 2016).

Tomamos en cuenta los dos pasos del marco de referencia de Kane. El primer paso lo dejamos igual, con la descripción del argumento de usos e interpretaciones, mientras que en el segundo paso integramos las tres etapas del modelo de Russell (Figura 9):

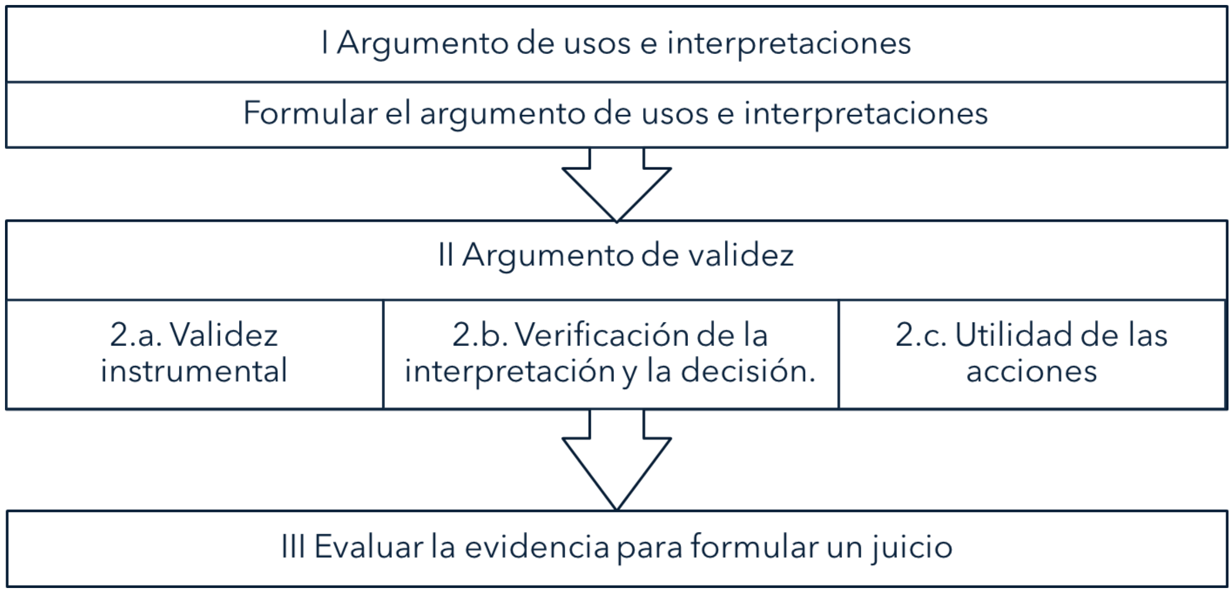


Figura 9. Pasos del método.

En la Tabla 6 se observan cuáles son las fuentes de evidencia de Messick que aportan para cada inferencia de Kane en las tres etapas del paso de Argumento de Validez.

1. Argumento de usos e interpretaciones

1.a. Formular el argumento de usos e interpretaciones.

Kane enfatizó en que el proceso de validación siempre debe iniciar con el argumento de usos e interpretaciones (Kane, 1992, 2009), momento en el que se especifica el objetivo del instrumento, los usuarios propuestos, los usos que se darán a las puntuaciones, el constructo que se pretende medir, y la interpretación de las puntuaciones (con referencia a norma o a criterio). Después de que se ha aplicado el instrumento, se describen las características de la población de sustentantes, como edad y sexo.

Messick (1989) indicó que cada proceso de validación debe iniciar con una hipótesis que debe probarse por medio de las diferentes fuentes de evidencia de validez. Por esta razón en este paso se establecen las hipótesis basadas en las inferencias de Kane.

El análisis de la evidencia se llevará a cabo a través de un argumento de validez, el que puede delinearse con base en las tres etapas de Russell. Estas se explican a continuación.

II.A. VALIDEZ INSTRUMENTAL

En esta etapa se debe cuestionar si el instrumento cumple su objetivo de apoyar una hipótesis acerca del constructo medido. Para comprobar esta pregunta, se debe reunir evidencia acerca de las características psicométricas de las puntuaciones, representación del constructo y los efectos relevantes e irrelevantes del constructo, tales como la SC y la VIC. Como esta etapa se debe completar antes de que se aplique la prueba, el responsable de proveer las evidencias es el desarrollador de la prueba. Considerando lo anterior, la fuente de evidencia de validez de contenido debe aportar la información necesaria para comprobar las hipótesis de las inferencias de puntuación, generalizabilidad y extrapolación de Kane, tal como se especifica en la Tabla 5. Esto será por medio de la información acerca del dominio que será evaluado: su definición, el nivel cognitivo, que la muestra de ítems o casos clínicos sea representativa, etc.

II. B. VERIFICACIÓN DE LA INTERPRETACIÓN Y DECISIÓN

Esta etapa implica el análisis sobre lo apropiado de la interpretación de los resultados de la prueba: si lo es, entonces esta interpretación puede informar para tomar una decisión acerca de las acciones subsecuentes. Este estudio se lleva a cabo una vez que se ha aplicado el instrumento, lo que permite el análisis de las puntuaciones para saber si son precisas y si pueden utilizarse para separar a los sustentantes en grupos y si las decisiones son apropiadas para cada grupo. También se debe verificar el grado en que las decisiones acerca de cada grupo correlacionan con sus habilidades o necesidades. Nuevamente son útiles las inferencias de puntuación, generalizabilidad y extrapolación, y su comprobación será a través de la estructura interna, los procesos de respuesta y las relaciones con otras variables. De esta manera se puede explorar si las puntuaciones reflejan el proceso cognitivo esperado, determinar la dimensionalidad del instrumento, la invarianza de la medida, su confiabilidad, y buscar correlaciones en las relaciones convergente-divergente o prueba-criterio.

II. C. Utilidad de las acciones

La última etapa implica evaluar la utilidad de las acciones, es decir, sus consecuencias, ya sean voluntarias o involuntarias y el grado en que las positivas superan a las negativas. La inferencia de implicaciones se puede comprobar a través de la fuente de evidencia de consecuencias al recolectar la información necesaria por medio de grupos focales o entrevistas, por ejemplo.

En este paso de validación ya se cuenta con las hipótesis para ser probadas y las fuentes de evidencia para lograrlo, de manera que debe desarrollarse un plan para reunir esta información y para analizarla. Esto depende de la cantidad de evidencia necesaria para cada tipo de instrumento y su trascendencia: mientras mayor sea el impacto de la evaluación, se necesita más evidencia y de mayor calidad.

2. c. Evaluar la evidencia y formular un juicio.

Finalmente, los resultados del análisis deben ser estudiados para producir un juicio acerca del grado de validez de los usos e interpretaciones de las puntuaciones. Esto se logra al determinar la aceptación o rechazo de las hipótesis establecidas al principio, es decir, la aceptación o rechazo de la validez de la interpretación de la prueba.

Cabe mencionar que durante todo el proceso de validación se deben evitar las amenazas a la validez de manera activa, no solamente durante la etapa de formulación de hipótesis.

	I Validez instrumental.					II. Verificación de la interpretación y la decisión.						III. Utilidad de las acciones		
	A. Puntuación			**B. Generalizabilidad**	**C. Extrapolación**	**A. Puntuación**		**B. Generalizabilidad**		**C. Extrapolación.**		**D. Implicaciones**		
Inferencias de Kane e hipótesis	La regla de puntuación es apropiada La calidad de redacción y del formato de los ítems es buena.			Los ítems son una muestra adecuada del dominio.	La puntuación observada se relaciona con un criterio relevante.	La puntuación refleja el proceso cognitivo apropiado. El instrumento midió la(s) dimensión(es) planeada(s).		El tamaño de la muestra fue adecuado. No existen/sí existen diferencias entre grupos. Confiabilidad. La puntuación observada se relaciona con la de instrumentos que miden el mismo constructo/no se relaciona con la de instrumentos que miden un constructo diferente.		El proceso de admisión predice el primer año de la carrera. La relación con otras variables se puede generalizar a un dominio más amplio del conocimiento.		La interpretación de los resultados es adecuada. Los alumnos tienen éxito académico en la carrera. Es razonable y justificable el tipo de escala de la prueba		
Fuente de evidencia de Messick	1. Contenido					2. Procesos de respuesta	3. Estructura interna			4. Relación con otras variables		5. Consecuencias		
Elemento	a. Definición del dominio	b. Proceso de desarrollo del instrumento	c. Relevancia del dominio	d. Representación del dominio	d. Representación del dominio	No aplica	a. Dimensionalidad	b. Invarianza de la medida	c. Confiabilidad	a. Relación prueba-criterio	b. Generalización de la validez	No aplica		
Evidencia para un Instrumento con POM	Ofrece detalles con respecto de lo que la prueba mide. Transforma el constructo teórico en un dominio de contenido concreto.	Propiedades de los ítems.	Grado en que cada ítem en una prueba es relevante para el dominio que se evalúa.	Expertos. Alineación con currículo.	Alineación de complejidad moderada: contenido y nivel cognitivo evaluado.	Demuestra que se están llevando a cabo los procesos cognitivos esperados al contestar la prueba.	Identificar cuántas dimensiones los ítems.	Análisis del ítem: dificultad y discriminación. Ausencia de sesgo sistemático: demuestra que las puntuaciones serán las mismas cuando se comparen entre grupos con diferentes características como raza, edad o sexo, de acuerdo con la taxonomía de equivalencia.	En cada ocasión que se aplica el instrumento se obtendrán resultados semejantes.	Las puntuaciones predicen el desempeño en un criterio relevante, de manera predictiva o concurrente.	Metaanálisis de los estudios de validación anteriores en condiciones semejantes.	Los usuarios están de acuerdo con la interpretación de los resultados.	Establecer la trayectoria académica de los alumnos y de la generación en conjunto.	Utilidad de las escalas con referencia a la norma.
Lista de cotejo para	La tabla de especificaciones de la prueba:	Revisión de los ítems por expertos que aseguren su exactitud técnica.	Revisión de ítems por expertos que evalúen:	Expertos externos e independientes califican cada ítem para determinar si: •	Explicación de los dominios del conocimiento que deben ser	Entrevistas cognitivas, pensar en voz alta.	AFC	Índice de dificultad por ítem y general de la prueba.	Alfa de Cronbach o KR20-21.	Análisis de senderos.	Metaanálisis de los estudios de validación anteriores en	Encuesta a los usuarios: alumnos,	Calificaciones de cada alumno para cada	Bibliografía en donde se demuestre la utilidad

demostrar las hipótesis	• Describe detalladamente las áreas del contenido y las habilidades cognitivas para cuya medición se ha diseñado el instrumento. • Enlista las subáreas y los niveles cognitivos que se miden. • Muestra los estándares específicos de contenido, objetivos curriculares, o habilidades contenidas dentro de los diferentes niveles cognitivos.	Revisión de los ítems por expertos de medición para determinar qué tan bien se adhieren a los estándares de principios de escritura de ítems de calidad. Revisión de sensibilidad para evitar varianza irrelevante al constructo. Piloteo de los ítems con análisis estadístico para seleccionar los ítems más apropiados para uso operativo.	• Que todos los aspectos importantes del dominio sean medidos por la prueba. • Si la prueba presenta contenido trivial o irrelevante.	Representa completa y suficientemente al dominio. • Concuerda con el estándar de contenido o un elemento de la tabla de especificaciones de la prueba. • Concuerda con el nivel cognitivo que se pretende alcanzar durante las clases.	evaluados para el ingreso a la licenciatura en Medicina y los porcentajes necesarios para cada campo. Demostrar la combinación de constructos que deben ser evaluados en los aspirantes a ingresar a la licenciatura en Medicina.	Tiempos de respuesta.		Funcionamiento diferencial del ítem.			condiciones semejantes	profesores y administrativos de la facultad.	asignatura de 1º a 5º año.	del uso de la escala con referencia a la norma en este tipo de evaluaciones.

Tabla 6. Pasos 2.a. Identificar y establecer las hipótesis y 2.b. Crear un plan para probarlas.

CONTEXTO DE LA UASLP

De 2011 a 2014 la demanda por espacios educativos en la UASLP incrementó en 16.46%, mientras que el número de espacios educativos ascendió en 17.08%. A pesar de que el aumento de espacios educativos superó al de la demanda, solo se aceptaba el 45.32% de los aspirantes en la UASLP en 2014 (UASLP, 2012, 2013, 2014b, 2015a, 2016) (Figura 10 y Tabla 7).

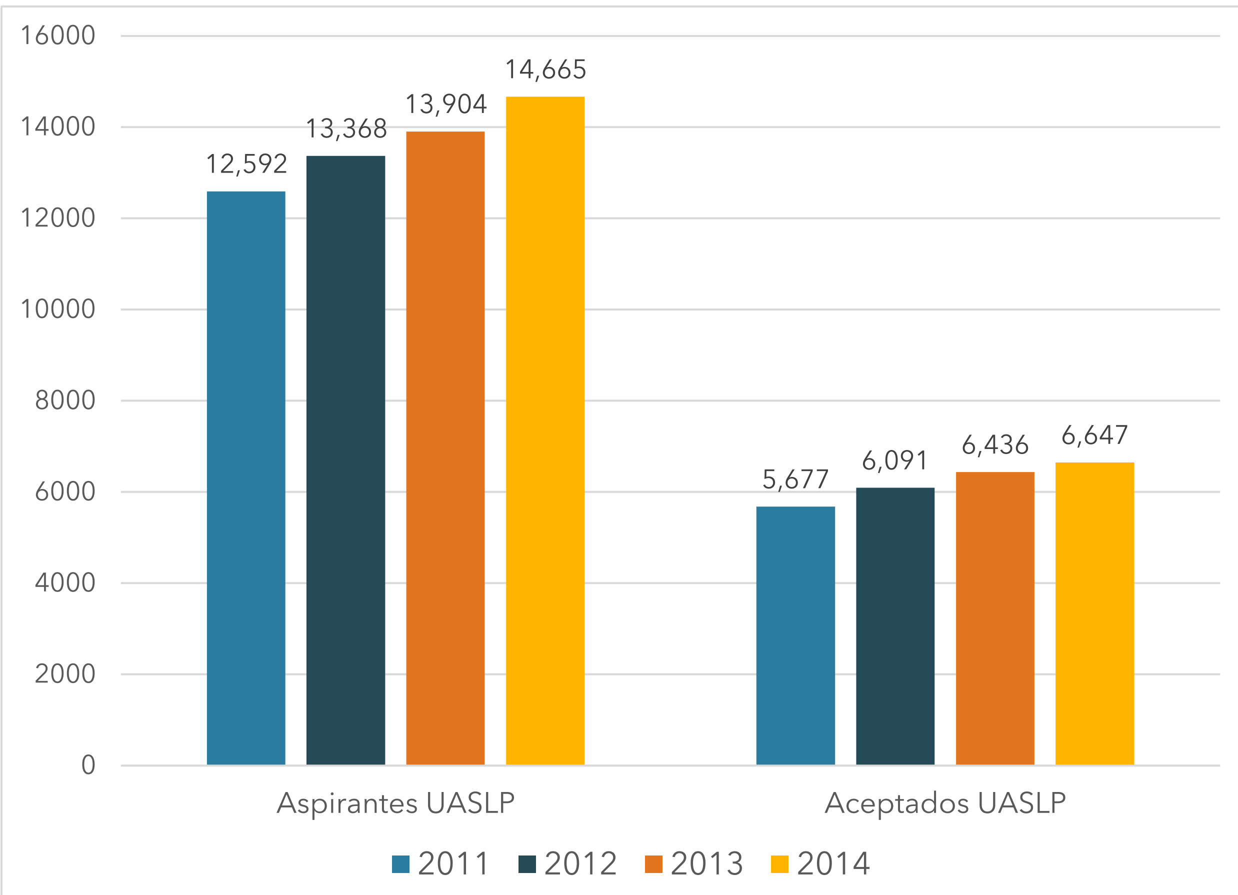


Figura 10. Número de aspirantes y aceptados en la UASLP de 2011 a 2014.

En el mismo lapso de tiempo, la demanda por espacios educativos en la Facultad de Medicina ha aumentado en un 29.69%; sin embargo, el incremento de lugares para estudiar la carrera de Médico

Cirujano solo ha sido de 10% , de 132 a 145 lugares, lo que corresponde al 8.97% de los aspirantes en 2014 (Figura 11 y Tabla 7). Esto ocurre así pues, a pesar de que se incrementó la cantidad de espacios educativos en la UASLP en 17% y para la Facultad de Medicina en 9.8%, la cantidad de solicitantes para la UASLP aumentó en 16% y para la Facultad de Medicina en 29.7%.

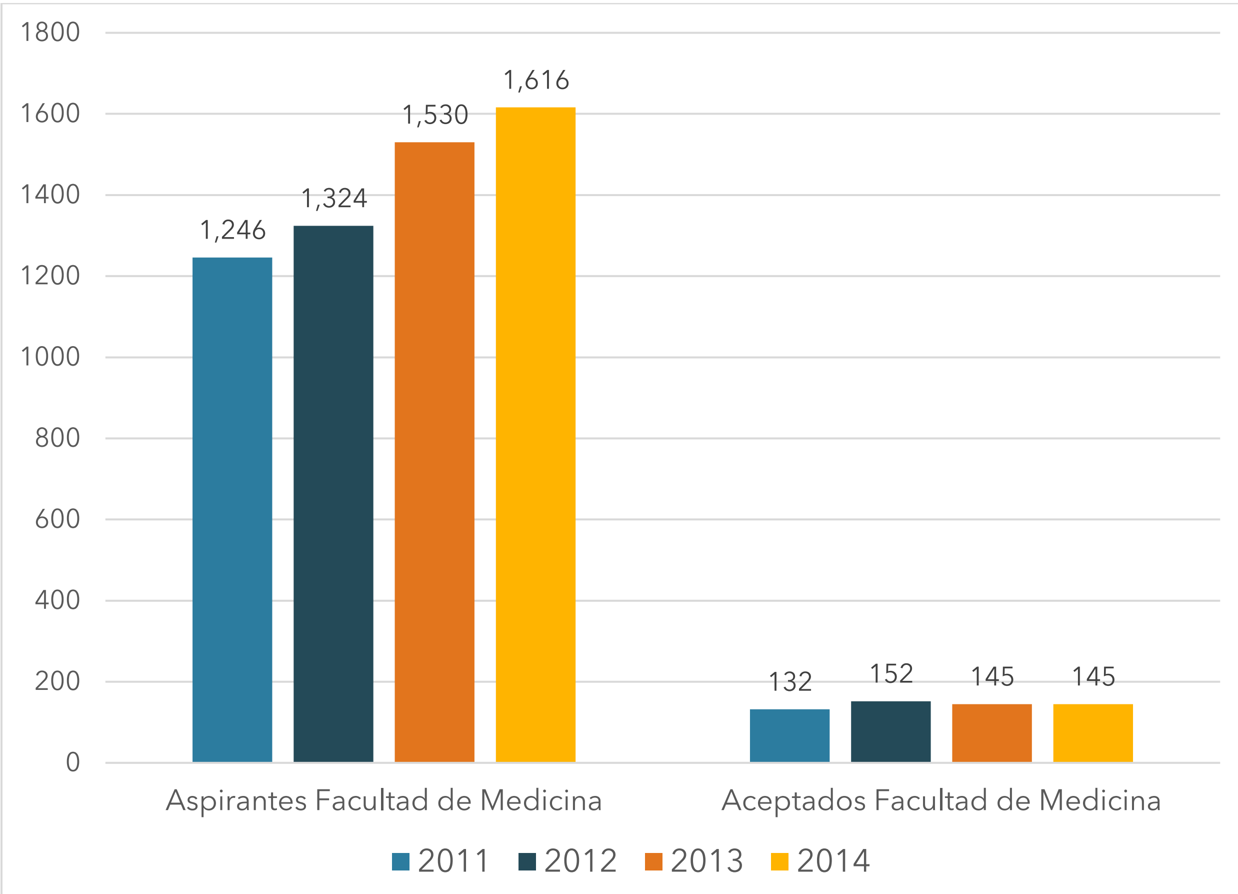


Figura 11. Número de aspirantes y aceptados en la Facultad de Medicina de la UASLP de 2011 a 2014.

	2013	**2014**
UASLP	46.28	45.32
FM	9.47	8.97

Tabla 7. Porcentaje de alumnos aceptados en total en la UASLP y en la Facultad de Medicina.

Para determinar la admisión de los aspirantes se lleva a cabo una evaluación que consta de tres componentes que aportan diferentes porcentajes para establecer la calificación final: el examen psicométrico (EP - 15%), examen de conocimientos (EC - 45%), y el EXANI-II del CENEVAL (40%) (UASLP, 2017a).

Estos tres componentes poseen objetivos diferentes y evalúan constructos diferentes, por lo que su desarrollo debe analizarse de manera independiente. El grado de validez de cada componente será valorado al final en conjunto para determinar el grado de validez del proceso en general (Figura 12).

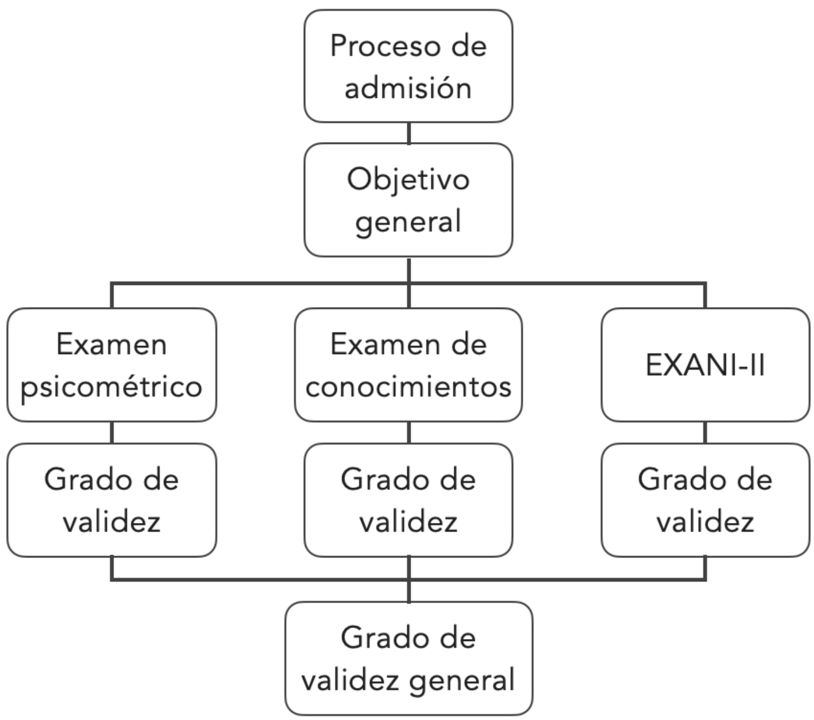


Figura 12. Componentes del proceso de admisión y su análisis.

De esta manera, se aplicó la metodología propuesta:

	UASLP	Examen psicométrico	Examen de conocimientos	EXANI-II
a. El objetivo de la prueba	"…asegurar que los aspirantes que ingresen a la universidad cuenten con las competencias requeridas para la realización de sus estudios, utilizando procesos confiables, rigurosos, certificados, transparentes y en constante modernización." (UASLP, 2014c).	Son "…evaluaciones psicológicas estandarizadas que miden las aptitudes básicas para el estudio" (UASLP, 2015b)	"Evalúa los conocimientos, las destrezas y las habilidades requeridas de los aspirantes a ingresar de acuerdo con el perfil del alumno pretendido en las licenciaturas." (UASLP, 2014c)	"…establecer el nivel de potencialidad de un individuo para lograr nuevos aprendizajes…" (Centro Nacional de Evaluación para la Educación Superior, 2013a).
b. Los usuarios propuestos	Facultad de Medicina de la UASLP			
c. Los usos	Seleccionar a "los aspirantes que cuentan con las competencias requeridas para la realización de sus estudios de licenciatura" en médico cirujano (UASLP, 2014c).			
d. El constructo medido		"1. Razonamiento verbal.- evalúa la capacidad de utilizar sinónimos y antónimos, paráfrasis incompletas, y definiciones de palabras. 2. Retención y comprensión.- mide la habilidad para comprender y retener por medio de lectura de comprensión. 3. Razonamiento abstracto.- valora la capacidad de razonamiento lógico e inmediato ante problemas cotidianos por medio de figuras y búsqueda de semejanzas y diferencias." (UASLP, 2015b).	"Explora la capacidad de comprensión y razonamiento en 5 diferentes áreas del conocimiento: Físico–Matemático, Biología, Química, Inglés y Español" (UASLP, 2017b).	"…evalúa la habilidad para analizar y resolver problemas con base en principios elementales de las matemáticas. el sustentante debe generalizar, abstraer, clasificar y emplear su imaginación espacial para solucionar expresiones matemáticas; situaciones que requieren operaciones algebraicas, aritméticas, trigonométricas y geométricas elementales; y problemas que involucran series con elementos visuales y alfanuméricos. Mide también la capacidad de comunicación del sustentante: su comprensión, interpretación y estructuración de mensajes con sentido, expresados en la lengua materna; así como su habilidad para el manejo de herramientas informáticas y computacionales que le permiten obtener, transmitir e intercambiar información en diferentes niveles." (Centro Nacional de Evaluación para la Educación Superior, 2013a).
e. Las interpretaciones de los resultados	Con referencia a norma.			

Tabla 8. Objetivo e interpretación de los resultados de los componentes del proceso de admisión.

- 2013

En 2013 1,373 aspirantes contestaron las tres evaluaciones, 748 mujeres y 625 hombres (Tabla 9). De ellos, ingresaron 145 alumnos, 76 mujeres y 69 hombres.

	Total	M	%	H	%
Aspirantes	1,373	748	54.5	625	45.5
Admitidos	145	76	52.4	69	47.6

Tabla 9. Sexo de aspirantes y alumnos admitidos en 2013.

M=mujeres, H=hombres.

Los sustentantes tenían una edad mínima de 16 años y máxima de 40 años con media de 18.92 y desviación estándar de 3.07. Presentaron una calificación del proceso total de admisión mínima de 13.17 y máxima de 82.09 con media de 56.05 y desviación estándar de 133.31 (Tabla 10 y Figuras 13 y 14).

	Mínimo	Máximo	Media	Desv. Std.	Varianza
Edad	16	40	18.92	1.75	3.07
Calificación total del proceso de admisión	13.17	82.09	56.05	11.54	133.31

Tabla 10. Estadística descriptiva de los sustentantes en 2013.

Los alumnos admitidos tenían una edad mínima de 17 años y máxima de 36 años con media de 19.33 y desviación estándar de 2.10. Presentaron una calificación del proceso total de admisión mínima de 70.54 y máxima de 82.09 con media de 74.31 y desviación estándar de 2.66 (Tabla 11 y Figuras 13 y 14).

	Mínimo	Máximo	Media	Desv. Std.	Varianza
Edad	17	36	19.33	2.10	4.40
Calificación total del proceso de admisión	70.54	82.09	74.31	2.66	7.08

Tabla 11. Estadística descriptiva de los alumnos admitidos en 2013.

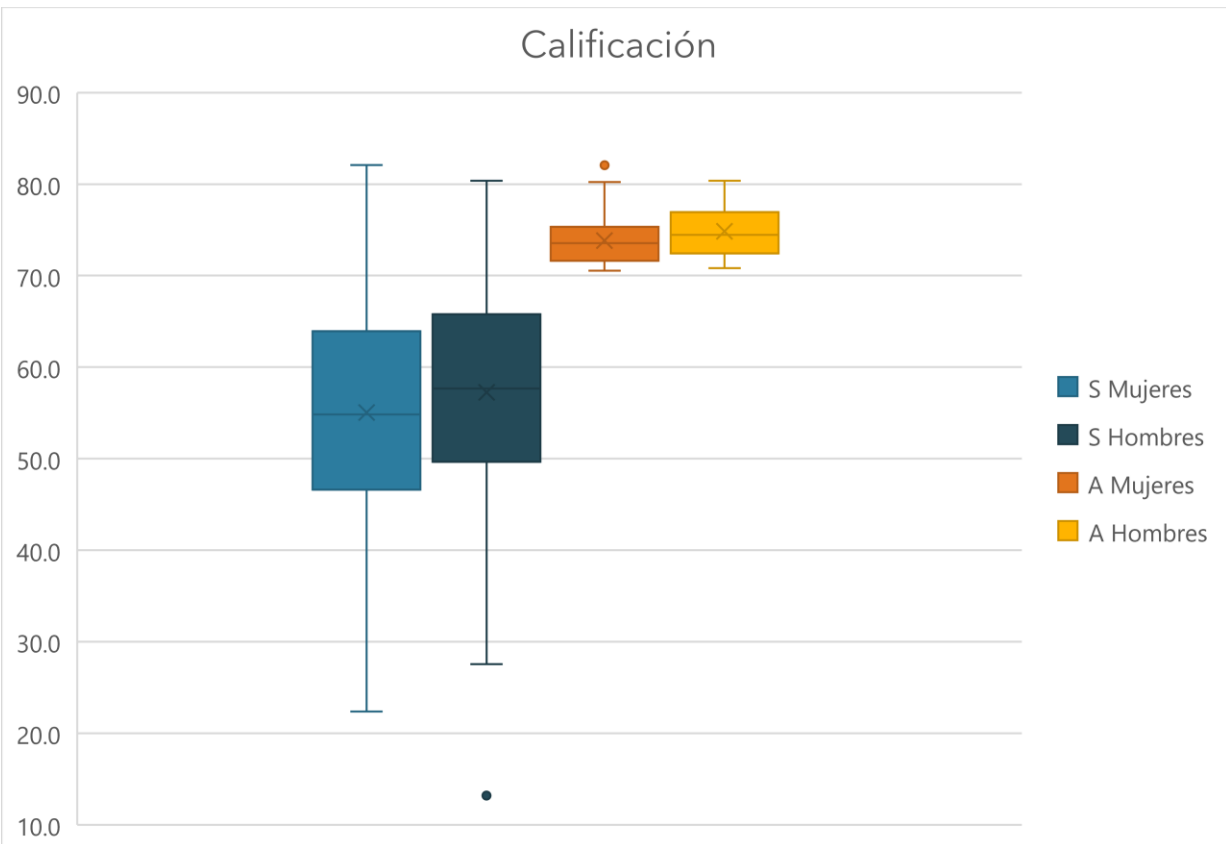


Figura 13. Calificaciones de los sustentantes y estudiantes admitidos por sexo en 2013.

(S=sustentantes, A=admitidos)

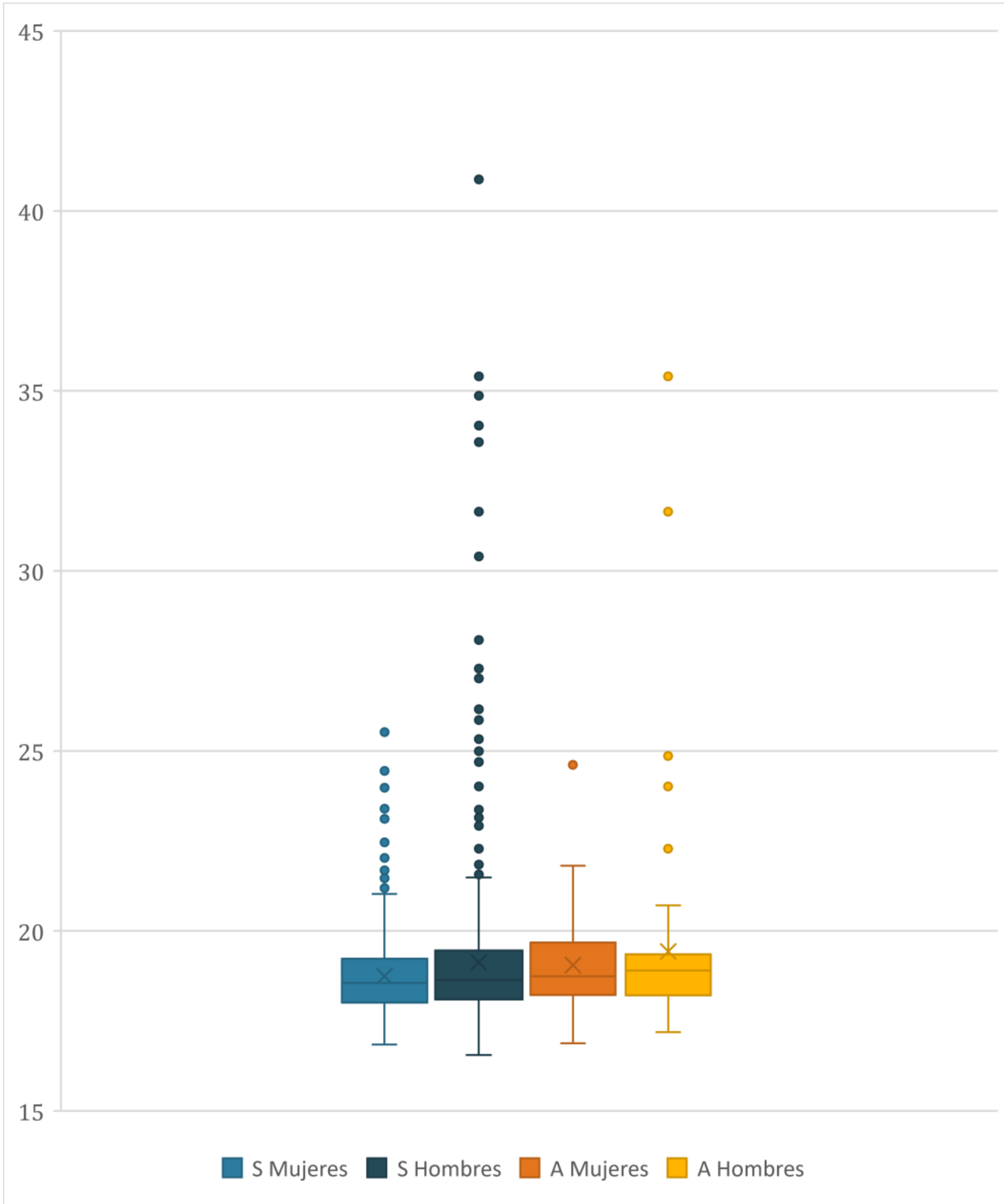


Figura 14. Edad de los sustentantes y estudiantes admitidos por sexo en 2013.

(S=sustentantes, A=admitidos)

- 2014

En la generación de 2014 se presentaron 1,554 sustentantes, 916 mujeres y 638 hombres. Ingresaron 145 alumnos, 66 mujeres y 79 hombres (Tabla 12).

	Total	M	%	H	%
Aspirantes	1,554	916	59.0	638	41.0
Admitidos	145	66	45.5	79	54.5

Tabla 12. Sexo de aspirantes y alumnos admitidos en 2014.

M = mujeres, H = hombres.

La edad mínima de los aspirantes fue de 16 años y la máxima de 55, con una media de 18.91 y desviación estándar de 1.79. La calificación mínima del proceso de admisión fue de 23.69 y la máxima de 85.42, con media de 56.11 y desviación estándar de 11.91 (Tabla 13 y Figuras 15 y 16).

	Mínimo	Máximo	Media	Desv. Std.	Varianza
Edad	16	55	18.91	1.79	3.2
Calificación total del proceso de admisión	23.69	85.42	56.11	11.91	141.97

Tabla 13. Estadística descriptiva de los sustentantes en 2014.

Con respecto de los alumnos admitidos, la edad mínima fue de 17 años y la máxima de 24, con una media de 19.25 y desviación estándar de 1.10. La calificación mínima del proceso de admisión fue de 71.86 y la máxima de 85.42, con media de 76.05 y desviación estándar de 3.12 (Tabla 14 y Figuras 15 y 16).

	Mínimo	Máximo	Media	Desv. Std.	Varianza
Edad	17	24	19.25	1.10	1.21
Calificación total del proceso de admisión	71.86	85.42	76.05	3.12	9.75

Tabla 14. Estadística descriptiva de los alumnos admitidos en 2014.

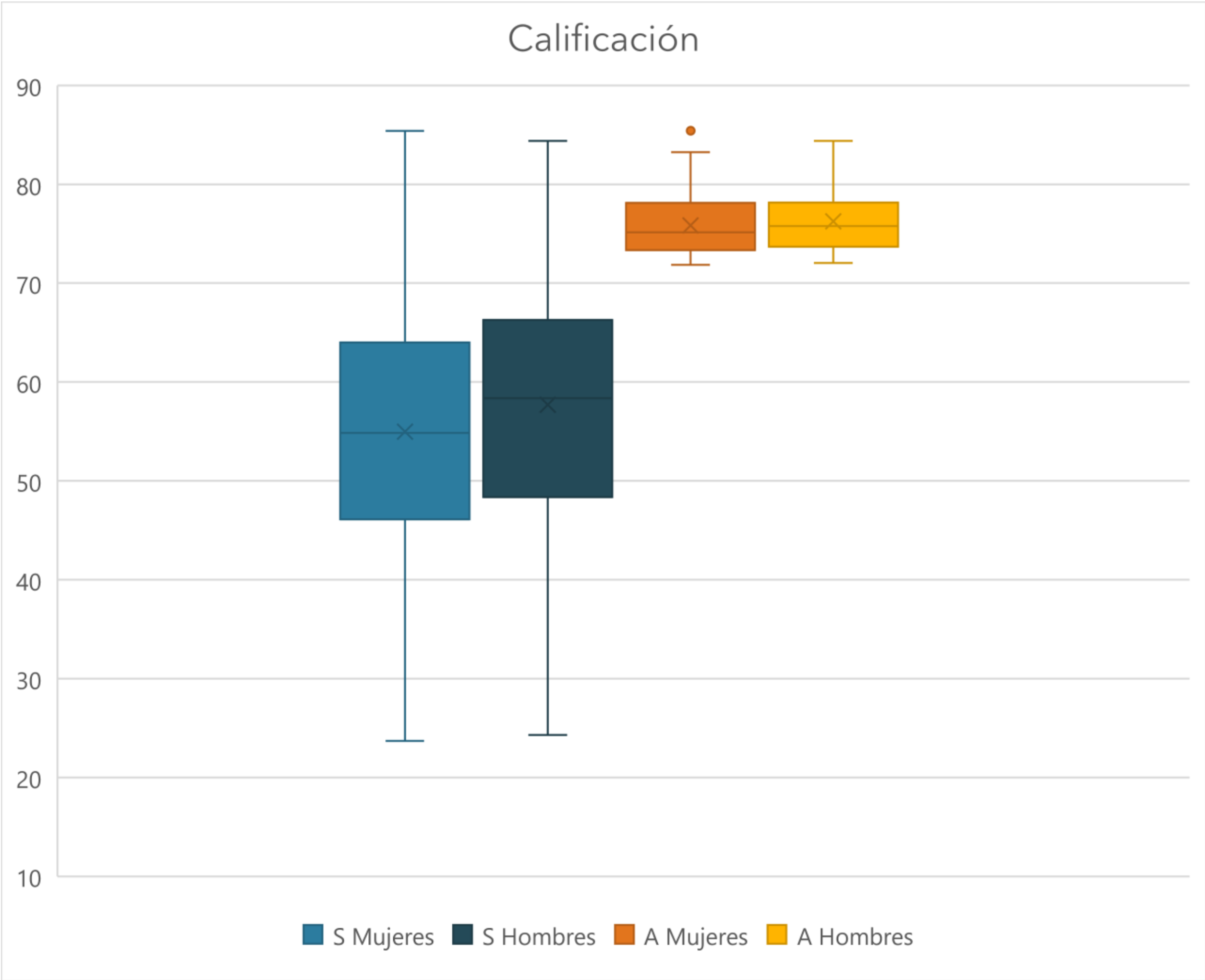


Figura 15. Calificaciones de los sustentantes y estudiantes admitidos por sexo en 2014.

(S=sustentantes, A=admitidos)

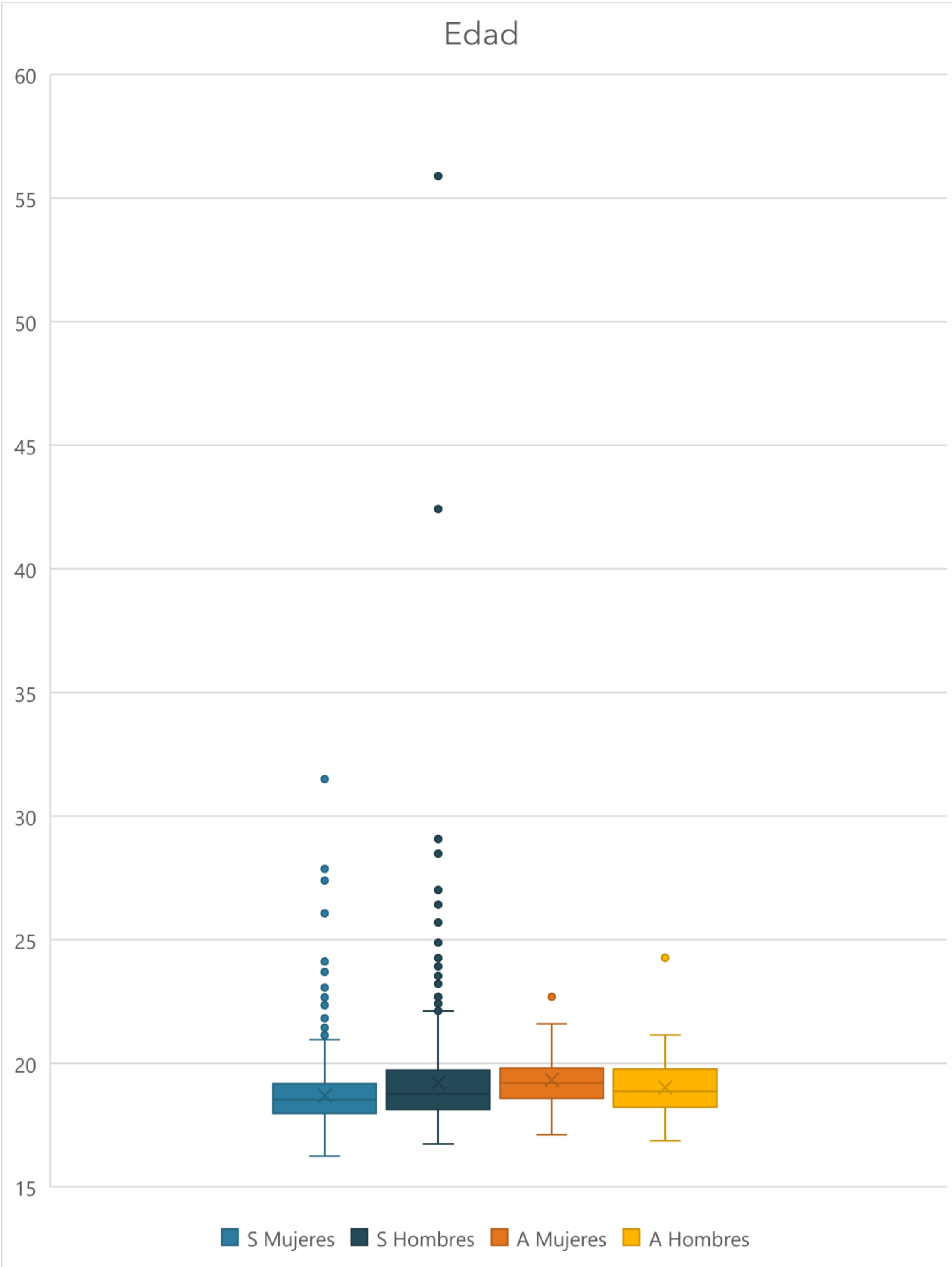


Figura 16. Edad de los sustentantes y estudiantes admitidos por sexo en 2014.

(S=sustentantes, A=admitidos)

Con respecto del tipo de preparatoria de origen, en 2013 el 58% de los estudiantes admitidos provenía de una escuela privada, mientras que en 2014 alcanzó casi el 63% (Figura 17).

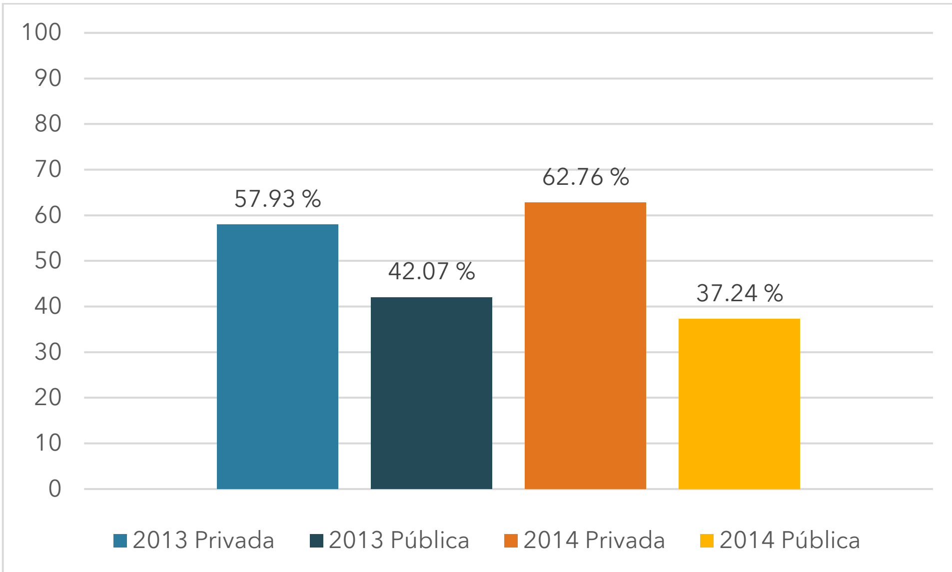


Figura 17. Tipo de preparatoria de origen de los alumnos admitidos por generación.

En este paso se identifican y establecen las hipótesis, se crea el plan de recolección de evidencia y se demuestran las hipótesis.

I. VALIDEZ DEL INSTRUMENTO

Las inferencias de Kane, sus hipótesis correspondientes, las fuentes de evidencia de Messick para comprobarlas, y el plan de recolección de evidencia se encuentran en la Tabla 15:

Kane	**Inferencias e hipótesis**	**A. Puntuación** La regla de puntuación es apropiada La calidad de redacción y del formato de los ítems es buena.			**B. Generalizabilidad** Los ítems son una muestra adecuada del dominio.	**C. Extrapolación** La puntuación observada se relaciona con un criterio relevante.
Messick	**Fuente de evidencia**	1. Contenido				
	Elemento	a. Definición del dominio	b. Proceso de desarrollo del instrumento	c. Relevancia del dominio	d. Representación del dominio	d. Representación del dominio
	Evidencia para un Instrumento con POM	Ofrece detalles con respecto de lo que la prueba mide. Transforma el constructo teórico en un dominio de contenido concreto.	Propiedades de los ítems.	Grado en que cada ítem en una prueba es relevante para el dominio que se evalúa.	Expertos. Alineación con currículo.	Alineación de complejidad moderada: contenido y nivel cognitivo evaluado.
	Lista de cotejo para demostrar las hipótesis	La tabla de especificaciones de la prueba: • Describe detalladamente las áreas del contenido y las habilidades cognitivas para cuya medición se ha diseñado el instrumento. • Enlista las subáreas y los niveles cognitivos que se miden. • Muestra los estándares específicos de contenido, objetivos curriculares, o habilidades contenidas dentro de los diferentes niveles cognitivos.	Revisión de los ítems por expertos que aseguren su exactitud técnica. Revisión de los ítems por expertos de medición para determinar qué tan bien se adhieren a los estándares de principios de escritura de ítems de calidad. Revisión de sensibilidad para evitar varianza irrelevante al constructo. Piloteo de los ítems con análisis estadístico para seleccionar los ítems más apropiados para uso operativo.	Revisión de ítems por expertos que evalúen: • Que todos los aspectos importantes del dominio sean medidos por la prueba. • Si la prueba presenta contenido trivial o irrelevante.	Expertos externos e independientes califican cada ítem para determinar si: • Representa completa y suficientemente al dominio. • Concuerda con el estándar de contenido o un elemento de la tabla de especificaciones de la prueba. • Concuerda con el nivel cognitivo que se pretende alcanzar durante las clases.	Explicación de los dominios del conocimiento que deben ser evaluados para el ingreso a la licenciatura en Medicina y los porcentajes necesarios para cada campo. Demostrar la combinación de constructos que deben ser evaluados en los aspirantes a ingresar a la licenciatura en Medicina.

Tabla 15. Plan de recolección de evidencias para la etapa 1. Validez del instrumento.

FUENTES DE EVIDENCIA PARA DEMOSTRAR LAS HIPÓTESIS.

Con base en la Tabla 15, se ofrecen las fuentes de evidencia para demostrar las hipótesis de la validez del instrumento:

I. A. 1. A. PUNTUACIÓN – CONTENIDO – DEFINICIÓN DEL DOMINIO.

<table>
<tr><th colspan="4">2013</th><th colspan="4">2014</th></tr>
<tr><td>Examen</td><td>Total de ítems</td><td>Sub-componente</td><td># de ítems</td><td>Examen</td><td>Total de ítems</td><td>Sub-componente</td><td># de ítems</td></tr>
<tr><td rowspan="4">EP</td><td rowspan="4">154 POM de 5 opciones</td><td>Razonamiento verbal</td><td>30</td><td rowspan="4">EP</td><td rowspan="4">163 POM de 5 opciones</td><td>Razonamiento verbal</td><td>50</td></tr>
<tr><td>Lectura</td><td>20</td><td>Lectura</td><td>26</td></tr>
<tr><td>Razonamiento abstracto 1</td><td>60</td><td>Razonamiento abstracto 1</td><td>60</td></tr>
<tr><td>Razonamiento abstracto 2</td><td>44</td><td>Razonamiento abstracto 2</td><td>27</td></tr>
<tr><td rowspan="5">EC</td><td rowspan="5">160 POM de 5 opciones</td><td>Biología</td><td>35</td><td rowspan="5">EC</td><td rowspan="5">160 POM de 5 opciones</td><td>Biología</td><td>35</td></tr>
<tr><td>Química</td><td>35</td><td>Química</td><td>35</td></tr>
<tr><td>Español</td><td>25</td><td>Español</td><td>25</td></tr>
<tr><td>Inglés</td><td>35</td><td>Inglés</td><td>35</td></tr>
<tr><td>Físico-matemático</td><td>30</td><td>Físico-matemático</td><td>30</td></tr>
<tr><td rowspan="4">EXANI-II</td><td rowspan="4">100 POM de 4 opciones</td><td>Razonamiento verbal</td><td>25</td><td rowspan="4">EXANI-II</td><td rowspan="4">100 POM de 4 opciones</td><td>Pensamiento matemático</td><td>25</td></tr>
<tr><td>Razonamiento numérico</td><td>25</td><td>Pensamiento analítico</td><td>25</td></tr>
<tr><td>Razonamiento matemático</td><td>25</td><td>Estructura de la lengua</td><td>25</td></tr>
<tr><td>Español</td><td>25</td><td>Comprensión lectora</td><td>25</td></tr>
</table>

Tabla 16. Dominios del conocimiento evaluados mediante los procesos de admisión de 2013 y 2014.

EP = examen psicométrico, EC = examen de conocimientos.

En la Tabla 16 se consignan los dominios del conocimiento evaluados mediante los tres componentes de los procesos de admisión de 2013 y de 2014, así como el número de preguntas correspondiente a cada dominio.

No se cuenta con la tabla de especificaciones de ninguno de los tres componentes del proceso de admisión. Sin embargo, se cuenta con el instructivo de cada uno de ellos en donde se especifican algunos puntos de importancia.

- **Examen psicométrico (EP).**

En el Instructivo (Anexo 4) se indica el porcentaje que aporta a la calificación final del proceso, los materiales y requisitos necesarios para presentarlo, cómo conocer la fecha y hora del mismo, y qué es lo que evalúa. Como piden un lápiz y una goma, se puede suponer que será escrito, pero no menciona el método de recolección de datos (hoja de respuestas, por ejemplo), no pone ejemplos de las preguntas ni especifica el número o tipo de reactivos. En entrevista con personal del Centro de Salud Universitario, la entidad encargada de su elaboración, no se pudo obtener más información acerca del marco teórico para su desarrollo; sin embargo, sí aclararon que cada año el número de reactivos es diferente, así como los ítems por sí mismos, y que no se ofrece más información al público (ni a la investigadora de esta tesis) para evitar sesgos en los resultados de exámenes futuros.

Se establecieron el promedio y la desviación estándar con base en la puntuación obtenida de las pruebas que evalúan los tres constructos a nivel universitario, es decir, los resultados obtenidos por todos los aspirantes a todas las carreras de toda la universidad. Con estos datos se estableció el rango medio, que abarca el promedio $\pm$ 0.5 desviaciones estándar. Sobre este rango medio se colocan dos rangos superiores, cada uno de una desviación estándar por encima del promedio. Debajo del rango medio, los rangos inferiores serán de una desviación estándar por debajo del promedio. Se agrupa a los estudiantes en rangos con base en su calificación en el EP, lo que determina su probabilidad de ingreso a la Facultad y su probable rendimiento y desempeño académico (Guerrero M, 1979):

La interpretación que se da a los rangos es:

Rango I: *Estudiantes con poca tolerancia a la frustración, por lo general siempre han sido buenos estudiantes con primeros lugares en sus clases.*

Rango II: *Son estudiantes estables las mejores calificaciones a nivel licenciatura y terminan la carrera en el tiempo establecido.*

Rango III+: *En este rango está la mayoría de los estudiantes, sus estudios en algunos casos encuentran dificultades por lo que algunas materias las presentan en exámenes extraordinarios o a título.*

Rango III-: *Son estudiantes que presentan dificultades para terminar sus estudios, por lo regular terminan su carrera en 1 o 2 años más que el previsto en el plan de estudios.*

Rango IV: *Los estudiantes que caen en este rango no terminan sus estudios en el tiempo establecido, reprueban constantemente materias y en algunos casos llegan a agotar sus oportunidades para acreditarlas.*

Rango V: *Los alumnos que están en este rango, si logran ingresar son los que en los primeros meses se dan de baja ya que sus aptitudes están enfocadas en otros objetivos* (Centro de Salud Universitario de la UASLP, s/f).

Este componente se aplicó en las instalaciones del Centro de Salud Universitario por medio de citas entre febrero y junio de 2013 y de 2014. Los aspirantes tuvieron tiempo límite para contestar esta evaluación (30 minutos, cinco a seis minutos por sección), por lo que podría considerarse una prueba de velocidad. Este componente aportó 15% de la calificación final del proceso de admisión (UASLP, 2014a), (Anexo 3).

- **Examen de Conocimientos (EC)**

La escala del EC es de 0 a 160 aciertos, de donde se obtuvo una calificación para este componente. Posteriormente, se ponderó para aportar el 45% de la calificación total del proceso de admisión (UASLP, 2014a), (Anexo 3).

Esta prueba estuvo conformada por 160 POM, con cuatro opciones incorrectas y solo una opción correcta. Los sustentantes contaron con 4 horas para contestarlo y se aplicó en las instalaciones del Centro Cultural Bicentenario de la UASLP el 06de julio de 2013 y el 05 de julio de 2014.

La Guía temática del Examen de Conocimientos de la Facultad de Medicina (UASLP, 2014a), proporcionada por esta entidad académica, menciona el aporte de cada componente a la calificación final: EC 45%, EXANI-II 40%, y EP 15%; estas ponderaciones se aplican exactamente como se indica en la Guía. Este mismo documento, que conforma el instructivo para el EC, establece que esta prueba consta de 160 reactivos y, aunque menciona que no existe penalización por contestar de manera equivocada alguno de ellos, no indica claramente el valor de cada una de las preguntas contestadas de manera correcta.

El Instructivo para aspirantes de primer ingreso (UASLP, 2015b) (Anexo 4) es el documento en donde se explican los pasos necesarios para presentar los componentes del proceso de admisión.

Especifica los requisitos para participar en el proceso de admisión y las ponderaciones correspondientes a cada componente.

Tanto en el Instructivo como en la Guía temática se indican la fecha y hora de aplicación del EC, el mecanismo para recolectar las respuestas (hoja de respuestas analizada por medio de lector óptico), así como el tipo de lápiz y marcas que han de hacerse para indicar la respuesta elegida. El tipo de preguntas se especifica como POM con 5 posibilidades de respuesta; también aporta una muestra de los diferentes tipos de modalidad de preguntas: frases incompletas, analogías y relaciones, construcción o reconstrucción de textos, clasificación y manejo de datos, comprensión de datos, e inferencias lógicas y silogísticas.

En los documentos mencionados también se explica que las condiciones de aplicación del examen fueron estandarizadas: todos los aspirantes se presentaron en el Centro Cultural Bicentenario, a la misma hora y bajo los mismos mecanismos de recolección de datos. No indica si se trataría del mismo examen para todos en una sola versión o en diferentes versiones.

- **EXANI-II**

La calificación para este componente se expresa por medio del índice Ceneval, que es de 700 a 1,300. Este resultado se ponderó para aportar el 40% de la calificación total del proceso de admisión (UASLP, 2014a), (Anexo 3). Los sustentantes contaron con dos horas para contestarlo y se aplicó en las instalaciones del Centro Cultural Bicentenario de la UASLP los mismos días que el EC, por la tarde.

El Centro Nacional de Evaluación para la Educación Superior, A.C. (Ceneval) es el órgano que se encarga de desarrollar el Examen Nacional de Ingreso a la Educación Superior (EXANI-II). Cada año publica la Guía del Examen Nacional de Ingreso a la Educación Superior, cuyo objetivo es "exponer las características y el contenido temático del EXANI-II y ofrecer información sobre la aplicación a quienes han de presentarlo". Sobre la puntuación del examen de selección o admisión (que es el que se utiliza en el proceso que se analiza), se indica que en 2013 se elaboraron "más de 60 versiones del EXANI-II de selección… todas equivalentes en contenido y grado de dificultad. Cada cuadernillo contiene una mezcla distinta de reactivos y opciones de respuesta…" (Centro Nacional de Evaluación para la Educación Superior, 2013a), mientras que en 2014 fueron más de 70 versiones (Centro Nacional de Evaluación para la Educación Superior, 2013b). También menciona que "El EXANI-II incluye únicamente preguntas del tipo opción múltiple, con cuatro opciones de respuesta.” (Centro Nacional de Evaluación para la Educación Superior, 2013a), y que cada respuesta correcta equivale a un punto. Incluye preguntas de prueba (ítems que están a prueba para conocer sus características psicométricas y saber si pueden ser consideradas para futuras versiones) y preguntas de control (para identificar la versión del cuadernillo y poder calificarlo adecuadamente).

I. A. 1. B. PUNTUACIÓN – CONTENIDO – PROCESO DE DESARROLLO DEL INSTRUMENTO.

La Guía temática del Examen de Conocimientos de la Facultad de Medicina (UASLP, 2014a), proporcionada por esta entidad académica, especifica se utilizan POM con cinco posibilidades de respuesta. Además, esta evaluación es elaborada por una comisión interna que verifica que cada ítem se adhiera a los estándares de principios de escritura de ítems de calidad.

Con respecto del Examen Psicométrico, el instructivo no menciona el método de recolección de datos (hoja de respuestas, por ejemplo), no pone ejemplos de las preguntas ni especifica el número o tipo de reactivos. En entrevista con personal del Centro de Salud Universitario, la entidad encargada de su elaboración, no se pudo obtener más información acerca del marco teórico para su desarrollo; y, aunque sí aclararon que cada año el número de reactivos es diferente así como los ítems por sí mismos, no ofrecen más información al público para evitar sesgos en los resultados de exámenes futuros.

Solo en la guía del EXANI-II se menciona que incluye preguntas de prueba (ítems que están a prueba para conocer sus características psicométricas y saber si pueden ser consideradas para futuras versiones) y preguntas de control (para identificar la versión del cuadernillo y poder calificarlo adecuadamente). En los otros dos componentes no se aclara este dato.

I. A. 1. C. PUNTUACIÓN – CONTENIDO – RELEVANCIA DEL DOMINIO.

La fuente de evidencia para este punto es la revisión de los ítems por parte de expertos. El EC es elaborado por una comisión especial que se asegura de la pertinencia de los ítems, mientras que en la guía del EXANI-II se aclara que son los expertos en evaluación del aprendizaje quienes se encargan de verificar la relevancia del dominio. Con respecto del examen psicométrico, no se cuenta con esta evidencia.

I. B. 1. d. Generalizabilidad – Contenido – Representación del dominio.

No se cuenta con evidencia para los EC y EP. En la Guía del Examen Nacional de Ingreso a la Educación Superior de 2013 indica que se elaboraron *"más de 60 versiones del EXANI-II de selección… todas equivalentes en contenido y grado de dificultad. Cada cuadernillo contiene una mezcla distinta de reactivos y opciones de respuesta…"* (Centro Nacional de Evaluación para la Educación Superior, 2013a), mientras que en 2014 fueron más de 70 versiones (Centro Nacional de Evaluación para la Educación Superior, 2013b). También menciona que *"El EXANI-II incluye únicamente preguntas del tipo opción múltiple, con cuatro opciones de respuesta"*, y que cada respuesta correcta equivale a un punto. También menciona que quienes elaboran los ítems son expertos en evaluación educativa (Centro Nacional de Evaluación para la Educación Superior, 2013b).

2. VERIFICACIÓN DE LA INTERPRETACIÓN Y LA DECISIÓN

Kane	**Inferencias e hipótesis**	**A. Puntuación** La puntuación refleja el proceso cognitivo apropiado. El instrumento midió la(s) dimensión(es) planeada(s)		**B. Generalizabilidad** El tamaño de la muestra fue adecuado No existen/sí existen diferencias entre grupos Confiabilidad La puntuación observada se relaciona con la de instrumentos que miden el mismo constructo/no se relaciona con la de instrumentos que miden un constructo diferente.		**C. Extrapolación.** El proceso de admisión predice el primer año de la carrera. La relación con otras variables se puede generalizar a un dominio más amplio del conocimiento.	
Messick	**Fuente de evidencia**	2. Procesos de respuesta	3. Estructura interna	3. Estructura interna		4. Relación con otras variables	
	Elemento		a. Dimensionalidad	b. Invarianza de la medida	c. Confiabilidad	a. Relación prueba-criterio	b. Generalización de la validez
	Evidencia para un Instrumento con POM	Demuestra que se están llevando a cabo los procesos cognitivos esperados al contestar la prueba.	Identificar cuántas dimensiones los ítems.	Análisis del ítem: dificultad y discriminación. Ausencia de sesgo sistemático: demuestra que las puntuaciones serán las mismas cuando se comparen entre grupos con diferentes características como raza, edad o sexo, de acuerdo con la taxonomía de equivalencia.	En cada ocasión que se aplica el instrumento se obtendrán resultados semejantes.	Las puntuaciones predicen el desempeño en un criterio relevante, de manera predictiva o concurrente.	Metaanálisis de los estudios de validación anteriores en condiciones semejantes.
	Lista de cotejo para demostrar las hipótesis	Entrevistas cognitivas, pensar en voz alta. Tiempos de respuesta.	Análisis factorial confirmatorio	Índice de dificultad por ítem y general de la prueba. Índice de discriminación por ítem y general de la prueba. Funcionamiento diferencial del ítem.	Alfa de Cronbach o KR20-21.	Análisis de senderos.	Metaanálisis de los estudios de validación anteriores en condiciones semejantes

Tabla 17. Plan de recolección de evidencias para la etapa 2. Verificación de la interpretación y la decisión.

En esta parte, y con base en la Tabla 17, se ofrecen las fuentes de evidencia para demostrar las hipótesis de la verificación de la interpretación y la decisión

II. A. 2. Puntuación – Procesos de respuesta.

El único componente en cuya guía se establece cuáles son las habilidades y los conocimientos que se evalúan es el EXANI-II. En este caso en particular, se debería comprobar que los ítems estimulan el desarrollo de dichos procesos. Sin embargo, no se realizó ninguna entrevista cognitiva, ni se tomaron los tiempos de respuesta.

II. A. 3. a. Puntuación – Estructura interna –Dimensionalidad.

- 2013

Los datos de 2013 se utilizaron de manera exploratoria. Primero se llevó a cabo AFE para determinar el número de factores por medio de un análisis paralelo. Este análisis indicó que básicamente hay un factor, al que se llamó factor común de inteligencia g, que causa gran parte de la varianza de cada subcomponente y que explica las calificaciones y las correlaciones entre las calificaciones de los 13 subcomponentes. También se identificó la influencia de cuatro subfactores: Verbal, Matemático, Día 1 y Día 2. Este proceso se llevó a cabo de la siguiente manera:

1. Análisis preliminar: adecuación de los datos

Este paso se lleva a cabo a través de dos etapas (Ferrando & Anguiano-Carrasco, 2010):

a. Estadística descriptiva.

Se observó que la muestra fue disminuyendo conforme pasó el tiempo. Probablemente esto fue debido a que el examen psicométrico se aplicó en un día (Día 1), el examen de conocimientos en otro día por la mañana (Día 2) y el EXANI-II por la tarde (Día 2), por lo que quizá algunos sustentantes iniciaron el proceso, pero no lo terminaron. Llama la atención que en los subcomponentes de EP-Lectura y EXANI-II-Español, la mínima fue de cero aciertos (Tabla 18).

Examen	Subcomponente	N	Media	Desviación estándar	Mínimo	Máximo
EP	Razonamiento verbal	1455	15.81	4.33	2	29
	Lectura		12.60	3.72	0	20
	Razonamiento abstracto 1		41.01	8.09	8	58
	Razonamiento abstracto 2		35.18	5.85	5	44
EC	Biología	1378	15.60	4.97	3	32
	Química		13.13	5.00	3	29
	Español		9.15	2.96	2	19
	Inglés		14.95	6.02	3	31
	Físico-matemáticas		13.88	5.01	1	28
EXANI-II	Razonamiento verbal	1373	13.91	3.23	2	20
	Razonamiento numérico		13.60	3.57	2	20
	Matemáticas		14.07	3.64	2	20
	Español		13.65	2.77	0	20

Tabla 18. Estadística descriptiva de los subcomponentes del proceso de admisión de 2013.

EP = examen psicométrico, EC = examen de conocimientos.

b. Evaluar el grado de relación conjunta entre las variables.

Se creó una matriz de correlaciones entre las variables, que es una tabla que en las columnas y las filas presenta las variables (número de variables) y sus correlaciones (Tabla 19). La mayor parte de las correlaciones entre subcomponentes son moderadas (+0.4 a +0.6), las más altas (≥0.7) son Química (EC) con Biología (EC) y con Físico-Matemático (EC), y de Matemáticas (EXANI-II) con Físico-Matemático (EC) y con Razonamiento Numérico (EXANI-II). Las relaciones más débiles (≤0.3) fueron de Lectura y Razonamiento Abstracto 1 (EP) con el resto de los subcomponentes.

		EP				EC					EXANI-II			
		RV	L	RA1	RA2	B	Q	E	I	F-M	RV	RN	M	E
EP	RV		0.38	0.49	0.38	0.5	0.46	0.34	0.46	0.46	0.53	0.45	0.43	0.41
	L			0.29	0.42	0.39	0.37	0.28	0.37	0.36	0.42	0.31	0.33	0.35
	RA1				0.51	0.38	0.38	0.24	0.35	0.39	0.39	0.5	0.42	0.35
	RA2					0.43	0.43	0.28	0.39	0.45	0.41	0.53	0.52	0.33
EC	F-M						0.65	0.47	0.59	0.64	0.56	0.52	0.57	0.45
	E							0.47	0.56	0.68	0.5	0.53	0.61	0.43
	Q								0.5	0.51	0.45	0.39	0.45	0.39
	B									0.6	0.57	0.49	0.54	0.47
	I										0.58	0.63	0.69	0.48
EXANI-II	RV											0.58	0.56	0.57
	RN												0.67	0.49
	M													0.5
	E													

Tabla 19. Matriz de correlaciones entre los subcomponentes del proceso de admisión de 2013.

EP = examen psicométrico, EC = examen de conocimientos, RV = razonamiento verbal, L = lectura, RA1 = razonamiento abstracto 1, RA2 = razonamiento abstracto 2, , FM = físico–matemático, E = español, Q = química, B = biología, I = inglés, RN = razonamiento numérico, M = matemáticas. Todos los valores tienen p <0.0001.

2. Factores.

El método de extracción se usa para obtener los autovalores (*eigenvalues*) de cada factor, necesarios para decidir cuántos factores se deben retener en el siguiente paso del AFE (Larsen & Warne, 2010). Los autovalores son el índice que indica la parte de la varianza total que se debe a un factor subyacente, y su símbolo es λ (American Psychological Association, 2009). Mientras mayor sea un autovalor, mayor es la varianza que causa ese factor; por lo tanto, si un autovalor es de 1.0 o más, quiere decir que tiene más poder de suma que una variable sola (Stellefson et al., 2009).

El método de extracción que se utilizó fue el de componentes principales, a través del que se encontraron 13 autovalores. Los dos primeros fueron de 6.7 (con diferencia de 5.69) y 1.00 (con diferencia de 0.18). El resto fue menor a 1.00.

A continuación, se obtuvo la comunalidad para el primer factor (Tabla 20):

Examen	Subcomponente	Estimado de comunalidad
EP	Razonamiento verbal	0.45
	Lectura	0.30
	Razonamiento abstracto 1	0.36
	Razonamiento abstracto 2	0.42
EC	Biología	0.61
	Química	0.60
	Español	0.39
	Inglés	0.56
	Físico-matemáticas	0.67
EXANI-II	Razonamiento verbal	0.60
	Razonamiento numérico	0.60
	Matemáticas	0.64
	Español	0.45

Tabla 20. Comunalidad del Factor 1.

EP = examen psicométrico, EC = examen de conocimientos.

Estos valores indican la proporción de cada subcomponente que es explicada por el factor común 1. Por ejemplo, el 45% de la varianza del subcomponente de Razonamiento Verbal se explica por el factor común 1. El subcomponente con menor porcentaje de varianza explicada por este factor es el subcomponente de Lectura, con 30%, mientras que el mayor porcentaje es para Físico-matemáticas del EC con 67%.

A través del AFE se busca recrear una matriz de correlaciones más simple, con base en su teorema fundamental, que indica que las correlaciones entre dos variables manifiestas dependen de lo que tienen en común con el (los) factor(es) subyacente(s). Así, el AFE busca estimar el patrón de coeficientes que mejor representa las correlaciones entre las variables observadas al revelar cuáles son las variables latentes (constructos latentes - factores) que pueden estar causando la covarianza de las variables manifiestas (p ej., calificaciones). Es decir, obtener una matriz de correlaciones reproducidas con base en las cargas factoriales, no en los resultados de cada ítem, aunque sí lo más cercano posible a la primera matriz (Costello & Osborne, 2005; Schreiber, 2021; Yela, 1996). Con base en esto, el siguiente paso consistió en elegir el número de factores comunes. Para ello se utilizó un método basado en un análisis de correlación residual, el análisis con χ^2, para saber si con el único factor identificado se puede mantener la hipótesis nula de que todos los residuos (diferencia entre la correlación real y la correlación reproducida) son cero en la población. En este caso se obtuvo χ^2 de 897.64 con $p<0.0001$ para la hipótesis de que un factor es suficiente.

Modelo de análisis factorial confirmatorio

A través del AFC, se probaron diferentes modelos para explorar el factor de inteligencia general "g" por default y dos ejes adicionales: uno fue sobre los días diferentes de aplicación de los componentes del proceso de admisión para explorar si hay diferencia entre los dos días. El otro eje es el factor verbal (lectura, español, inglés, razonamiento verbal) y el factor matemático. Entre los componentes de razonamiento verbal se incluyó al subcomponente de EC- Biología (a posteriori).

Para obtener el modelo final, antes se evaluaron seis modelos diferentes por medio de los índices de bondad de ajuste (Tabla 21). Los valores de referencia son los siguientes: para χ^2 /gl se recomiendan proporciones menores de 2 o 3; para RMR el modelo ajusta mejor mientras más cercano a 0 es el valor. El modelo también ajusta cuando SRMR ≤ 0.08, RMSEA ≤ 0.06 y GFI $\geq$ 0.95 (Abad et al., 2011).

	Modelo						
	1	2	3	4	5	6	Final
χ^2 /gl	62	64	49	51	37	35	37
RMR	0.85	0.98	0.62	0.62	0.30	0.27	0.27
SRMR	0.03	0.04	0.02	0.03	0.01	0.01	0.01
RMSEA	0.07	0.08	0.05	0.07	0.03	0.02	0.02
GFI	0.73	0.91	0.96	0.62	0.98	0.99	0.97

Tabla 21. Índices de bondad de ajuste de los modelos evaluados.

Con base en ello, el modelo final presenta un factor común de inteligencia "g", así como cuatro subfactores cuya influencia también se identificó: Verbal, Matemático, Día 1 y Día 2. El peso de cada subfactor y del factor común sobre cada subcomponente se puede observar en la Figura 18.

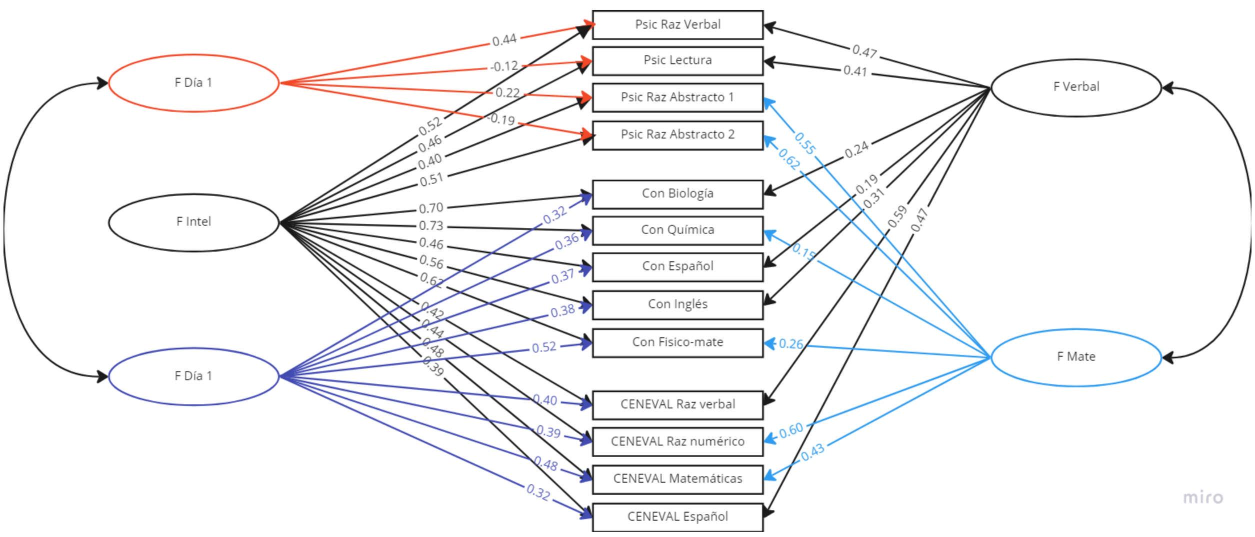


Figura 18. Peso de cada subfactor y del factor común sobre cada subcomponente y para la generación de 2013.

- 2014

En la matriz de correlaciones para los subcomponentes del proceso de admisión se encuentra que Razonamiento Verbal, Razonamiento Abstracto 1 y Razonamiento Abstracto 2 (EP) correlacionan débilmente con todos los subcomponentes. El resto de correlaciones son moderadas, excepto por Físico-Matemático (EC) con Química (EC) y con Pensamiento Matemático (EXANI-II), Química con Biología (EC), y Pensamiento Matemático con Pensamiento Analítico (EXANI-II), que son fuertes (Tabla 22).

		EP				EC					EXANI-II			
	2014	RV	L	RA1	RA2	B	Q	E	I	F-M	PM	PA	EL	CL
EP	RV													
	L	0.28												
	RA1	0.1	0.34											
	RA2	0.2	0.36	0.43										
EC	F-M	0.27	0.44	0.24	0.25									
	E	0.28	0.38	0.18	0.21	0.49								
	Q	0.31	0.46	0.26	0.25	0.74	0.49							
	B	0.3	0.46	0.26	0.24	0.67	0.49	0.74						
	I	0.32	0.49	0.24	0.26	0.06	0.51	0.63	0.61					
EXANI-II	PM	0.32	0.44	0.28	0.03	0.73	0.43	0.67	0.56	0.56				
	PA	0.31	0.42	0.03	0.34	0.65	0.45	0.6	0.53	0.52	0.74			
	EL	0.35	0.49	0.23	0.27	0.61	0.49	0.58	0.55	0.59	0.62	0.58		
	CL	0.36	0.51	0.23	0.27	0.55	0.42	0.5	0.47	0.54	0.58	0.56	0.64	

Tabla 22. Matriz de correlaciones entre los subcomponentes del proceso de admisión de 2014.

EP = examen psicométrico, EC = examen de conocimientos. RV = razonamiento verbal, L = lectura, RA1 = razonamiento abstracto 1, RA2 = razonamiento abstracto 2, FM = físico–matemático, E = español, Q = química, B = biología, I = inglés, PM = pensamiento matemático, PA = pensamiento analítico, EL = estructura de la lengua, CL = comprensión lectora. Todos los valores tienen p <0.0001.

Al utilizar el modelo final que se obtuvo para la generación de 2013, los pesos del factor común y de los subfactores para los subcomponentes del proceso de admisión de 2014 se observan en la figura 19.

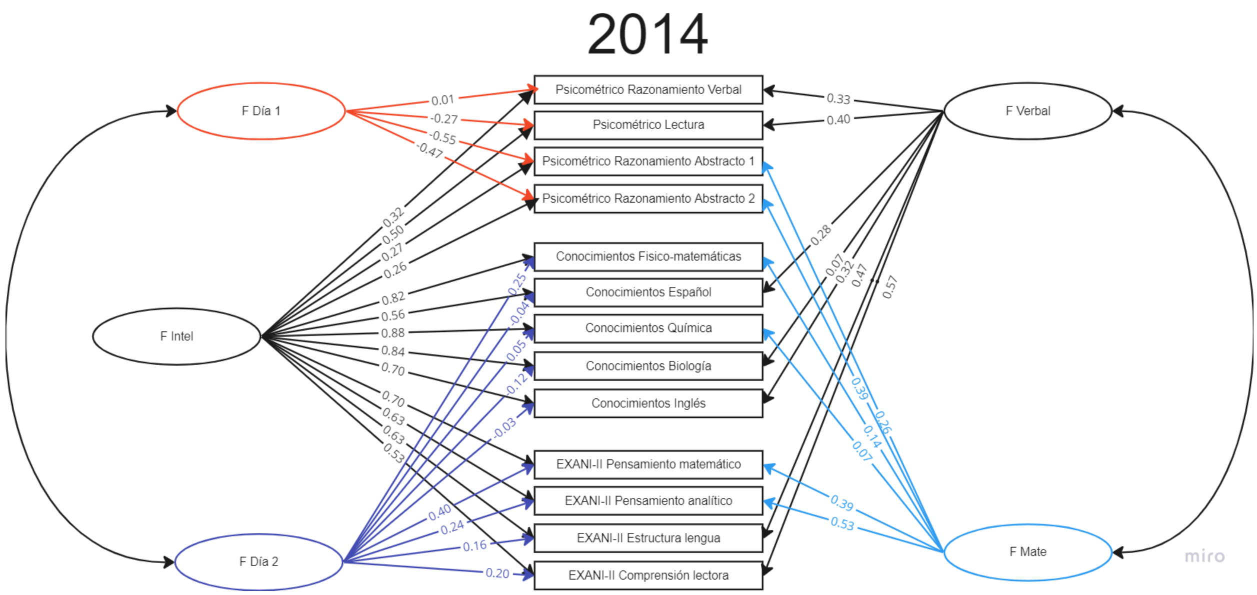


Figura 19. Pesos del factor de inteligencia y los subfactores en los subcomponentes del proceso de admisión de 2014.

Con base en los índices de bondad de ajuste, existe evidencia de equivalencia estructural entre ambos procesos de admisión.

II. B. 3. A. GENERALIZABILIDAD – ESTRUCTURA INTERNA – INVARIANZA DE LA MEDIDA.

La prueba de la diferencia de parámetros entre hombres y mujeres en cada subcomponente y para cada generación mostró diferencias importantes entre hombres y mujeres en el EP, sobre todo con respecto de las cargas del subfactor Día 1 para ambas generaciones (Tablas 23 y 24). Además, en el caso de Biología en 2014, el efecto de inteligencia global sí tiene diferencia entre hombres y mujeres (p=0.014).

		Fg		FV		FM		D1		D2	
		H	M	H	M	H	M	H	M	H	M
EP	Raz verbal	0.58*	0.43	0.51*	0.47*	-	-	-0.40	-0.50*	-	-
	Lectura	0.42*	0.50	0.40*	0.43*	-	-	0.20	0.01	-	-
	Raz Abs1	0.40*	0.32	-	-	0.52*	0.53*	0.05	-0.47*	-	-
	Raz Abs2	0.44*	0.52	-	-	0.60*	0.66*	0.38	0.01	-	-
EC	Biología	0.70*	0.65	0.22*	0.27*	-	-	-	-	0.31*	0.40*
	Química	0.71*	0.72*	-	-	0.17*	0.15*	-	-	0.37*	0.40*
	Español	0.44*	0.45*	0.22*	0.15*	-	-	-	-	0.38*	0.41*
	Inglés	0.52*	0.54*	0.33*	0.30*	-	-	-	-	0.40*	0.44*
	Fisico-mate	0.60*	0.61*	-	-	0.30*	0.22*	-	-	0.54*	0.55*
CENEVAL	Raz verbal	0.40*	0.37*	0.60*	0.57*	-	-	-	-	0.41*	0.48*
	Raz numérico	0.35*	0.34*	-	-	0.63*	0.58*	-	-	0.42*	0.52*
	Mate	0.50*	0.48*	-	-	0.42*	0.44*	-	-	0.50*	0.52*
	Español	0.30*	0.30*	0.50*	0.43*	-	-	-	-	0.37*	0.46*

Tabla 23. Cargas de los factores por sexo para la generación de 2013.

*EP = examen psicométrico, EC = examen de conocimiento, Fg = Factor g, FV = Factor verbal, FM = Factor matemático, D1= Día 1, D2= Día 2. * = p< 0.05*

		Fg		FV		FM		D1		D2	
		H	M	H	M	H	M	H	M	H	M
EP	Raz verbal	0.31*	0.33*	0.36*	0.32*			0.008	0.004		
	Lectura	0.49*	0.52*	0.46*	0.35*			-.22*	-0.30*		
	Raz Abs1	0.24*	0.26*	-		0.15	0.23*	-0.77*	-0.59*		
	Raz Abs2	0.24*	0.24*	-		0.44*	0.36*	-0.37*	-0.51*		
EC	Fisico-mate	0.75*	0.85*	-		0.10	0.16			0.40*	0.07
	Español	0.54*	0.56*	0.28*	0.26*					0.08	-0.15 *
	Química	0.88*	0.86*	-		0.006	0.07			0.24	-0.10
	Biología	0.84*	0.80*	0.08	0.06					0.06	-0.31*
	Inglés	0.69*	0.68*	0.29*	0.32*					0.13	-0.17
CENEVAL	PensaMatema	0.60*	0.75*			0.21*	0.44*			0.61*	0.23*
	PensaAnalit	0.54*	0.64*			0.33*	0.58*			0.53*	0.03
	LenguaEstr	0.62*	0.66*	0.41*	0.48*					0.33*	0.04
	ComprenLect	0.47*	0.57*	0.54*	0.55*					0.36*	0.06

Tabla 24. Cargas de los factores por sexo para la generación de 2014.

*EP = examen psicométrico, EC = examen de conocimiento,. Fg = Factor g, FV = Factor verbal, FM = Factor matemático, D1= Día 1, D2= Día 2. * = p< 0.05.*

II. B. 3. b Generalizabilidad – Estructura interna – Confiabilidad.

El proceso de admisión constituye un test compuesto, pues combina las puntuaciones obtenidas de varias pruebas (EP, EC y EXANI-II). Se podría analizar la confiabilidad de todo el proceso como un test compuesto per se si cada prueba tuviera atributos similares, pero cada prueba evalúa constructos diferentes y tiene número de ítems diferente (Martínez Arias et al., 2014).

Por parte de la comisión de admisión se reportó, para el EC, índice de confiabilidad de Kuder Richardson de 0.78 en 2014. No hay información para 2013, y no se cuenta con los datos necesarios para hacer el análisis por nuestra parte.

II. C. 4. c. Extrapolación – Relación con otras variables – Relación prueba-criterio.

Para la generación 2013, el resultado del proceso de admisión explica 4% de la varianza de las calificaciones de todos los alumnos de primer año ($p=0.01$), y 7% de la varianza de solo los aprobados ($n=126$, $p=0.0024$). Las calificaciones del proceso de admisión de la generación 2014 explican 24% de la varianza de las calificaciones de todos los alumnos de primer año ($p<0.001$), y 31% de la varianza en el caso de solo los aprobados ($n=115$, $p<0.001$).

Los resultados del análisis de regresión logística para 2013 y 2014 por componente y subcomponente se registran en la Tabla 25.

Para conocer la relación del resultado del proceso de admisión con el criterio (calificaciones de primer año) se realizó análisis de senderos, que mostró que el puntaje del primer año en 2013 sí predice fuertemente el puntaje en el segundo año; sin embargo, incluir los puntajes del examen de admisión no añade fuerza de predicción significativa a la basada únicamente en el puntaje del primer año. El valor de la predicción es de 0.95, lo que indica que el alumno que tenga un punto más en el promedio de primero, tendrá 0.95 más en segundo año. Para 2014 los componentes del examen de admisión no explican la varianza en el promedio de segundo año, si se controla por el puntaje del primer año.

		2013		2014	
		Parámetro	*p*	Parámetro	*p*
	EP	-0.011	0.04	0.010	0.02
	EXANI-II	0.009	0.53	0.053	<0.001
	EC	0.025	<0.001	0.030	<0.001
EP	Razonamiento verbal	-0.014	0.38	-0.003	0.72
	Lectura	-0.018	0.35	-0.018	0.40
	Razonamiento Abstracto 1	-0.007	0.48	0.018	0.01
	Razonamiento Abstracto 2	-0.022	0.17	0.026	0.16
EC	Biología	0.027	0.13	0.040	0.002
	Química	0.037	0.02	0.019	0.38
	Español	0.001	0.93	0.012	0.60
	Inglés	0.017	0.21	0.025	0.06
	Físico-matemático	0.028	0.14	0.046	0.01
EXANI-II	Razonamiento verbal	-0.032	0.40		
	Razonamiento lógico-matemático	0.035	0.30		
	Matemáticas	0.018	0.65		
	Español	0.0004	0.98		
	Pensamiento matemático			0.052	0.05
	Pensamiento analítico			0.047	0.76
	Estructura de la lengua			0.010	0.005
	Comprensión lectora			0.084	0.002

Tabla 25. Resultados del análisis de regresión logística para 2013 y 2014, por componente y subcomponente.

EP = examen psicométrico, EC = examen de conocimientos.

3. UTILIDAD DE LAS ACCIONES

Kane	Inferencias e hipótesis	**D. Implicaciones** La interpretación de los resultados es adecuada. Los alumnos tienen éxito académico en la carrera. Es razonable y justificable el tipo de escala de la prueba		
Messick	**Fuente de evidencia**	5. Consecuencias		
	Evidencia para un Instrumento con POM	Los usuarios están de acuerdo con la interpretación de los resultados.	Establecer la trayectoria académica de los alumnos y de la generación en conjunto.	Utilidad de las escalas con referencia a la norma.
	Lista de cotejo para demostrar las hipótesis	Encuesta a los usuarios: alumnos, profesores y administrativos de la facultad.	Calificaciones de cada alumno para cada asignatura de 1º a 5º año.	Bibliografía en donde se demuestre la utilidad del uso de la escala con referencia a la norma en este tipo de evaluaciones.

Tabla 26. Plan de recolección de evidencias para la etapa 3. Utilidad de las acciones.

Para conocer la utilidad de las acciones, se aplicó una encuesta del 23 de septiembre al 09 de octubre de 2021 (anexo E) a los alumnos aceptados en los años 2013 y 2014 (Miranda López et al., 2018). Sus objetivos fueron:

a) Conocer el nivel de aceptabilidad con el que los sustentantes perciben el cumplimiento de los instructivos para cada componente del proceso de admisión.
b) Medir la satisfacción conseguida en las diferentes etapas del proceso de admisión.

Los encuestados de ambas generaciones tenían entre 16 y 21 años cuanto presentaron el examen de admisión, y entre 25 y 29 años edad cuando contestaron la encuesta. Del total de 78 personas que respondieron, fueron 40 mujeres y 38 hombres.

El instrumento se divide en tres momentos: antes de llevar a cabo el proceso de admisión, durante la aplicación de los exámenes de admisión, y la entrega de resultados. Con respecto del primer momento, 93% de los encuestados está satisfecho o muy satisfecho con el tiempo de emisión y vigencia de la convocatoria, 76.9% con los medios de difusión de la convocatoria, 67.9% con la utilidad del Instructivo para Aspirantes de Nuevo Ingreso para conocer el proceso en general, 67.9% con la claridad con la que la UASLP resolvió las dudas sobre la convocatoria, 92.3% con la atención de la autoridad educativa para realizar el registro, recepción y revisión de la documentación, 75.6% con la utilidad del Instructivo para Aspirantes de Nuevo Ingreso para conocer cómo se va a evaluar, 48.7% con la utilidad de Guía temática del Examen de Conocimientos de la Facultad de Medicina para conocer qué se va a evaluar, 74.3% con el tiempo con el que contó para tener acceso a la bibliografía y guía de estudios, 52.5% con la relación de la guía de estudios y la bibliografía, con el contenido de los exámenes (Figuras 20 a 31).

En el siguiente momento, el de aplicación de los exámenes de admisión, 80.7% de los encuestados está satisfecho o muy satisfecho con los aspectos que se evalúan en los exámenes, 78.2% con la precisión de la redacción de los planteamientos en las preguntas, 89.7% con la cantidad total de preguntas del examen, 91% con la extensión de las preguntas del examen, 80.7% con la contextualización de las preguntas del examen, 97.4% con la localización de la sede, 96.1% con la accesibilidad de la sede, 65.3% con la comodidad del mobiliario, 85.8% con la iluminación y la temperatura de las aulas, 96.1% con la precisión de las indicaciones brindadas por el aplicador durante el examen, 91% con la atención del aplicador ante las dudas de los sustentantes, y 94.8% con el trato brindado a los sustentantes por el aplicador (Figuras 32 a 43).

En el momento posterior a la evaluación, entre 61.5 y 67.9% sabían el número de espacios educativos disponibles, cómo se califican los exámenes y cómo se conforman las listas de aceptados antes de presentar las evaluaciones (Figura 44).

No se aplicó una encuesta semejante a los académicos o administrativos a cargo, ya que varios de ellos ya no se encontraban laborando en la institución en el momento de este estudio de validación.

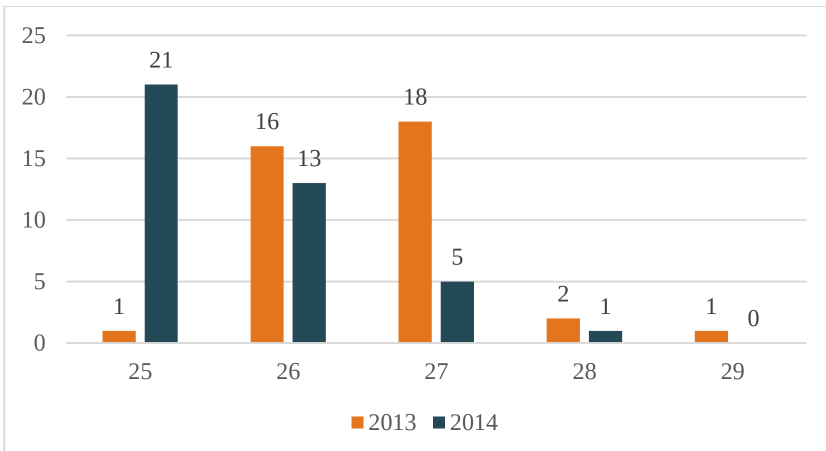


Figura 20. Edad actual de los encuestados.

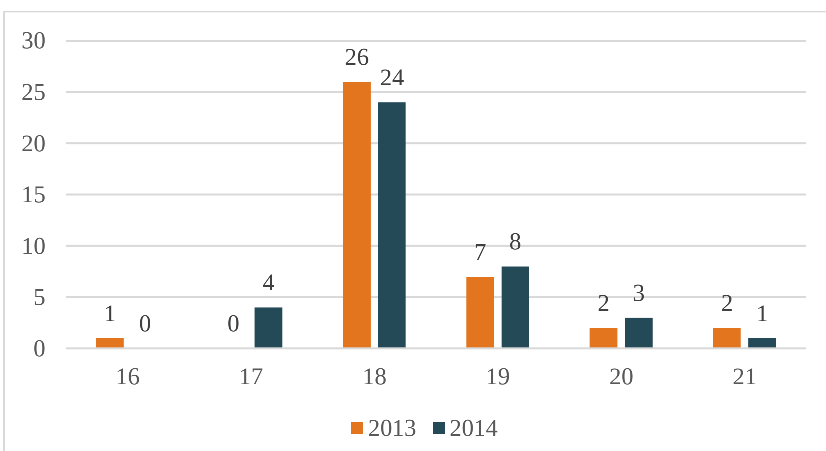


Figura 21. Edad de los encuestados cuando presentaron el examen de admisión.

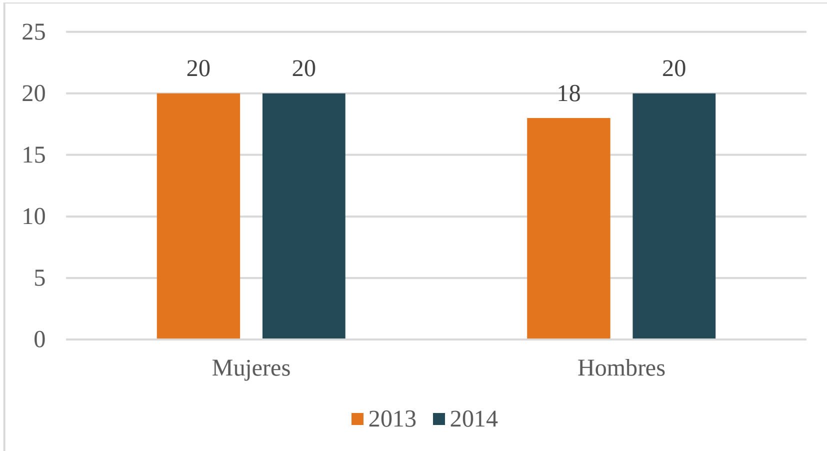


Figura 22. Sexo de los encuestados.

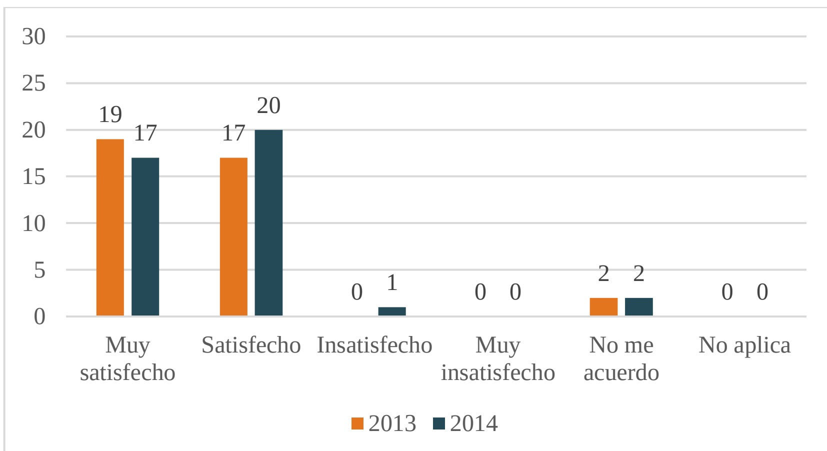


Figura 23. El tiempo de emisión y vigencia de la convocatoria para el proceso de admisión.

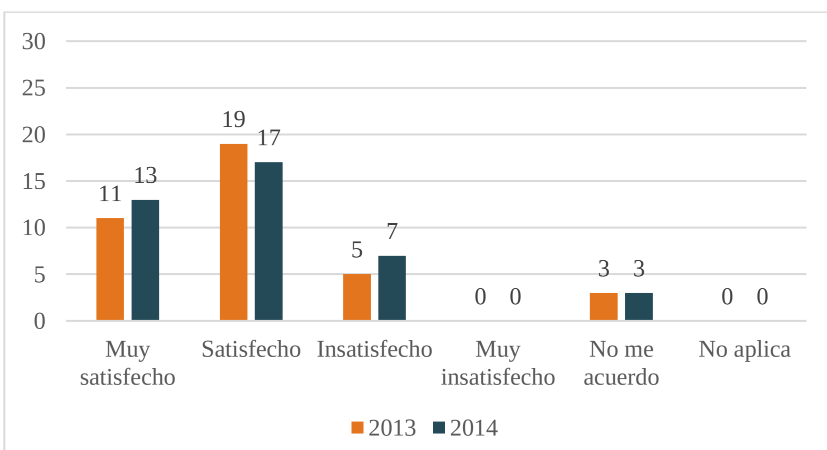


Figura 24. Los medios de difusión de la convocatoria del proceso de admisión.

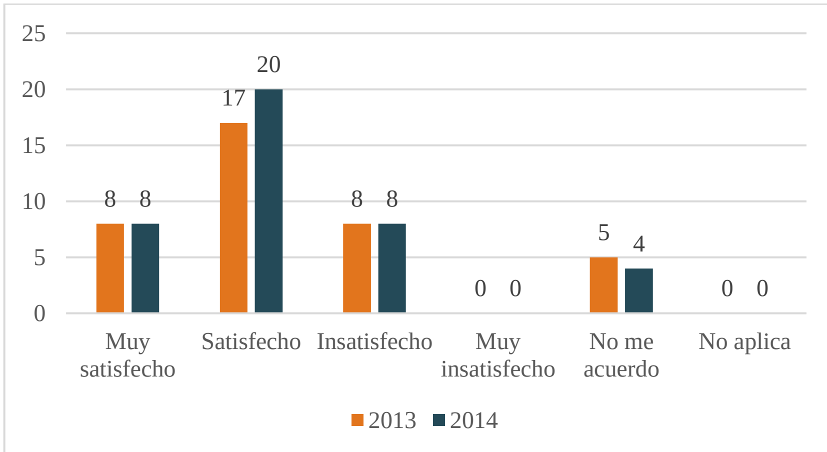


Figura 25. La utilidad del Instructivo para Aspirantes de Nuevo Ingreso para conocer el proceso en general.

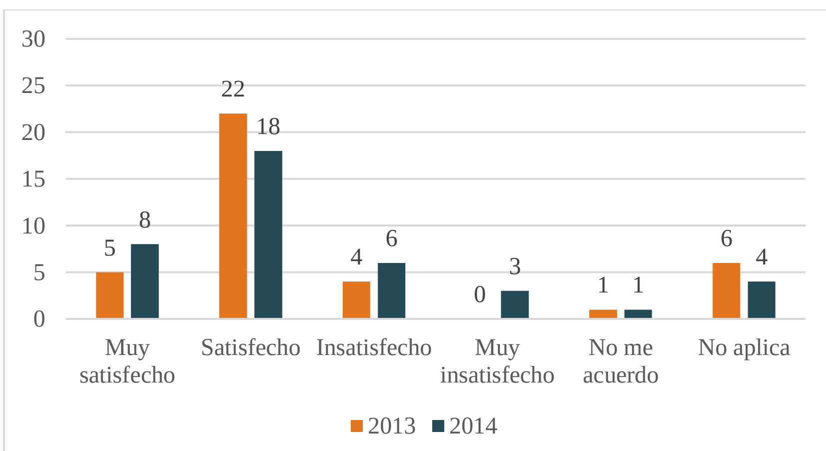


Figura 26. La claridad con la que la UASLP resolvió las dudas sobre la convocatoria.

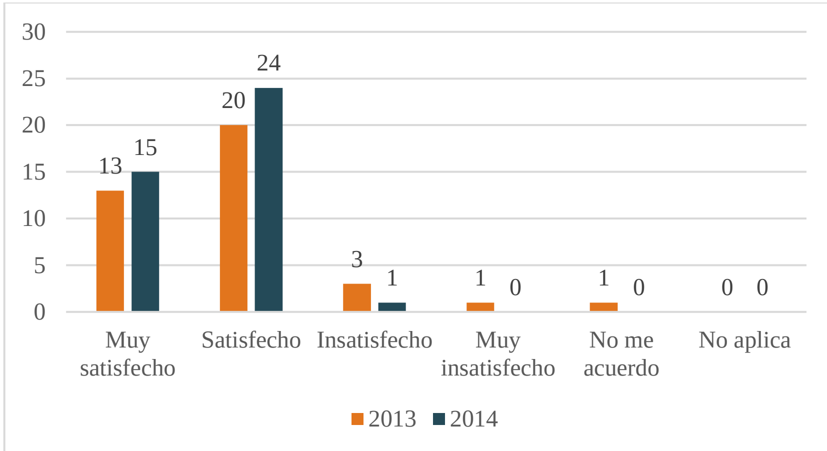


Figura 27. La atención de la autoridad educativa para realizar el registro, recepción y revisión de la documentación.

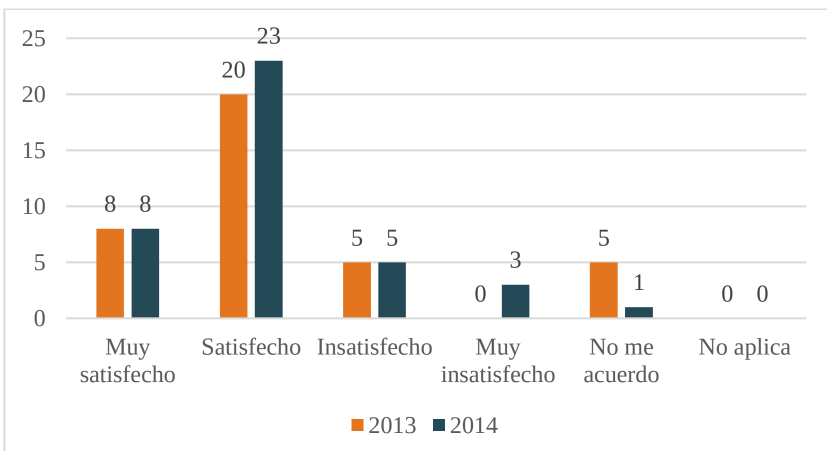


Figura 28. La utilidad del Instructivo para Aspirantes de Nuevo Ingreso para conocer cómo se va a evaluar.

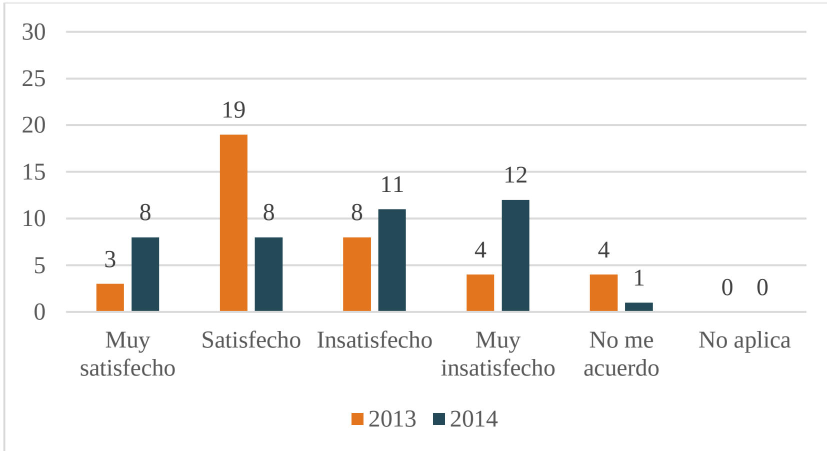


Figura 29. La utilidad de la Guía temática del Examen de Conocimientos de la Facultad de Medicina para conocer qué se va a evaluar.

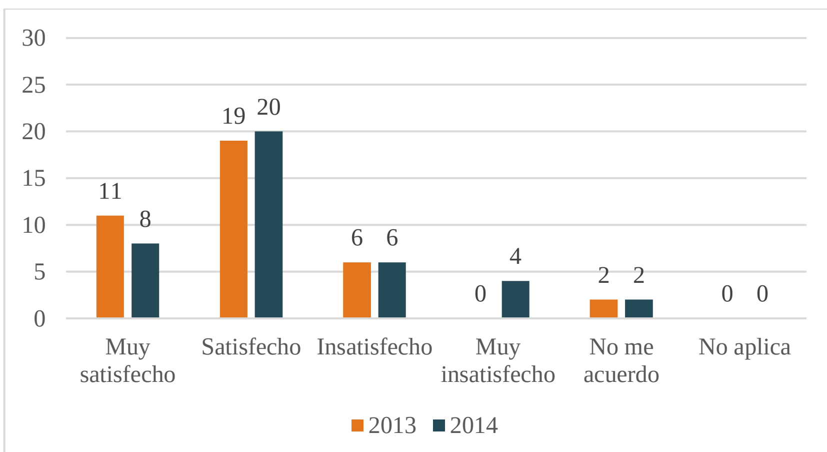


Figura 30. El tiempo con el que contó para tener acceso a la bibliografía y guía de estudios.

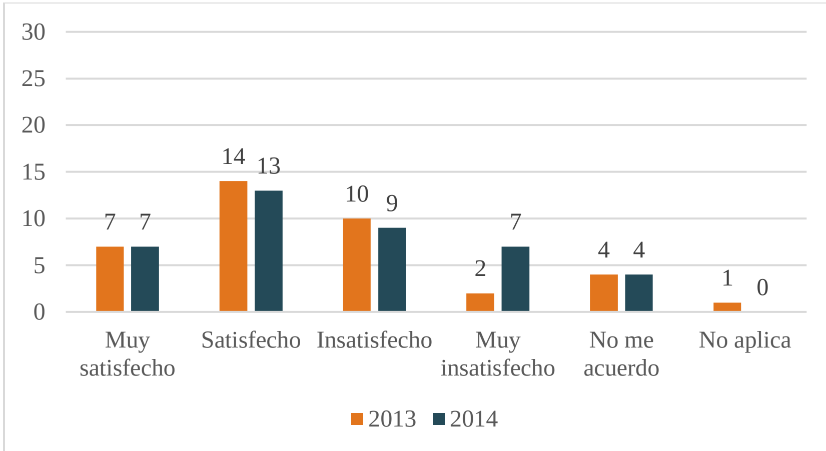


Figura 31. La relación de la guía de estudios y la bibliografía, con el contenido de los exámenes.

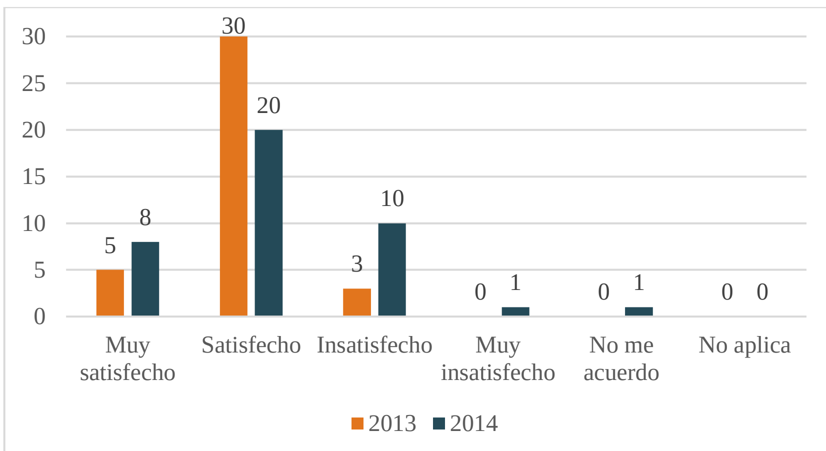


Figura 32. Los aspectos que se evalúan en los exámenes.

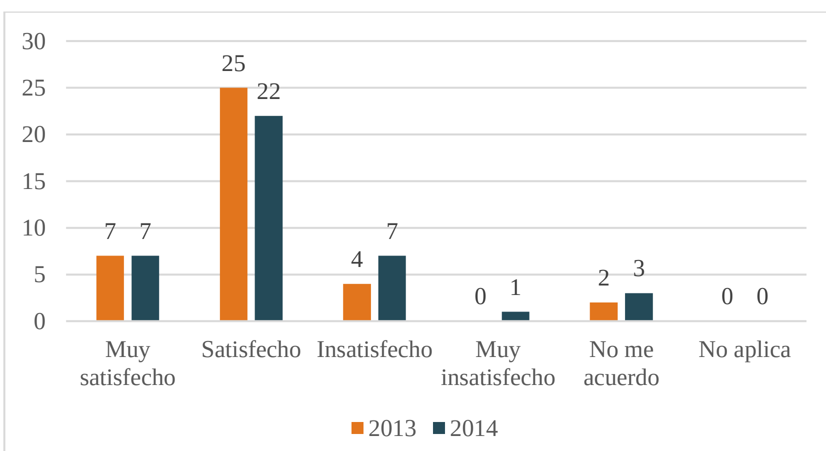


Figura 33. La precisión de la redacción de los planteamientos en las preguntas.

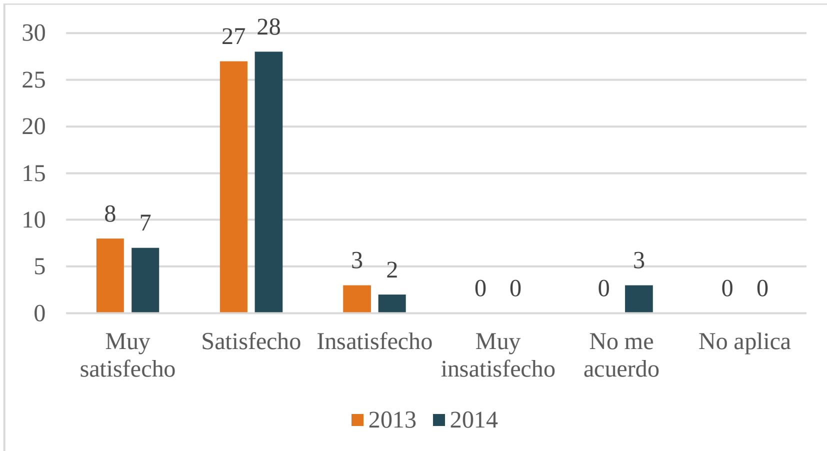


Figura 34. La cantidad total de preguntas del examen.

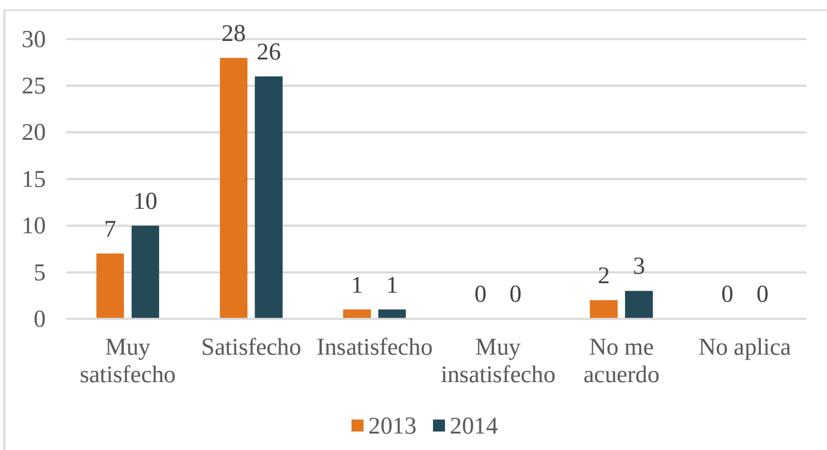


Figura 35. La extensión de las preguntas del examen.

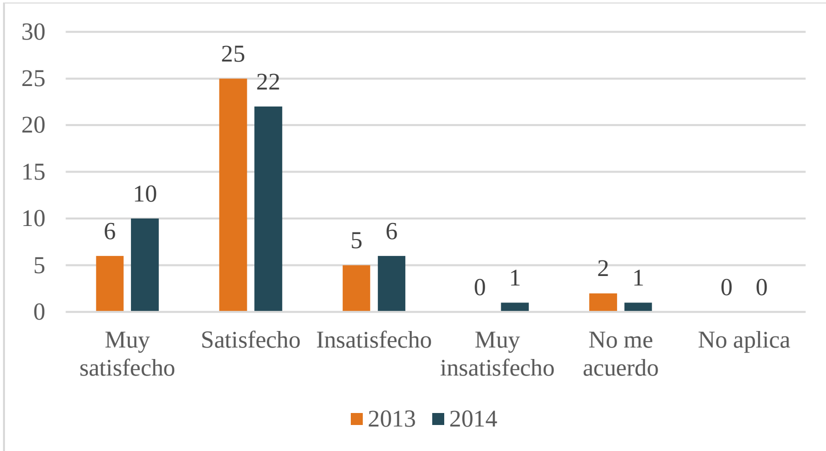


Figura 36. La contextualización de las preguntas del examen.

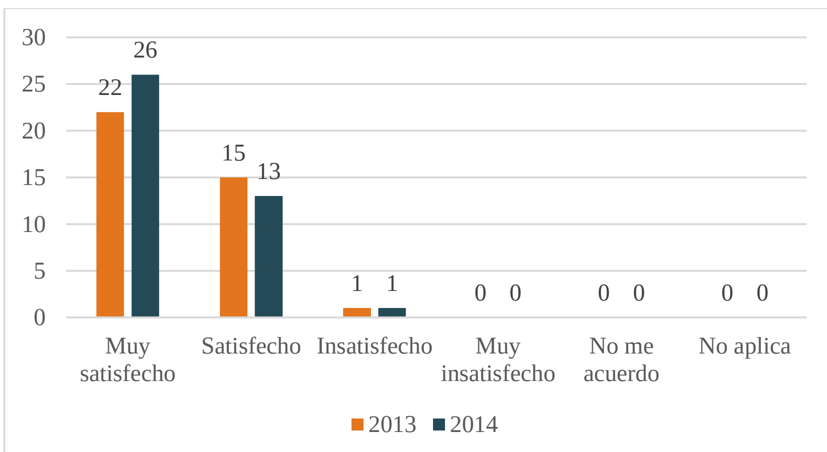


Figura 37. La localización de la sede.

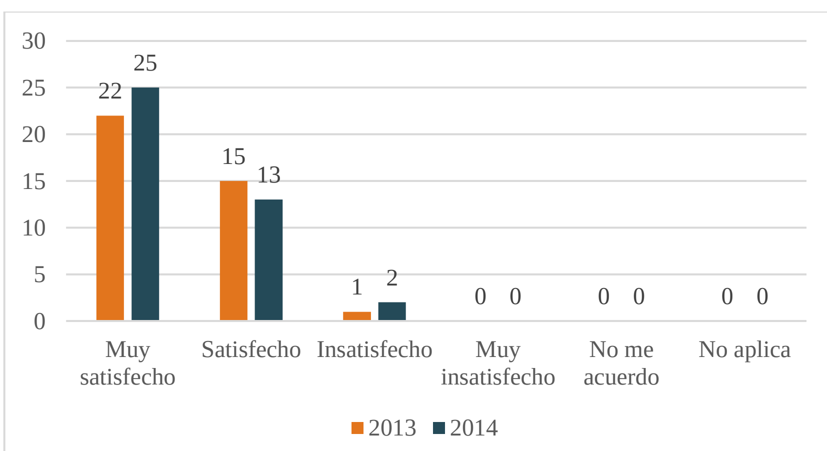


Figura 38. La accesibilidad de la sede.

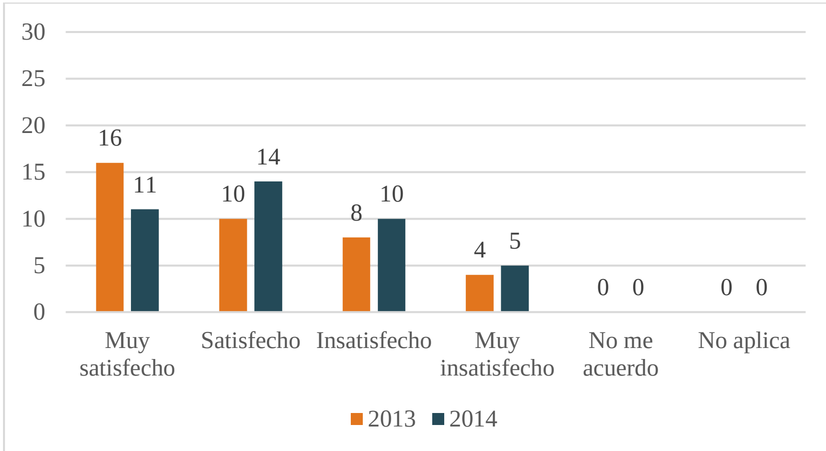


Figura 39. La comodidad del mobiliario de las aulas.

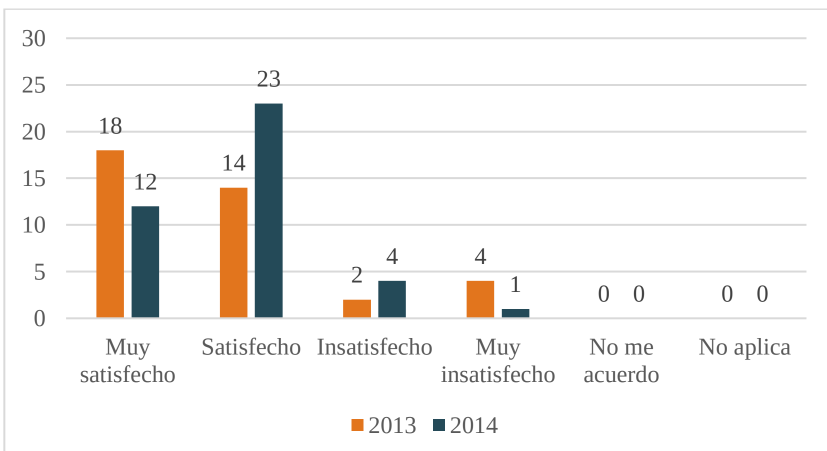


Figura 40. La iluminación y la temperatura de las aulas

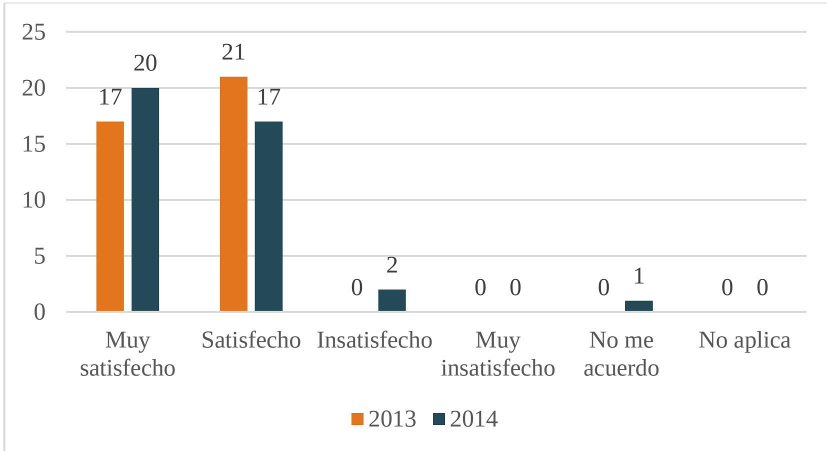


Figura 41. La precisión de las indicaciones brindadas por el aplicador durante el examen.

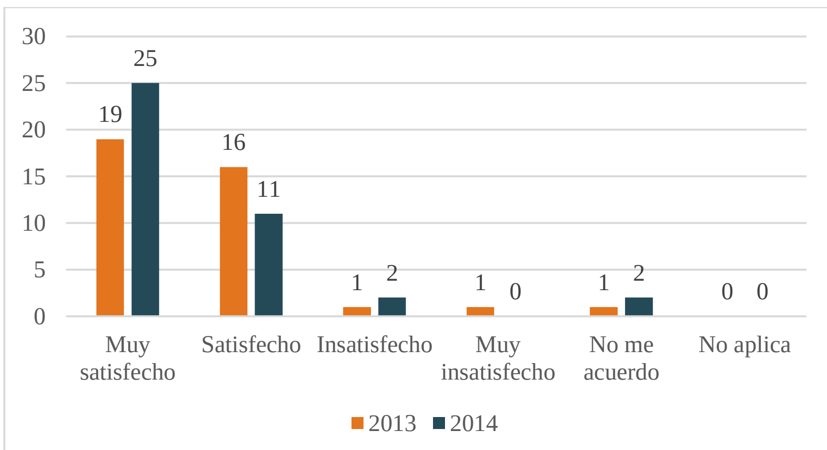


Figura 42. La atención del aplicador ante las dudas de los sustentantes.

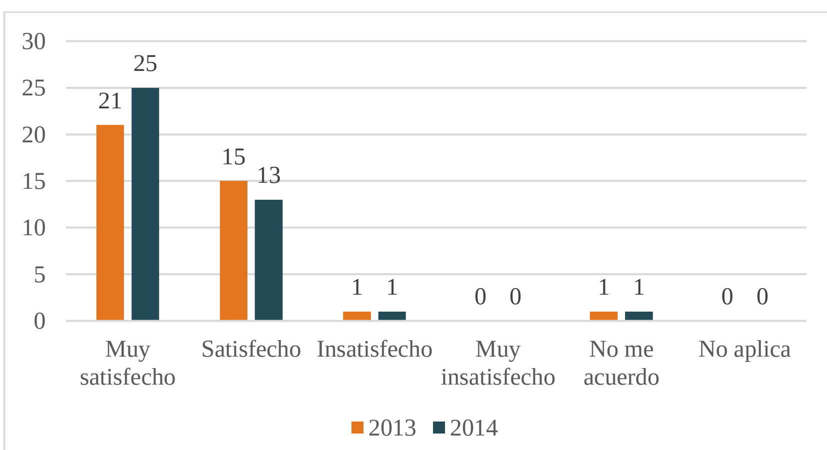


Figura 43. El trato brindado a los sustentantes por el aplicador.

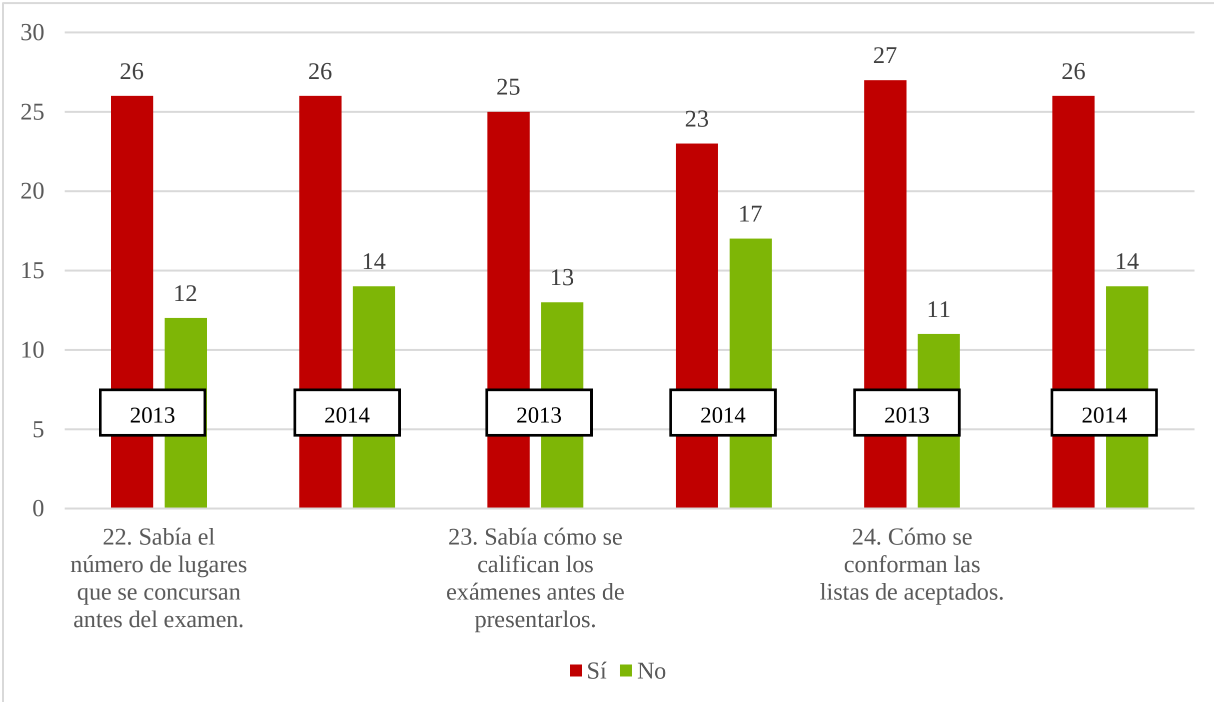


Figura 44. Momento posterior a la evaluación (tercera fase)

- Los alumnos tienen éxito académico en la carrera

Para saber si los alumnos tienen éxito académico en la carrera, como fuente de evidencia de la utilidad de las acciones se hizo un seguimiento de la trayectoria académica por generación. De los 145 alumnos que ingresaron en 2013, 89 terminaron hasta el internado; de ellos 77 aprobaron el EGEL de Medicina General (primera parte del examen profesional), y 72 aprobaron el ECOE (segunda parte del examen profesional) y se titularon a tiempo (Figura 45).

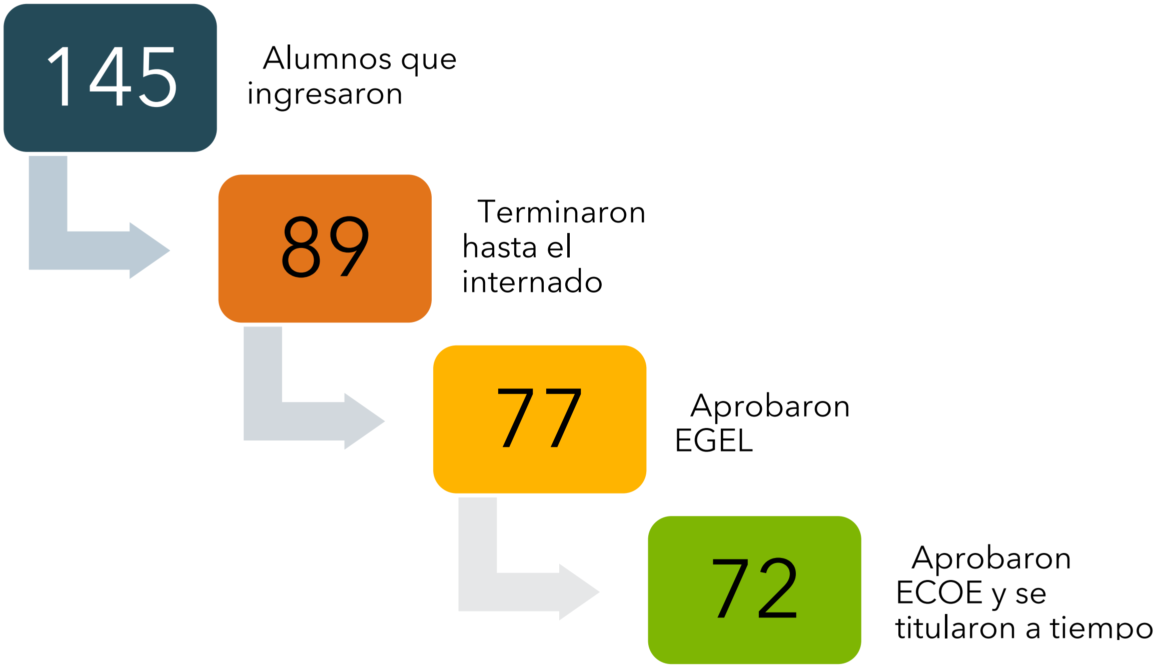


Figura 45. Trayectoria académica de la generación 2013.

De los 145 alumnos que ingresaron en 2014, 80 terminaron hasta el internado, 77 aprobaron el EGEL y 72 aprobaron el ECOE y se titularon a tiempo (Figura 46).

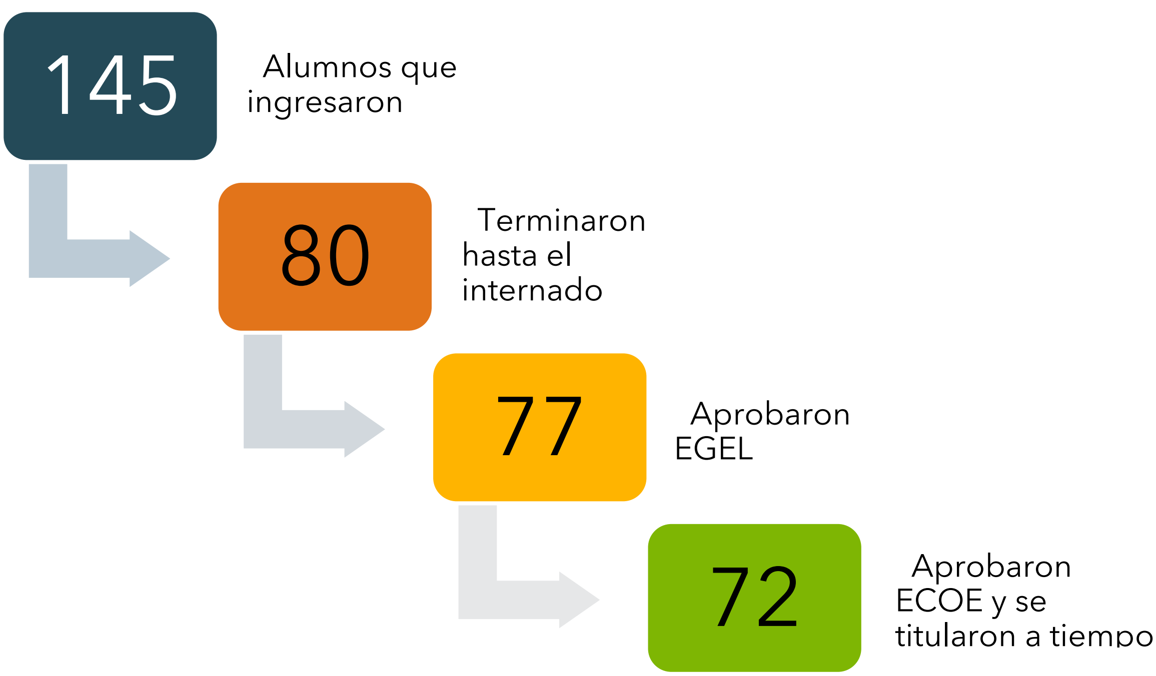


Figura 46. Trayectoria académica de la generación 2014.

- Es razonable y justificable el tipo de escala de la prueba

Con base en el objetivo del proceso de admisión, el uso de la escala con referencia a norma puede justificarse, ya que es una evaluación estandarizada a gran escala con propósitos de selección de estudiantes y que discrimina entre los aspirantes de alto y bajo desempeño. Además, el reporte numérico de las calificaciones permite conocer el grado de desempeño en los diferentes subcomponentes para cada alumno (Lok et al., 2016; K. D. Royal & Guskey, 2015; S. G. Sireci & Greiff, 2019; Tavakol & Dennick, 2017).

7. DISCUSIÓN Y CONCLUSIONES

7.1 Discusión

La validez como un concepto abstracto es difícil de entender, pues carece de referentes perceptibles; para hacer frente a este reto, las teorías de representación múltiple ofrecen un camino. Estas teorías explican que los conceptos abstractos se incorporan por medio de la percepción, la acción y las emociones, así como por la recreación de la experiencia lingüística. De esta manera, la internalización de los conceptos abstractos implica la activación de diferentes áreas del sistema nervioso central, además de que está influenciada por las experiencias sociales previas de cada persona (Borghi et al., 2017).

En el caso de pruebas de altas consecuencias, como los exámenes de admisión, se recomienda realizar un análisis integral con un marco de referencia sólido para aportar evidencias de validez (American Educational Research Association et al., 2018); sin embargo, es común que las publicaciones sobre validez de los exámenes de admisión se enfoquen en solo uno de los aspectos a evaluar, como la validez predictiva, y no suelen especificar el marco de referencia utilizado (Bala et al., 2022; Barajas-Ochoa et al., 2019; Crawford et al., 2021; Cunningham et al., 2019; Hafferty et al., 2020; Panda et al., 2021; Paton et al., 2022; Salehi et al., 2021; Simpson et al., 2014; Violato et al., 2020). En México estos análisis son escasos y estudian un solo tipo de prueba (Bárquez Antillán & Vivian Mascareño, 2009; García Domínguez, 2016; Martínez Villarreal, 2013; Morales Ibarra et al., 2009; Villegas Vizcaíno, 2017). Es así que es necesario indagar sobre cuál marco de referencia es apropiado utilizar para reunir y evaluar las fuentes de evidencia de validez del proceso de admisión a una escuela de medicina mexicana.

Acerca de la propuesta del modelo integrador

La validez es un tema complejo, y no existe consenso acerca de muchos de sus aspectos, por ejemplo, su definición (Borsboom et al., 2004; Camargo et al., 2018; Cizek, 2012). Tampoco lo hay con respecto a la interpretación que se le da, pues puede verse como una característica de la prueba, como un argumento basado en una cadena de evidencias y/o como un imperativo social (St-Onge et al., 2017). A pesar de lo anterior, las principales asociaciones internacionales en evaluación han alcanzado un acuerdo que incluye cómo demostrar la validez de los usos y las interpretaciones de las puntuaciones de la prueba (American Educational Research Association et al., 2018; S. G. Sireci, 2016a). Esta falta de consenso ocasiona diferencias en el entendimiento y aplicación de los marcos de referencia de validez (M. Young et al., 2018), por lo tanto, el proceso de validación generalmente implica un gran trabajo e incluso un reto mayor para los educadores en profesiones de la salud que no cuentan con amplias habilidades en evaluación educativa. En este trabajo hemos desarrollado una propuesta para ordenar la información de manera que tenga sentido para los educadores en ciencias de la salud, ya sean legos al tema de validez o no, y que permita entender el resultado del análisis de manera congruente con la necesidad de validar los usos de las puntuaciones y el significado de las puntuaciones, de acuerdo con Cizek (Cizek, 2020) y Sireci(S. G. Sireci, 2016b).

Al unir los marcos de referencia con el objetivo de enriquecerlos y potenciarlos, se amplifica la visión y el entendimiento de los interesados en validez. Además, debido a que es difícil simplificar demasiado la validez debido a su complejidad, la contribución de este trabajo no solamente es intentar construir conexiones entre los marcos de referencia más importantes a través de los elementos de las fuentes de evidencia de validez, sino también incorporar un modelo que considera las diferentes etapas del desarrollo de instrumentos, lo que hace más fácil entender en qué punto se encuentra el evaluador en cada paso del desarrollo de las pruebas en términos de la validez.

Existen varios esfuerzos en el mismo tenor, por ejemplo, el de Kinnear y colaboradores (Kinnear et al., 2021), quienes proponen un mapa de validez en el que ellos muestran cómo unir los marcos de referencia de Messick y de Kane. Particularmente en este caso, consideramos que las inferencias de Kane se alinean de manera un poco diferente con las fuentes de evidencia de Messick, considerando lo que ya se ha descrito en esta tesis con respecto de los diferentes elementos de cada fuente de evidencia y cómo coinciden con las inferencias de Kane y sus propias fuentes de evidencia.

Por otro lado, (Cook & Hatala, 2016) sugirieron un acercamiento práctico a la validación que implica diez puntos a considerar. Estas recomendaciones se pueden llevar a cabo durante cada una de las tres etapas de Russell: Etapa I: definir el constructo y la interpretación propuestos, y hacer explícita la decisión determinada. Etapa II: desarrollar el argumento. Para todas las etapas: priorizar la evidencia de validez necesaria. También se considera importante separar el AUI, como lo indica Kane (Kane, 2013a), y que la especificación de las inferencias y las suposiciones debe ser parte de cada etapa del segundo paso que se propone aquí.

Durante la validación es frecuente preguntarse cuánta evidencia es necesaria para cada instrumento y cómo estructurar la argumentación. No existen guías claras de qué hacer cuando llega el momento de juzgar la evidencia y determinar el grado de validez de los usos e interpretaciones de las puntuaciones. En este último paso debe reconocerse que el grado de validez es tan fuerte como el eslabón más débil, así que se sugiere reconocer las fuentes más importantes para el caso y asegurarse de que se encuentren en el mejor estado posible (Crooks et al., 1996; Gasmalla & Tahir, 2020). También se aconseja asignar la tarea de recolección de datos a personas específicas. Por ejemplo, los diseñadores de la prueba deben proveer evidencia acerca de las evaluaciones de altas consecuencias porque su impacto puede ser grande sobre los usuarios (Crooks et al., 1996; Gasmalla & Tahir, 2020; Kane, 2013b).

En muchas ocasiones, las consecuencias de la prueba se consideran hasta el final en la validación; al parecer algunos autores consideran que no es crítico llevar a cabo un análisis completo (Moss, 1998; Nichols & Williams, 2009), ya que existen pocos reportes que consideren esta fuente de evidencia (Lyons-Thomas et al., 2014). Sin embargo, Haertel (2013) hace énfasis en que las consecuencias involuntarias deben estudiarse, e ilustra esto con un ejemplo de pruebas de admisión a las universidades: si su uso específico es llevar a una mejor toma de decisiones, entonces la validación debería incluir, pero no limitarse a, el estudio de validez predictiva. Además, es importante reconocer que exponer la evidencia no es lo mismo que argumentar, así que se debe utilizar una estructura de argumentación de validez explícita y dirigirla hacia una audiencia específica, como sugieren Kinnear et al. (2022).

Como se mencionó antes, la validez de los procesos de admisión es un área muy poco estudiada en México, lo que conforma un vacío del conocimiento. Otro de los objetivos de este manuscrito, fue implementar el modelo que proponemos para llevar a cabo un estudio de validación en donde se identificó información de diversas fuentes de evidencia de validez en el proceso de admisión de una escuela de medicina de una universidad pública, la FMUASLP, encontrando evidencias de soporte, pero también áreas de oportunidad.

Para valorar el grado de validez de los usos e interpretaciones de las puntuaciones se analizaron los datos de los procesos de admisión de 2013 y 2014 para ingresar a la FM UASLP. En esta sección discutiremos la evidencia a favor y en contra de la validez de la interpretación de los resultados del proceso de admisión, así como la evidencia faltante.

La evidencia fuerte a favor es la basada en la estructura interna (correlaciones entre componentes y el AFC), la relación entre el EC y los resultados en el primer año, y la satisfacción de los usuarios. Solo se encontraron correlaciones fuertes (+0.7 a +0.9, -0.7 a -0.9) (Dancey & Reidy, 2017; Rios & Wells, 2014) entre subcomponentes de EP y de EXANI-II (matemáticas) en 2014. Existe relación moderada (+0.4 a +0.6, -0.4 a -0.6) entre los subcomponentes de ciencias naturales del EC y entre los subcomponentes de matemáticas del EXANI-II, y débil (+0.1 a +0.3) o moderada entre los subcomponentes que evalúan constructos semejantes dentro del EP. Sobre todo en 2014, el factor "g" tuvo una fuerte influencia en la varianza de las calificaciones de los subcomponentes orientados a las ciencias naturales y el de inglés. La influencia de este factor sobre el subcomponente de Inglés puede deberse a que en este se evalúan el dominio de la lengua y también se hacen preguntas sobre otros subcomponentes en inglés. Además, la influencia de este factor sobre los subcomponentes del EP fue moderada (2013) o baja (2014), y sobre los del EXANI-II moderada (2013) o alta (2014). En contraste, en un estudio en que se compararon dos versiones del MCAT, encontraron correlaciones moderadas entre componentes que miden constructos semejantes y además identificaron cuatro factores que nombraron conocimiento biomédico, conceptos de ciencia básica, razonamiento cognitivo y desempeño general (Violato et al., 2020).

La diferencia entre coeficientes entre hombres y mujeres con respecto de Biología en 2014 es interesante, ya que en otras publicaciones lo que se ha reportado es con respecto de ciencias naturales en general o de Física. Además, en estos análisis también han observado que entre los aspirantes hay más mujeres que hombres, pero son admitidos más hombres que mujeres (lo que se observa en la generación de 2014). En varios estudios se ha demostrado que en realidad puede tratarse de VIC, tal vez producida por ansiedad u otras variables que, aunque impactan en el desempeño de las mujeres en evaluaciones de alto impacto, no implican un mal desempeño académico posterior (Arenas & Calsamiglia, 2022; Ganjoo et al., 2020; Habersack et al., 2015; Haladyna & Downing, 2004; Leiner et al., 2018; Luschin-Ebengreuth et al., 2016).

No se hizo un análisis de diferencia de coeficientes entre los grupos de estudiantes de la UASLP provenientes de una institución privada o de una institución pública. Sin embargo, se observó que los promedios de las escuelas privadas y públicas son casi iguales, aunque ingresan más alumnos de escuelas privadas. Es probable que quienes cursaron la preparatoria en una institución privada acuden más a cursos de preparación para el examen de admisión, lo que les da ventajas porque no solo refuerzan conocimientos, sino que también pueden aprender estrategias y trucos para contestar las preguntas, por lo que tal vez no estén mejor preparados o posean más habilidades que quienes no toman los cursos mencionados. Los *Estándares para pruebas educativas y psicológicas* (American Educational Research Association et al., 2018) indican que se debe documentar si es posible que el desempeño durante la prueba que se está evaluando cambie con base en práctica y entrenamiento. En este caso no es posible seguir este estándar, ya que no se cuenta con información sobre los sustentantes que llevaron algún curso de preparación para sustentar el proceso de admisión, cuál o cuáles cursos, cuántas veces, etc.

En cuanto a la relación con otras variables, el dominio de relación de la prueba con el criterio se examinó a través de regresión logística. El componente de EC produce gran varianza en las calificaciones del primer año (por cada punto en el EC, el promedio sube 0.025 en primer año en 2013 y 0.030 en 2014), mientras que el EP tiene influencia negativa y el EXANI-II, nula. También se observó que cada vez es menor la predicción de cada año sobre el siguiente, lo que es de esperar porque las calificaciones de la población se van homogeneizando. En la literatura se encuentran resultados variables que incluyen mayor impacto del razonamiento verbal sobre el desempeño académico durante y hasta el final de la carrera (Violato & Donnon, 2005), o mayor importancia de los conocimientos científicos (semejante al EC) sobre el desempeño académico durante la carrera (Davies et al., 2022).

La satisfacción de los estudiantes con respecto del proceso de admisión fue buena en general; el resultado bajo en el nivel de satisfacción con la guía para el examen es motivo de reflexión sobre su elaboración y la información que contiene. El acceso a materiales de ayuda puede ser motivo de inequidad entre los aspirantes, favoreciendo a quienes tienen mejor situación económica (Lambe et al., 2012). Desde el punto de vista de aceptabilidad del EC como prueba de admisión, hay reportes con resultados mixtos; es más frecuente que les parezca irrelevante el examen de razonamiento no verbal (Patterson, Prescott-Clements, et al., 2016).

La evidencia débil a favor está presente, pero no es comprobable. Por ejemplo, la verificación de la calidad de los ítems debe hacerse para probar que no hay VIC por medios como poner ítems a prueba (American Educational Research Association et al., 2018). Solo el EXANI-II utiliza ítems a prueba, pero el EC, que sí causa varianza en las calificaciones del primer año, no aporta información al respecto. Otra fuente débil es la confiabilidad, ya que solo contamos con el reporte de la UASLP en 2014, pero no tenemos datos de 2013. Esto es una muestra de la importancia de contar con archivos históricos en las instituciones educativas de nuestro país, con el objetivo de proveer de información organizada que permita realizar consultas e investigación.

En cuanto a la representación del dominio para extrapolación, aunque en Estados Unidos se demostró la importancia de evaluar los campos del conocimiento de ciencias naturales y la capacidad de razonamiento verbal y matemático para el proceso de admisión a medicina (AAMC-HHMI., 2009), no existe este tipo de análisis en nuestra institución o en nuestro país. Otros estudios sugieren llevar a cabo pruebas de juicio de situaciones para aportar escenarios de la vida real (Kelly et al., 2018; UKCAT Consortium, 2023).

Entre la evidencia en contra se observó que el EXANI-II tiene un efecto nulo en las calificaciones del primer año y el EP lo tiene negativo en 2013; esto podría significar que altas puntuaciones en el EP predicen menor desempeño en el primer año. Los diferentes estudios de validez predictiva con respecto de pruebas de admisión han mostrado que la relación entre el proceso de admisión y las calificaciones se va perdiendo después del primer año, pues existen muchas variables que tienen impacto sobre las calificaciones de la carrera (Ganjoo et al., 2020; Hernández-Mata et al., 2005). Sin embargo, en la literatura existe evidencia del poder predictivo del examen de razonamiento verbal y de pruebas semejantes al EC (Bala et al., 2022), así como de la prueba de razonamiento verbal por sí sola (Violato & Donnon, 2005).

Si se cumplieran los objetivos del EP y del EXANI-II, se estarían admitiendo estudiantes aptos para estudiar y con gran potencial para aprender cosas nuevas, lo que los llevaría a tener un buen desempeño académico. Nuestros resultados no demuestran esta predictibilidad, por lo que sugerimos reevaluar el desarrollo del EP, así como proporcionar más información a los sustentantes acerca de su estructura y de la modalidad de las preguntas. En cuanto al EXANI-II, su diseño ha cambiado por completo recientemente, por lo que sugerimos llevar a cabo un análisis como el que presentamos pero con la nueva versión.

Faltó evidencia documental que demostrara la definición del dominio, su revisión por expertos, demostración de que la combinación de constructos evaluados es la apropiada, datos para confiabilidad, índices de confiabilidad, discriminación y FDI, y entrevistas cognitivas.

El conocimiento, como constructo evaluado, es un dominio muy específico ya que el sustentante puede tener amplios conocimientos sobre química y no tenerlos sobre física, por ejemplo. Por este motivo es importante contar con un número adecuado de preguntas para obtener la confiabilidad necesaria, sobre todo en un examen como este, que es de altas consecuencias (Schuwirth & Van Der Vleuten, 2011).

El proceso de admisión es con propósitos de selección, por lo que se puede argumentar la justificación de usar la referencia a norma; sin embargo, sería deseable demostrar que las tres pruebas discriminan entre aspirantes de alto y bajo desempeño. Este es un aspecto complejo, pues en realidad son los espacios físicos y número de profesores, entre otros factores, los que determinan la cantidad de estudiantes que se pueden recibir (Lok et al., 2016; K. D. Royal & Guskey, 2015; S. G. Sireci & Greiff, 2019; Tavakol & Dennick, 2017).

Es preferible que las entrevistas cognitivas se lleven a cabo en cuanto ha terminado la prueba. En el caso que analizamos, no se realizó ninguna entrevista de este tipo. Los objetivos de las pruebas, en donde se establecen los usos e interpretaciones de los resultados, especifican los procesos que se evalúan, por lo que se requiere de evidencia de que, en efecto, se están demostrando estos procesos (Embretson, 1998).

Como se comentó antes, (Cook & Hatala, 2016) identificaron varios errores que pueden cometerse durante el proceso de validación. A continuación, se hará referencia a cada uno de estos errores, en qué consisten, y si se presentaron o no y en qué medida durante el análisis que se consigna en esta tesis.

1. Crear una nueva evaluación cada vez.- Es mejor realizar un análisis de validez en un instrumento ya existente que empezar de cero con la creación de un instrumento nuevo. Esto permite a los lectores comparar sus resultados con los nuestros, ya que se trata de una evaluación con la que ya están familiarizados. Sin embargo, es poco común encontrar estudios de validez completos en la literatura nacional e internacional. Incluso en cuanto a estudios de validez, son pocos los que se han encontrado que hagan referencia a los procesos en México hasta el momento y solo se enfocan en el valor predictivo de los exámenes de admisión(García Domínguez, 2016; Martínez Villarreal, 2013; Morales Ibarra et al., 2009; Villegas Vizcaíno, 2017). Por otro lado, también es complejo hacer comparaciones, ya que en cada universidad el proceso implica exámenes diferentes.

2. No utilizar un marco de referencia.- Los marcos de referencia proveen de rigor y de sistematización para obtener las fuentes de evidencia de validez necesarias; también dejan ver cuáles son las fuentes faltantes y cómo se puede subsanar esta ausencia. Como se mencionó antes, el propósito de este trabajo es proponer un modelo que permite verlos de manera más integrada para su mejor comprensión.

3. Hacer de las comparaciones entre los resultados de grupos de expertos e inexpertos la base del argumento de validez.- Los resultados que se obtienen de estas comparaciones no suelen ser útiles, ya que pueden deberse a múltiples causas. Es más interesante si se obtienen diferencias entre grupos en los que no se esperaban o, al contrario, que no existan diferencias entre grupos en los que sí deberían existir. En este caso se encontró diferencia entre grupos (hombres y mujeres) con respecto del subcomponente de Biología en 2014, lo que no debería existir.

4. Concentrarse en la evidencia de validez más accesible en lugar de en la más importante.- Esto podría hacer que hubiera más evidencia sobre una fuente y muy poca sobre otra. Se deben identificar las ausencias y subsanarlas con base en el marco de referencia, de esta manera se obtendrá un argumento de validez claro y completo. Este punto incide nuevamente en la necesidad de hacer más accesible la validez a los educadores en ciencias de la salud y la importancia de saber cuáles son las fuentes indispensables para probarla. Este punto conforma una limitante en este estudio, ya que las fuentes de evidencia estuvieron incompletas.

5. Enfocarse en el instrumento en vez de hacerlo en los usos e interpretaciones de las puntuaciones.- El análisis de validez depende del contexto en que se aplica el instrumento, ya que

esto definirá sus usos e interpretaciones. Para evitar cometer este error se sugiere, al iniciar el desarrollo de las pruebas, establecer los usos e interpretaciones y tenerlas en cuenta en cada paso del proceso de validación. Se ha seguido este consejo al iniciar el estudio de validación con el AUI en donde se anotan claramente los usos e interpretaciones de los subcomponentes del proceso de admisión.

6. Falla para sintetizar o criticar la evidencia de validez.- Después de llevar a cabo los pasos del proceso de validación no debemos olvidar elaborar las conclusiones sobre el mismo, reconociendo la información obtenida y el resultado del análisis. Esto se logra mediante la elaboración de un argumento de validez completo y sistemático, que permita identificar las ausencias y cómo fueron subsanadas. Es difícil este paso debido a que puede ser una crítica directa a los desarrolladores y aplicadores de la prueba; sin embargo, es fundamental para entender dónde se puede mejorar. En este caso, se puede concluir que el grado de validez del proceso en general es aceptable, pero hacen falta más datos para ofrecer un juicio bien informado.

7. Ignorar las mejores prácticas para el desarrollo de la evaluación.- El instrumento de evaluación que provee de las puntuaciones cuya interpretación está sujeta a validación debe desarrollarse considerando que debe evaluar el constructo para el que fue hecho, con un número de ítems suficiente y representativo. También debe ser un instrumento que favorezca la objetividad y disminuya la subjetividad. Este punto se relaciona con el que habla acerca de tener claros los usos y las interpretaciones de la prueba; sin estas bases, no solo el instrumento no funcionará para evaluar lo que se pretende, sino que podría tener consecuencias negativas sobre los usuarios (sustentantes, por ejemplo). Este es un punto que corresponde a los desarrolladores de las evaluaciones. Con respecto del EXANI-II, CENEVAL ha presentado informes en los que detalla cómo desarrolla esta prueba (Centro Nacional de Evaluación para la Educación Superior, 2013b, 2013a, 2016); sin embargo, no hay datos completos con respecto de los otros componentes.

8. Omitir detalles acerca del instrumento.- Al reportar que la interpretación de los resultados de un instrumento es válida, se debe compartir este instrumento con el mayor detalle posible. Esto permitirá la comparación con los de los lectores, ser citado y que el instrumento sea más utilizado. Este punto es al que se hace referencia con el reporte de los resultados de la validación de los procesos de admisión que analizamos en este manuscrito y que se consigna en el Anexo 10.

9. Permitir que la disponibilidad del instrumento de evaluación guíe la evaluación.- El instrumento debe alinearse con el constructo a evaluar, no al revés. Por ello, la validación debe ser un proceso planeado desde el principio del desarrollo de la prueba. Como se ha comentado, la falta de fuentes de evidencia completas para este proceso de admisión (principalmente para el EC y el EP) podrían hacer pensar que los desarrolladores no tenían planeado llevar a cabo la validación, o, más bien, que no se comparten los datos para que investigadores independientes la realicen.

10. Etiquetar a un instrumento como validado.- La validez es una propiedad de las puntuaciones, las interpretaciones y las decisiones, y se expresa en grados, no como variable dicotómica. Además, la validez es un proceso, no un punto final. Este punto permite mejorar las evaluaciones futuras, por lo que aumenta su importancia: si no se mide y no se evalúa, no se sabe dónde y cómo mejorar.

El marco de referencia de Kane está fundamentado en la formulación de un argumento de validez, el que explica basándose en el modelo de argumentación de Toulmin (Johnson, 2011). Las publicaciones dentro del campo de educación en profesiones de la salud (EPS), no suelen explicar de manera evidente cuál es la estructura del argumento que se utiliza, con la consecuencia de presentar las evidencias como si fueran el argumento, sin discutir por qué o cómo aportan al proceso de validación. Esto hace más complicado determinar el grado de validez de la interpretación de los resultados. Para mejorar esta situación Kinnear et al. (2022) sugieren manifestar claramente cuál es tipo de argumento que se está usando al momento de analizar las evidencias, ya sea retórica nueva o lógica informal. Además, proponen dirigir este argumento a una audiencia específica, tanto para que el proceso de validación se comprenda mejor, como para que tenga más utilidad. Entre estas audiencias se consideran a los editores y lectores de revistas especializadas, así como los usuarios de las pruebas como los sustentantes y las instituciones. La claridad del tipo de argumento, su desarrollo, y su resultado, cobran mayor relevancia mientras más altas sean las consecuencias de la prueba cuyos resultados se están analizando.

Como se comentó antes, son escasas las publicaciones con respecto de análisis de validez de pruebas de altas consecuencias en educación en ciencias de la salud. Es probable que una de las causas sea su gran complejidad. Por ello se enumeran a continuación de algunas recomendaciones que pueden ser de interés para quienes deseen realizarlo en su institución (Cizek, 2020).

1. **Desarrollar un plan.**

Antes de empezar la elaboración de la evaluación se debe contemplar la evidencia que se necesita para la validación: cuál es y en qué momento se debe recolectar. El propósito de esto es llevar a cabo un análisis prospectivo y completo; así, se podrán corregir errores antes de la implementación final de la prueba. Establecer con claridad el objetivo de la evaluación y el constructo que se desea medir durante el primer paso es sumamente importante, pues conforman la guía para reconocer cuál es la evidencia más importante. Tener claros los momentos de recolección también permiten identificar al responsable de recogerla: el desarrollador de la prueba actúa durante la etapa I. Validez instrumental, quienes aplican la prueba actúan en la etapa II. Verificación de la interpretación y la decisión, y quienes usan la prueba actúen a la etapa III. Utilidad de las acciones.

2. **Integrar un equipo multidisciplinario.**

Contar con un equipo multidisciplinario es de gran ayuda para realizar este proceso de manera rápida y eficiente. Para saber a cuáles expertos debemos reclutar, se debe revisar el plan de recolección de evidencias y los momentos en que estas se recogen. Así, se puede identificar la necesidad de expertos como especialistas en psicometría, expertos en evaluación, matemáticos y expertos en estadística, coordinadores de elaboradores de reactivos, personas que apliquen las encuestas de satisfacción, etc. También es importante involucrar a los directivos de la institución en donde se implementa la evaluación, pues además de que forman parte de los usuarios de los resultados de la prueba, son quienes manejan los recursos que permiten mejorar su calidad. Otro importante miembro del equipo será un especialista que haya llevado a cabo antes este tipo de análisis: su experiencia es invaluable sobre todo para quienes son noveles en el tema, así que no se debe dudar en pedir ayuda en otras instancias dentro de la misma institución o a otras instituciones con más práctica.

3. **Tener infraestructura.**

Sin duda también es indispensable contar con espacios físicos y material para llevar a cabo los análisis (equipo de cómputo y software de análisis estadístico) y la discusión de los resultados. En ocasiones los espacios serán los sitios con los que ya se cuenta para otras actividades cotidianas en la institución, aunque sí debe delimitarse un tiempo particular para revisar el avance del análisis y los resultados que va arrojando.

4. **Tener la mente abierta.**

A veces el resultado del análisis demostrará que los procesos de evaluación que se están llevando a cabo tienen muchas áreas de oportunidad. Es importante tener la mente abierta y ser humildes para reconocerlas y tomar medidas para mejorarlas o, de plano, volver a empezar de cero.

5. **Saber cuándo se debe repetir la validación.**

Se recomienda volver a realizar la validación completa cuando cambian el objetivo o el constructo que se desea medir. Sin embargo, debe recordarse que, ya que el proceso de validación es continuo, es conveniente volver a llevar a cabo ciertos análisis cada vez que se aplica la prueba, sobre todo cuando ocurren eventos como la reciente pandemia por COVID-19, durante la que hubo cambios en la manera de aplicar las pruebas de altas consecuencias a nivel mundial. También mientras haya cambios en las teorías con respecto de los constructos que se pretenden medir, en la población a la que se apliquen la prueba, en los usos de la prueba o en las inferencias que se desean probar.

6. **Utilizar un marco de referencia.**

Por último, y lo más importante, utilizar un marco de referencia consolidado, ya sea el de Messick, el de Kane, la propuesta que en esta tesis se expone o cualquier otro, de manera exhaustiva. Es necesario fundamentar en un marco de referencia los análisis de validez y, con base en ello, deben realizarse integralmente para obtener una visión completa de la prueba y de sus resultados.

7.2 Limitaciones del estudio

Los procesos de admisión que se evaluaron ocurrieron hace 10 y 11 años, por lo que parte de la información ya no se encuentra disponible o podría presentar sesgo de tiempo (por ejemplo, las respuestas de la encuesta de satisfacción). La muestra para el análisis de regresión es pequeña, y va disminuyendo conforme se avanza en la carrera. Al mismo tiempo, el efecto de la restricción de rango (solo analizamos los resultados de los alumnos admitidos, los 145 sustentantes con mayor calificación del proceso) implica que los resultados del análisis de regresión, por ejemplo, puedan parecer insignificantes, aunque para el tamaño de la muestra sea adecuado.

También hubo limitaciones de transparencia y de disponibilidad de información que dificultan el proceso de validación por personas externas como, por ejemplo, las características de los ítems, es decir, por sustentante, qué contestó en cada ítem. Además, el estudio se realizó en una sola institución pública, por lo que los resultados podrían no ser extrapolables a otras universidades nacionales públicas o privadas, o extranjeras.

Otra limitación de este estudio es que, aunque la validación debe realizarse preferiblemente de manera prospectiva, llevamos a cabo el análisis retrospectivamente, cuando los instrumentos ya se habían aplicado y sus puntuaciones habían sido interpretadas y utilizadas. La utilidad de estos resultados puede ser limitada para esos procesos en particular, pero permite reflexionar sobre los procesos futuros y en dónde incidir para mejorar la selección de los estudiantes.

7.3 Conclusiones y líneas de investigación a futuro

Un marco de referencia de validez claro permite entender mejor los usos y las interpretaciones de los resultados de las evaluaciones de alto impacto, como los exámenes de admisión para las carreras en ciencias de la salud. También es importante planearlo con antelación, para recabar las fuentes de evidencia de manera ordenada y así tener información completa y relevante para un estudio de validación veraz y que ofrezca una visión amplia del examen que se está evaluando.

Una ventaja de utilizar las tres etapas de Russell, es que favorece la identificación clara de las ausencias y de las fuentes más débiles en cada momento de la validación. El valor de esta propuesta es continuar con la discusión acerca de validez y de validación, y de contribuir a su entendimiento y aplicación en la educación en ciencias de la salud.

A través de los conocimientos generados de esta tesis se abren nuevas líneas de investigación, tales como aplicar el método en un proceso de admisión desde el momento que inicia el desarrollo de las pruebas que lo conformen, a otras evaluaciones de altas consecuencias como el ECOE como examen de titulación, y en otros procesos de admisión dentro de las profesiones de la salud.

8. REFERENCIAS

AAMC-HHMI. (2009). *Association of American Medical Colleges and the Howard Hughes Medical Institute, Report of Scientific Foundations for Future Physicians Committee.* 46. http://www.hhmi.org/grants/pdf/08-209_AAMC-HHMI_report.pdf

Abad, F. J., Olea, J., Ponsoda, V., & García, C. (2011). *Medición en ciencias sociales y de la salud*. Editorial Síntesis.

Al Alwan, I., Al Kushi, M., Tamim, H., Magzoub, M., & Elzubeir, M. (2013). Health sciences and medical college preadmission criteria and prediction of in-course academic performance: A longitudinal cohort study. *Advances in Health Sciences Education*, *18*(3), 427–438. https://doi.org/10.1007/s10459-012-9380-1

Alcocer Andalón, A. (1976). *Historia de la escuela de medicina de la Universidad Autónoma de San Luis Potosí, S.L.P. (México) 1877-1977*. Aconcagua Ediciones y Publicaciones, S. A.

Alfonso, V. C., Flanagan, D. P., & Radwan, S. (2005). The impact of the Cattell-Horn-Carroll theory on test development and interpretation of cognitive and academic abilities. *Contemporary intellectual assessment: Theories, tests and issues*, 185–202.

American Educational Research Association, American Psychological Association, & National Council on Measurement in Education. (2018). *Estándares para pruebas educativas y psicológicas*. American Educational Research Association.

American Psychological Association. (2009). Ee. En *APA college dictionary of psychology* (1st ed., pp. 120–144). American Psychological Association.

Arenas, A., & Calsamiglia, C. (2022). Gender Differences in High-Stakes Performance and College Admission Policies. *IZA Discussion Paper*, *15550*, 1–47. www.iza.org

Association of American Medical Colleges. (2020). *What's on the MCAT ® Exam?* www.aamc.org/mcat

Association of American Medical Colleges. (2023). *2023 Official Guide to Medical School Admissions*.

Association of American Medical Colleges, & Howard Hughes Medical Institute. (2009). *Scientific Foundations for Future Physicians Report of the AAMC-HHMI Committee*. https://www.aamc.org/system/files?file=2020-02/scientificfoundationsforfuturephysicians.pdf

Bala, L., Pedder, S., Sam, A. H., & Brown, C. (2022). Assessing the predictive validity of the UCAT—A systematic review and narrative synthesis. *Medical Teacher*, *44*(4), 401–409. https://doi.org/10.1080/0142159X.2021.1998401

Barajas-Ochoa, A., Ramos-Remus, C., Castillo-Ortiz, J. D., Yáñez, J., Barajas-Ochoa, Z., Sánchez-González, J. M., Hernández-Ávila, M., Córdova-Villalobos, J. Á., & Bustamante-Montes, L. P. (2019). Flaws in the design of the Examen Nacional para Aspirantes a Residencias Médicas produce inequity. *Salud Publica de Mexico*, *61*(2), 125–135. https://doi.org/10.21149/9790

Bárquez Antillán, I. L., & Vivian Mascareño, M. F. (2009). La validez predictiva de la prueba de aptitud académica (PAA) respecto al desempeño académico de los estudiantes de la Universidad La Salle Noroeste , A.C. En A. C. Consejo Mexicano de Investigación Educativa (Ed.), *X Congreso Nacional de Investigación Educativa*. Consejo Mexicano de Investigación Educativa, A.C. https://www.comie.org.mx/congreso/memoriaelectronica/v10/contenido/contenido0101T.htm

Bhola, D. S., Impara, J. C., & Buckendahl, C. W. (2003). Aligning Tests with States' Content Standards: Methods and Issues. *Educational Measurement: Issues and Practice*, *22*(3), 21–29. https://doi.org/10.1111/j.1745-3992.2003.tb00134.x

Blair, C. (2006). How similar are fluid cognition and general intelligence? A developmental neuroscience perspective on fluid cognition as an aspect of human cognitive ability. *Behavioral and Brain Sciences*, *29*(2), 109–125. https://doi.org/10.1017/S0140525X06009034

Boer, D., Hanke, K., & He, J. (2018). On Detecting Systematic Measurement Error in Cross-Cultural Research: A Review and Critical Reflection on Equivalence and Invariance Tests. *Journal of Cross-Cultural Psychology*, *49*(5), 713–734. https://doi.org/10.1177/0022022117749042

Borghi, A. M., Binkofski, F., Castelfranchi, C., Cimatti, F., Scorolli, C., & Tummolini, L. (2017). The challenge of abstract concepts. *Psychological Bulletin*, *143*(3), 263–292. https://doi.org/10.1037/bul0000089

Borsboom, D., Mellenbergh, G. J., & Van Heerden, J. (2004). The concept of validity. *Psychological Review*, *111*(4), 1061–1071. https://doi.org/10.1037/0033-295X.111.4.1061

Brennan, R. (2013). Commentary on "Validating the Interpretations and Uses of Test Scores". *Journal of Educational Measurement*, *50*(1), 74–83. https://doi.org/10.1111/jedm.12001

Brualdi, A. (1999). Traditional and Modern Concepts of Validity. En *ERIC/AE Digest*. https://files.eric.ed.gov/fulltext/ED435714.pdf

Camargo, S. L., Herrera, A. N., & Traynor, A. (2018). Looking for a Consensus in the Discussion about the Concept of Validity: A Delphi Study. *Methodology*, *14*(4), 146–155. https://doi.org/10.1027/1614-2241/a000157

Campbell, D., & Fiske, D. (1959). Convergent and discriminant validation by the multitrait-multimethod matrix. *Psychological Bulletin*, *56*(2), 81–105.

Carrillo Avalos, B. A., Banda Lara, M. I., & Zavala Cruz, G. G. (2016). Valor predictivo del examen de admisión en el desempeño académico en las ciencias morfológicas. *V CONGRESO INTERNACIONAL DE EDUCACIÓN MÉDICA. "Educación Médica en las Américas"*.

Centro de Salud Universitario de la UASLP. (s/f). *Características del examen psicométrico*.

Centro Nacional de Evaluación para la Educación Superior. (2013a). *Guía del examen nacional de ingreso a la educación superior (EXANI-II)* (Número 1).

Centro Nacional de Evaluación para la Educación Superior. (2013b). *Guía del examen nacional de ingreso a la educación superior (EXANI-II) 2014* (19a ed.). Centro Nacional de Evaluación para la Educación Superior, A. C.

Centro Nacional de Evaluación para la Educación Superior. (2014). *Examen General para el Egreso de la Licenciatura en Médico General*. http://www.ceneval.edu.mx/documents/20182/35042/Contenidodelaprueba.pdf/8f2ca90a-9868-4cd6-8512-9d04dfdc00e6

Centro Nacional de Evaluación para la Educación Superior. (2016). *Resultados del examen nacional de ingreso a la educación superior en el año 2016*.

Centro Nacional de Evaluación para la Educación Superior. (2018). *Medicina General - Ceneval*. http://www.ceneval.edu.mx/medicina-general

Centro Nacional de Evaluación para la Educación Superior. (2023). *Módulo de conocimientos disciplinares específicos de Premedicina EXANI-II*. https://online.flippingbook.com/view/278838270/

Chalhoub-Deville, M. (2016). Validity theory: Reform policies, accountability testing, and consequences. *Language Testing*, *33*(4), 453–472. https://doi.org/10.1177/0265532215593312

Chiavaroli, N. (2017). Negatively-worded multiple choice questions: An avoidable threat to validity. *Practical Assessment, Research and Evaluation*, *22*(3), 1–14.

Cizek, G. J. (2012). Defining and distinguishing validity: Interpretations of score meaning and justifications of test use. *Psychological Methods*, *17*(1), 31–43. https://doi.org/10.1037/a0026975

Cizek, G. J. (2020). *Validity: an integrated approach to test score meaning and use.* Routledge.

Clauser, B. E., Margolis, M. J., & Swanson, D. B. (2008). Issues of validity and reliability for assessments in medical education. En E. Holmboe & R. Hawkins (Eds.), *A Practical Guide to the Evaluation of Clinical Competence* (pp. 10–23). Elsevier.

Coates, H. (2008). Establishing the criterion validity of the Graduate Medical School Admissions Test (GAMSAT). *Medical Education*, *42*(10), 999–1006. https://doi.org/10.1111/j.1365-2923.2008.03154.x

Collins, J. (2006). Education techniques for lifelong learning: Writing multiple-choice questions for continuing medical education activities and self-assessment modules. *Radiographics*, *26*(2), 543–551. https://doi.org/10.1148/rg.262055145

Comisión de Admisión de la Escuela de Medicina de la UASLP. (1977a). Informe de la comisión de admisión 1a parte. *Boletín Informativo de la Escuela de Medicina*, *20*(2), 25–45.

Comisión de Admisión de la Escuela de Medicina de la UASLP. (1977b). Informe de la comisión de admisión 2a parte. *Boletín Informativo de la Escuela de Medicina*, *20*(3), 49–92.

Comisión de Admisión de la Facultad de Medicina de la UASLP. (2009). *Reglamento de la comisión de admisión*.

Consejo Mexicano para la Acreditación de la Educación Médica (COMAEM). (2018). *Estatutos*. http://www.comaem.org.mx/?page_id=72

Consejo Mexicano para la Acreditación de la Educación Médica (COMAEM). (2023). *Estado Global de Acreditación*. https://www.comaem.org.mx/?page_id=2150

Consejo Técnico Consultivo de la Facultad de Medicina de la UASLP. (2018). *Sección Primera De la Estructura Organizacional*.

Cook, D. A., Brydges, R., Ginsburg, S., & Hatala, R. (2015). A contemporary approach to validity arguments: A practical guide to Kane's framework. *Medical Education*, *49*(6), 560–575. https://doi.org/10.1111/medu.12678

Cook, D. A., & Hatala, R. (2016). Validation of educational assessments: a primer for simulation and beyond. *Advances in Simulation*, *1*(1), 1–12. https://doi.org/10.1186/s41077-016-0033-y

Cook, D. A., Kuper, A., Hatala, R., & Ginsburg, S. (2016). When assessment data are words: Validity evidence for qualitative educational assessments. *Academic Medicine*, *91*(10), 1359–1369. https://doi.org/10.1097/ACM.0000000000001175

Costello, A. B., & Osborne, J. W. (2005). Best practices in exploratory factor analysis: Four recommendations for getting the most from your analysis. *Practical Assessment, Research and Evaluation*, *10*(7).

Crawford, C., Black, P., Melby, V., & Fitzpatrick, B. (2021). An exploration of the predictive validity of selection criteria on progress outcomes for pre-registration nursing programmes—A systematic review. *Journal of Clinical Nursing*, *30*(17–18), 2489–2513. https://doi.org/10.1111/jocn.15730

Cronbach, L. J. (1988). Five perspectives on validity argument. En H. Wainer & H. I. Braun (Eds.), *Test validity* (pp. 3–17). Routledge. https://doi.org/10.4324/9780203056905

Cronbach, L. J., & Meehl, P. E. (1955). Construct validity in psychological tests. *Psychological Bulletin*, *52*(4), 281–302. https://doi.org/10.1037/h0040957

Crooks, T. J., Kane, M. T., & Cohen, A. S. (1996). Threats to the valid use of assessments. *International Journal of Phytoremediation*, *21*(1), 265–286. https://doi.org/10.1080/0969594960030302

Cunningham, C., Patterson, F., & Cleland, J. (2019). A literature review of the predictive validity of European dental school selection methods. *European Journal of Dental Education*, *23*(2), 73–87. https://doi.org/10.1111/eje.12405

Cuschieri, A., Gleeson, F. A., Harden, R. M., & Wood, R. A. B. (1979). A new approach to a final examination in surgery. Use of the objective clinical examinations. *Annals of the Royal College of Surgeons of England*, *61*(5), 400–405.

Dancey, C., & Reidy, J. (2017). *Statistics Without Maths for Psychology* (7th ed.). Pearson.

Davies, D. J., Sam, A. H., Murphy, K. G., Khan, S. A., Choe, R., & Cleland, J. (2022). BMAT's predictive validity for medical school performance: A retrospective cohort study. *Medical Education*, *56*(9), 936–948. https://doi.org/10.1111/medu.14819

DGAE UNAM. (2017). *¿Qué onda con el PASE Reglamentado?*

DGAE UNAM. (2018). *Resultados concurso Febrero 2018*. https://www.dgae.unam.mx/Febrero2018/resultados/2/2080125.html

Downing, S. M. (2002a). Construct-irrelevant variance and flawed test questions: Do multiple-choice item-writing principles make any difference? *Academic Medicine*, *77*(10 SUPPL.), 103–104. https://doi.org/10.1097/00001888-200210001-00032

Downing, S. M. (2002b). Threats to the validity of locally developed multiple-choice tests in medical education: Construct-irrelevant variance and construct underrepresentation. *Advances in Health Sciences Education*, *7*(3), 235–241. https://doi.org/10.1023/A:1021112514626

Downing, S. M. (2003). Validity: On the meaningful interpretation of assessment data. *Medical Education*, *37*(9), 830–837. https://doi.org/10.1046/j.1365-2923.2003.01594.x

Downing, S. M. (2004). Reliability : on the reproducibility of assessment data. *Medical Education, 38*, 1006–1012. https://doi.org/10.1046/j.1365-2929.2004.01932.x

Downing, S. M. (2005). The effects of violating standard item writing principles on tests and students: The consequences of using flawed test items on achievement examinations in medical education. *Advances in Health Sciences Education*, *10*(2), 133–143. https://doi.org/10.1007/s10459-004-4019-5

Downing, S. M. (2006). Face validity of assessments: Faith-based interpretations or evidence-based science? En *Medical Education* (Vol. 40, Número 1, pp. 7–8). https://doi.org/10.1111/j.1365-2929.2005.02361.x

Downing, S. M., & Haladyna, T. M. (2004). Validity threats: Overcoming interference with proposed interpretations of assessment data. *Medical Education*, *38*(3), 327–333. https://doi.org/10.1046/j.1365-2923.2004.01777.x

Downing, S. M., & Yudkowski, R. (Eds.). (2009). *Assessment in health professions education*. Routledge.

Education Resources Information Center. (1966). *Test validity*. https://eric.ed.gov/?qt=validity&ti=Test+Validity

Edwards, D., Friedman, T., & Pearce, J. (2013). Same admissions tools , different outcomes : a critical perspective on predictive validity in three undergraduate medical schools. *BMC Medical Education*, *13*. https://doi.org/https://doi.org/10.1186/1472-6920-13-173

Embretson, S. E. (1998). A Cognitive Design System Approach to Generating Valid Tests: Application to Abstract Reasoning. *Psychological Methods*, *3*(3), 380–396. https://doi.org/10.1037/1082-989X.3.3.380

Espinoza del Río, M. Alberto. (2017). *Análisis de los resultados de los alumnos de nuevo ingreso del ciclo escolar 2014-2015 para la licenciatura de Médico cirujano de la facultad de medicina de la UASLP*. Universidad Autónoma de San Luis Potosí.

Facultad de Medicina de la UASLP. (2015). *Perfil de Ingreso: Características Deseables en el Estudiante*. http://www.medicina.uaslp.mx/Oferta_Educativa/Medico_Cirujano

Facultad de Medicina de la UASLP. (2017). *Mapa curricular de la carrera de médico cirujano*. http://www.medicina.uaslp.mx/Oferta_Educativa/Medico_Cirujano

Ferguson, E., James, D., & Madeley, L. (2002). Factors associated with success in medical school: systematic review of the literature. *BMJ: British Medical Journal*, *324*(April), 952–957.

Ferrando, P. J., & Anguiano-Carrasco, C. (2010). El análisis factorial como técnica de investigación en psicología. *Papeles del Psicólogogo*, *31*(1), 18–33.

Finnerty, P. E. (2010). The Role and Value of the Basic Sciences in Medical Education: An Examination of Flexner's Legacy. *Medical Science Educator, 85*(2), 349–355.

Fraenkel, J. R., Wallen, N. E., & Hyun, H. H. (2019). *How to Design and Evaluate Research in Education* (10th ed.). Mc Graw-Hill Education.

Ganjoo, R., Schwartz, L., Boss, M., McHarg, M., & Dobrydneva, Y. (2020). Predictors of success on the MCAT among post-baccalaureate pre-medicine students. *Heliyon*, *6*(4). https://doi.org/10.1016/j.heliyon.2020.e03778

García Bonilla, C. E., & Noyola Bernal, J. E. (2001). Informe de la Comisión de Admisión 1994-2001. *Boletín Informativo de la Facultad de Medicina*, *44*, 184–186.

García Domínguez, L. A. (2016). Pruebas de selección como predictores del rendimiento académico de estudiantes de Medicina. *Investigación en Educación Médica*, *5*(18), 88–92. https://doi.org/10.1016/j.riem.2016.01.018

Garrocho Sandoval, C., & Torre López, E. (1970). Nuevo enfoque de las pruebas de admisión a la escuela de medicina. *Boletín Informativo de la Escuela de Medicina*, *15*(1), 1–15.

Gasmalla, H. E. E., & Tahir, M. E. (2020). The validity argument: Addressing the misconceptions. *Medical Teacher*, *0*(0), 1–8. https://doi.org/10.1080/0142159x.2020.1856802

Gómez-Benito, J., Sireci, S., Padilla, J. L., Dolores Hidalgo, M., & Benítez, I. (2018). Differential item functioning: Beyond validity evidence based on internal structure. *Psicothema*, *30*(1), 104–109. https://doi.org/10.7334/psicothema2017.183

Griffin, B., Horton, G. L., Lampe, L., Shulruf, B., & Hu, W. (2021). The change from UMAT to UCAT for undergraduate medical school applicants: impact on selection outcomes. *Medical Journal of Australia*, *214*(2), 84–89. https://doi.org/10.5694/mja2.50877

Guerrero M, A. (1979). Evaluación de aptitudes en aspirantes a la carrera de medicina de la Universidad Autónoma de San Luis Potosí. *Boletín Informativo de la Escuela de Medicina, 22*(1), 19–24.

Habersack, M., Dimai, H. P., Ithaler, D., & Reibnegger, G. (2015). Time: an underestimated variable in minimizing the gender gap in medical college admission scores. *Wiener Klinische Wochenschrift*, *127*(7–8), 241–249. https://doi.org/10.1007/s00508-014-0649-7

Hadie, S. N. H. (2018). The Application of Learning Taxonomy in Anatomy Assessment in Medical School. *Education in Medicine Journal*, *10*(1), 13–23. https://doi.org/10.21315/eimj2018.10.1.3

Haertel, E. (2013). Getting the Help We Need. *Journal of Educational Measurement*, *50*(1), 84–90. https://doi.org/10.1111/jedm.12002

Hafferty, F. W., O'Brien, B. C., & Tilburt, J. C. (2020). Beyond High-Stakes Testing: Learner Trust, Educational Commodification, and the Loss of Medical School Professionalism. *Academic Medicine*, *95*(6), 833–837. https://doi.org/10.1097/ACM.0000000000003193

Haladyna, T. M., & Downing, S. M. (2004). Construct-Irrelevant Variance in High-Stakes Testing. *Educational Measurement: Issues and Practice*, *23*(1), 17–27. https://onlinelibrary.wiley.com/doi/abs/10.1111/j.1745-3992.2004.tb00149.x

Hatala, R., Cook, D. A., Brydges, R., & Hawkins, R. (2015). Constructing a validity argument for the Objective Structured Assessment of Technical Skills (OSATS): a systematic review of validity evidence. *Advances in Health Sciences Education*, *20*(5), 1149–1175. https://doi.org/10.1007/s10459-015-9593-1

Hawkins, R. E., Margolis, M. J., Durning, S. J., & Norcini, J. J. (2010). Constructing a validity argument for the mini-clinical evaluation exercise: A review of the research. *Academic Medicine*, *85*(9), 1453–1461. https://doi.org/10.1097/ACM.0b013e3181eac3e6

Hernández-Mata, J. M., Hernández-Castro, R., Nieto-Caraveo, A., & Hernández Sierra, J. F. (2005). Factores de riesgo para la deserción de estudiantes en la Facultad de Medicina de la Universidad Autónoma de San Luis Potosí (UASLP), México. *Gaceta Médica de México*, *141*(5), 445–447. https://www.scielo.org.mx/scielo.php?script=sci_arttext&pid=S0016-38132005000500016

Hicks, N. A. (2011). Guidelines for identifying and revising culturally biased multiple-choice nursing examination items. *Nurse Educator*, *36*(6), 266–270. https://doi.org/10.1097/NNE.0b013e3182333fd2

Instituto Nacional para la Evaluación de la Educación. (2017). Criterios técnicos para el desarrollo, uso y mantenimiento de instrumentos de evaluación. En *Diario Oficial de la Federación*. https://www.inee.edu.mx/wp-content/uploads/2019/04/P1E104.pdf

Johnson, R. C. (2011). *Assessing the assessments: Using an argument-based validity framework to assess the validity and use of an English placement system in a foreign language context*. Macqarie University.

Jozefowicz, R. F., Koeppen, B. M., Case, S., Galbraith, R., Swanson, D., & Glew, R. H. (2002). The quality of in-house medical school examinations. *Academic Medicine*, *77*(2), 156–161. https://doi.org/10.1097/00001888-200202000-00016

Julian, E. R. (2005). Validity of the Medical College Admission Test for predicting medical school performance. *Academic Medicine*, *80*(10), 910–917. https://doi.org/10.1097/00001888-200510000-00010

Jurado-Núñez, A., & Leenen, I. (2016). Reflexiones sobre adivinar en preguntas de opción múltiple y cómo afecta el resultado del examen. *Investigación en Educación Médica*, *5*(17), 55–63. https://doi.org/10.1016/j.riem.2015.07.004

Kane, M. T. (1992). An argument-based approach to validation. *Psychological Bulletin*, *112*(3), 527–535. https://doi.org/10.1037/0033-2909.112.3.527

Kane, M. T. (2009). Validating the interpretations and uses of test scores. En R. W. Lissitz (Ed.), *Validity: Revisions, New Directions and Applications* (pp. 39–64). Information Age Publishing, Inc.

Kane, M. T. (2011). Validating score interpretations and uses: Messick Lecture, Language Testing Research Colloquium, Cambridge, April 2010. *Language Testing*, *29*(1), 3–17. https://doi.org/10.1177/0265532211417210

Kane, M. T. (2013a). Validating the Interpretations and Uses of Test Scores. *Journal of Educational Measurement*, *50*(1), 1–73. https://doi.org/10.1111/jedm.12000

Kane, M. T. (2013b). Validation as a Pragmatic, Scientific Activity. *Journal of Educational Measurement*, *50*(1), 115–122. https://doi.org/10.1111/jedm.12007

Kelly, M. E., Patterson, F., O'flynn, S., Mulligan, J., & Murphy, A. W. (2018). A systematic review of stakeholder views of selection methods for medical schools admission. *Medical Education*, *18*(July), 139–164. https://doi.org/10.1186/s12909-018-1235-x

Kinnear, B., Kelleher, M., May, B., Sall, D., Schauer, D. P., Schumacher, D. J., & Warm, E. J. (2021). Constructing a Validity Map for a Workplace-Based Assessment System: Cross-Walking Messick and Kane. *Academic Medicine*, *96*(7), S64–S69. https://doi.org/10.1097/ACM.0000000000004112

Kinnear, B., Varpio, L., Schumacher, D. J., & Driessen, E. W. (2022). How argumentation theory can inform assessment validity : A critical review. *Medical Education*, 1–12. https://doi.org/10.1111/medu.14882

Lambe, P., Waters, C., & Bristow, D. (2012). The UK clinical aptitude Test: Is it a fair test for selecting medical students? *Medical Teacher*, *34*(8). https://doi.org/10.3109/0142159X.2012.687482

Lane, S. (2014). Evidencia de validez basada en las consecuencias del uso del test. *Psicothema*, *26*(1), 127–135. https://doi.org/10.7334/psicothema2013.258

Lane, S. (2020). Test-Based Accountability Systems: The Importance of Paying Attention to Consequences. *ETS Research Report Series*, *2020*(1), 1–22. https://doi.org/10.1002/ets2.12283

Lane, S., Raymond, M., & Haladyna, T. (2016). Handbook of Test Development. En S. Lane, M. Raymond, & T. Haladyna (Eds.), *International Journal of Testing* (2nd ed.). Routledge. https://doi.org/10.1080/15305050701813433

Larsen, R., & Warne, R. T. (2010). Estimating confidence intervals for eigenvalues in exploratory factor analysis. *Behavior Research Methods*, *42*(3), 871–876. https://doi.org/10.3758/BRM.42.3.871

Leenen, I. (2014). Virtudes y limitaciones de la teoría de respuesta al ítem para la evaluación educativa en las ciencias médicas. *Investigación en Educación Médica*, *3*(9), 40–55.

Leiner, J. E. M., Scherndl, T., & Ortner, T. M. (2018). How do men and women perceive a high-stakes test situation? *Frontiers in Psychology*, *9*(DEC), 1–14. https://doi.org/10.3389/fpsyg.2018.02216

Leiva Garza, J. L. (2003). 6. Desarrollo Académico. En J. E. Noyola Bernal & E. R. Zazueta Quirarte (Eds.), *Historia de la Facultad de Medicina de la Universidad Autónoma de San Luis Potosí 1977-2002* (pp. 118–220). Editorial Universitaria Potosina.

Lin, J. C., Lokhande, A., Margo, C. E., & Greenberg, P. B. (2022). Best practices for interviewing applicants for medical school admissions: a systematic review. *Perspectives on Medical Education*, *11*(5), 239–246. https://doi.org/10.1007/s40037-022-00726-8

Llamas, E., & Escalante, V. W. (2001). Predicting results of morphology teaching in non-English-speaking countries. *Anatomical Record*, *265*(4), 161–162. https://doi.org/10.1002/ar.1148

Lok, B., McNaught, C., & Young, K. (2016). Criterion-referenced and norm-referenced assessments: compatibility and complementarity. *Assessment and Evaluation in Higher Education*, *41*(3), 450–465. https://doi.org/10.1080/02602938.2015.1022136

Luschin-Ebengreuth, M., Dimai, H. P., Ithaler, D., Neges, H. M., & Reibnegger, G. (2016). Medical University admission test: a confirmatory factor analysis of the results. *Wiener Klinische Wochenschrift*, *128*(9–10), 376–383. https://doi.org/10.1007/s00508-015-0911-7

Lynch, B., MacKenzie, R., Dowell, J., Cleland, J., & Prescott, G. (2009). Does the UKCAT predict Year 1 performance in medical school? *Medical Education*, *43*(12), 1203–1209. https://doi.org/10.1111/j.1365-2923.2009.03535.x

Lyons-Thomas, J., Liu, Y., & Zumbo, B. D. (2014). Validity and Validation in Social, Behavioral, and Health Sciences: A Synthesis of Syntheses. En B. D. Zumbo & E. K. H. Chan (Eds.), *Validity and Validation in Social, Behavioral, and Health Sciences* (1a ed., pp. 313–319). Springer International Publishing. https://doi.org/10.1007/978-3-319-07794-9

Manterola, C., Rivadeneira, J., Delgado, H., Soteldo, C., & Otzen, T. (2023). ¿Cuántos Tipos de Revisiones de la Literatura Existen? Enumeración, descripción y clasificación. Revisión cualitativa. *International Journal of Morphology*, *41*(4), 1240–1253. http://dx.doi.org/10.4067/S0717-95022023000401240

Martínez Arias, M. R., Hernández Lloreda, M. V., & Hernández Lloreda, M. J. (2014). *Psicometría*. Alianza Editorial.

Martínez Villarreal, R. T. (2013). *Valor predictivo del examen nacional de ingreso en la licenciatura de medicina en la Universidad Autónoma de Nuevo León, México* [Facultad de Medicina de la Universidad Complutense de Madrid]. https://hdl.handle.net/20.500.14352/38125

McManus, I. C., Ferguson, E., Wakeford, R., Powis, D., & James, D. (2011). Predictive validity of the Biomedical Admissions Test: An evaluation and case study. *Medical Teacher*, *33*(1), 53–57. https://doi.org/10.3109/0142159X.2010.525267

Messick, S. (1989). Validity. En R. L. Linn (Ed.), *Educational Measurement* (3rd ed., pp. 13–104). Macmillan. https://doi.org/10.1002/j.2330-8516.1987.tb00244.x

Meyer, H., Zimmermann, S., Hissbach, J., Klusmann, D., & Hampe, W. (2019). Selection and academic success of medical students in Hamburg, Germany. *BMC Medical Education*, *19*(1), 1–15. https://doi.org/10.1186/s12909-018-1443-4

Miller, G. E. (1990). The assessment of clinical skills/competence/performance. *Academic Medicine*, *65*(9), S63–S67. https://doi.org/10.1097/00001888-199009000-00045

Miranda López, F., Pérez Güemes, E. E., Villamil Serrano, E., Márquez Gutiérrez, Y., González Chávez, G. A., & Hernández Gómez, J. E. (2018). *Encuesta de satisfacción de los procesos de evaluación de ingreso y promoción en educación básica y media superior 2017*. https://historico.mejoredu.gob.mx/wp-content/uploads/2018/12/P1F224.pdf

Moore, K., Dailey, A., & Agur, A. (2002). *Anatomía con orientación clínica* (4a ed.). Lippincot Williams & Wilkins/Editorial Médica Panamericana.

Moore, K., Dailey, A., & Agur, A. (2013). *Anatomía con orientación clínica* (7a ed.). Wolters Kluwer Health, S.A., Lippincot Williamns & Wilkins.

Morales Ibarra, R., Barrera Baca, A., & Mandujano Garnett, E. (2009). Validez predictiva y concurrente del EXANI-II, en la Universidad Autónoma del Estado de México. En A. C. Consejo Mexicano de Investigación Educativa (Ed.), *X Congreso Nacional de Investigación Educativa*. Consejo Mexicano de Investigación Educativa, A.C. https://www.comie.org.mx/congreso/memoriaelectronica/v10/contenido/contenido0116T.htm

Moreno, R., Martínez, R. J., & Muñiz, J. (2004). Directrices para la construcción de ítems de elección múltiple. *Psicothema*, *16*(3), 490–497. https://www.redalyc.org/articulo.oa?id=72716324

Moss, P. A. (1998). The role of consequences in validity theory. *Educational Measurement: Issues and Practice*, *17*(2), 6–12. https://doi.org/10.1111/j.1745-3992.1998.tb00826.x

Muñoz-Comonfort, A., Leenen, I., & Fortoul-van der Goes, T. I. (2014). Correlación entre la evaluación diagnóstica y el rendimiento académico de los estudiantes de medicina. *Investigación en Educación Médica*, *3*(10), 85–91. https://doi.org/10.1016/s2007-5057(14)72731-0

National Board of Medical Examiners. (2016). *Cómo elaborar preguntas para evaluaciones escritas en el área de ciencias básicas y clínicas* (M. A. Paniagua & K. A. Swygert, Eds.; 4th ed.). National Board of Medical Examiners.

NCBI. (s/f). *MeSH term Medical School*.

NCBI. (1991a). *MeSH term College Admission Test*. https://www.ncbi.nlm.nih.gov/mesh/?term=college+admission+test

NCBI. (1991b). *MeSH term School Admission Criteria*. https://www.ncbi.nlm.nih.gov/mesh/?term=School+Admission+Criteria

Nichols, P. D., & Williams, N. (2009). Consequences of test score use as validity evidence: Roles and responsibilities. *Educational Measurement: Issues and Practice*, *28*(1), 3–9. https://doi.org/10.1111/j.1745-3992.2009.01132.x

Norcini, J. J. (2003). Setting standards on educational tests. *Medical Education*, *37*(5), 464–469. https://doi.org/10.1046/j.1365-2923.2003.01495.x

Norman, G., van der Vleuten, C., & Newble, D. (2002). *International Handbook of Research in Medical Education* (G. Norman, C. van der Vleuten, & D. Newble, Eds.). Springer.

Padilla, J. L., & Benítez, I. (2014). Validity evidence based on response processes. *Psicothema*, *26*(1), 136–144. https://doi.org/10.7334/psicothema2013.259

Panda, N., Bahdila, D., Abdullah, A., Ghosh, A. J., Lee, S. Y., & Feldman, W. B. (2021). Association between USMLE Step 1 Scores and In-Training Examination Performance: A Meta-Analysis. *Academic Medicine*, *96*(12), 1742–1754. https://doi.org/10.1097/ACM.0000000000004227

Paton, L. W., McManus, I. C., Cheung, K. Y. F., Smith, D. T., & Tiffin, P. A. (2022). Can achievement at medical admission tests predict future performance in postgraduate clinical assessments? A UK-based national cohort study. *BMJ Open*, *12*(2), 1–12. https://doi.org/10.1136/bmjopen-2021-056129

Patterson, F., Prescott-Clements, L., Zibarras, L., Edwards, H., Kerrin, M., & Cousans, F. (2016). Recruiting for values in healthcare: a preliminary review of the evidence. *Advances in Health Sciences Education*, *21*(4), 859–881. https://doi.org/10.1007/s10459-014-9579-4

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. *Medical Teacher*, 1–11. https://doi.org/10.1080/0142159X.2018.1498589

Patterson, F., Zibarras, L., & Ashworth, V. (2016). Situational judgement tests in medical education and training: Research, theory and practice: AMEE Guide No. 100. *Medical Teacher*, *38*(1), 3–17. https://doi.org/10.3109/0142159X.2015.1072619

Powis, D. (2015). Selecting medical students: An unresolved challenge. *Medical Teacher*, *37*(3), 252–260. https://doi.org/10.3109/0142159X.2014.993600

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

Rios, J., & Wells, C. (2014). Evidencia de validez basada en la estructura interna. *Psicothema*, *26*(1), 108–116. https://doi.org/10.7334/psicothema2013.260

Royal, K. (2016). "Face validity" is not a legitimate type of validity evidence! *The American Journal of Surgery*, *212*, 1026–1027. https://doi.org/10.1016/j.amjsurg.2016.02.018

Royal, K. D., & Guskey, T. R. (2015). Editorial: On the appropriateness of norm-and criterion-referenced assessments in medical education. *Ear, Nose & Throat Journal*, *94*(7), 150–152. https://doi.org/https://doi.org/10.1177/014556131509400701

Russell, M. (2022). Clarifying the Terminology of Validity and the Investigative Stages of Validation. *Educational Measurement: Issues and Practice*, *41*(2), 25–35. https://doi.org/10.1111/emip.12453

Salehi, P. P., Azizzadeh, B., & Lee, Y. H. (2021). Pass/Fail Scoring of USMLE Step 1 and the Need for Residency Selection Reform. *Otolaryngology - Head and Neck Surgery (United States)*, *164*(1), 9–10. https://doi.org/10.1177/0194599820951166

Schreiber, J. B. (2021). Issues and recommendations for exploratory factor analysis and principal component analysis. *Research in Social and Administrative Pharmacy*, *17*(5), 1004–1011. https://doi.org/10.1016/j.sapharm.2020.07.027

Schreurs, S., Cleland, J., Muijtjens, A. M. M., oude Egbrink, M. G. A., & Cleutjens, K. (2018). Does selection pay off? A cost–benefit comparison of medical school selection and lottery systems. *Medical Education*, *52*(12), 1240–1248. https://doi.org/10.1111/medu.13698

Schuwirth, L. W. T., & Van Der Vleuten, C. P. M. (2011). General overview of the theories used in assessment: AMEE Guide No. 57. *Medical Teacher*, *33*(10), 783–797. https://doi.org/10.3109/0142159X.2011.611022

Schuwirth, L. W. T., & van der Vleuten, C. P. M. (2012). Programmatic assessment and Kane's validity perspective. *Medical Education*, *46*(1), 38–48. https://doi.org/10.1111/j.1365-2923.2011.04098.x

Shulruf, B., Bagg, W., Begun, M., Hay, M., Lichtwark, I., Turnock, A., Warnecke, E., Wilkinson, T. J., & Poole, P. J. (2018). The efficacy of medical student selection tools in Australia and New Zealand. *The Medical Journal of Australia*, *208*(5), 214–218. https://doi.org/10.5694/mja17.00400

Shulruf, B., Poole, P., Wang, G. Y., Rudland, J., & Wilkinson, T. (2012). How well do selection tools predict performance later in a medical programme? *Advances in Health Sciences Education*, *17*(5), 615–626. https://doi.org/10.1007/s10459-011-9324-1

Simpson, P. L., Scicluna, H. A., Jones, P. D., Cole, A. M. D., O'Sullivan, A. J., Harris, P. G., Velan, G., & McNeil, H. P. (2014). Predictive validity of a new integrated selection process for medical school admission. *BMC Medical Education*, *14*(1), 1–10. https://doi.org/10.1186/1472-6920-14-86

Sireci, S., & Faulkner-Bond, M. (2014). Validity evidence based on test content. *Psicothema*, *26*(1), 100–107. https://doi.org/10.7334/psicothema2013.256

Sireci, S. G. (2016a). Comments on valid (and invalid?) commentaries. *Assessment in Education: Principles, Policy and Practice*, *23*(2), 319–321. https://doi.org/10.1080/0969594X.2016.1158694

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

Sireci, S. G., & Greiff, S. (2019). Editorial: On the importance of educational tests. *European Journal of Psychological Assessment*, *35*(3), 297–300. https://doi.org/10.1027/1015-5759/a000549

Soemantri, D., Karunathilake, I., Yang, J. H., Chang, S. C., Lin, C. H., Nadarajah, V. D., Nishigori, H., Samarasekera, D. D., Lee, S. S., Tanchoco, L. R., & Ponnamperuma, G. (2020). Admission policies and methods at crossroads: A review of medical school admission policies and methods in seven Asian countries. *Korean Journal of Medical Education*, *32*(2), 243–256. https://doi.org/10.3946/KJME.2020.169

Stellefson, M. L., Hanik, B. W., Chaney, B. H., & Chaney, J. D. (2009). Factor retention in EFA: Strategies for health behavior researchers. *American Journal of Health Behavior*, *33*(5), 587–599. https://doi.org/10.5993/AJHB.33.5.12

St-Onge, C., Young, M., Eva, K. W., & Hodges, B. (2017). Validity: one word with a plurality of meanings. *Advances in Health Sciences Education*, *22*(4), 853–867. https://doi.org/10.1007/s10459-016-9716-3

Tamimi, A., Hassuneh, M., Tamimi, I., Juweid, M., Shibli, D., AlMasri, B., & Tamimi, F. (2023). Admission criteria and academic performance in medical school. *BMC Medical Education*, *23*(1), 1–9. https://doi.org/10.1186/s12909-023-04251-y

Tarrant, M., Knierim, A., Hayes, S. K., & Ware, J. (2006). The frequency of item writing flaws in multiple-choice questions used in high stakes nursing assessments. *Nurse Education Today*, *26*(8), 662–671. https://doi.org/10.1016/j.nepr.2006.07.002

Tavakol, M., & Dennick, R. (2017). The foundations of measurement and assessment in medical education. *Medical Teacher*, *39*(10), 1010–1015. https://doi.org/10.1080/0142159X.2017.1359521

Torre López, J. M. (1960a). Instructivo para admisión de alumnos. *Boletín Informativo de la Escuela de Medicina*, *2*(6), 61–63.

Torre López, J. M. (1960b). Selección de alumnos para estudiar la carrera de médico. *Boletín Informativo de la Escuela de Medicina*, *2*(1), 1–10.

Torre López, J. M. (1962). Resultado de la selección de alumnos efectuada en 1961. *Boletín Informativo de la Escuela de Medicina*, *4*(7), 89–94.

Torre López, J. M. (1963a). La selección de alumnos de primer ingreso para 1964. *Boletín Informativo de la Escuela de Medicina*, *5*(6), 89–91.

Torre López, J. M. (1963b). Resultados de la selección de alumnos efectuada en 1962. *Boletín Informativo de la Escuela de Medicina*, *5*(7), 105–111.

Torre López, J. M. (1964). Resultados de la selección de alumnos efectuada en 1963. *Boletín Informativo de la Escuela de Medicina*, *6*(10), 153–158.

Torre López, J. M. (1966). La selección de estudiantes de primer ingreso a la escuela de medicina. *Boletín Informativo de la Escuela de Medicina*, *8*(9), 125–139.

Tweed, M., & Cookson, J. (2001). The face validity of a final professional clinical examination. *Medical Education*, *35*(5), 465–473. https://doi.org/10.1046/j.1365-2923.2001.00895.x

UASLP. (2009). *Reglamento de transparencia y acceso a la información pública de la UASLP*.

UASLP. (2012). *Informe 2011-2012. Docencia.* (UASLP, Ed.).

UASLP. (2013). *Informe 2012-2013* (UASLP, Ed.).

UASLP. (2014a). *Guía Temática del Examen de Conocimientos Facultad de Medicina*.

UASLP. (2014b). *Informe 2013-2014*.

UASLP. (2014c). *Plan institucional de desarrollo 2013-2023*. http://www.uaslp.mx/PIDE/Documents/PIDE_2013_2023.pdf

UASLP. (2015a). *Informe 2014-2015*. UASLP.

UASLP. (2015b). *Instructivo para aspirantes de nuevo ingreso 2015-2016*. http://www.uaslp.mx/ServiciosEscolares/Documents/Instructivo_aspirantes2015.pdf

UASLP. (2016). *Informe 2012-2016* (UASLP, Ed.). http://www.cenam.gob.mx/sm2010/info/pviernes/sm2010-vp04d.pdf

UASLP. (2017a). *De Nuevo Ingreso 2017-2018*. http://www.uaslp.mx/ServiciosEscolares/Documents/Instructivo-para-aspirantes-preinscritos-.pdf

UASLP. (2017b). *Guía Temática del Examen de Conocimientos Facultad de Medicina*. http://www.uaslp.mx/ServiciosEscolares/Documents/pdfs-admisiones2017-2018/guias-2017-2018/Medicina-2017.pdf

UKCAT Consortium. (2023, abril 5). *Test Statistics*. https://www.ucat.ac.uk/results/test-statistics/

Villegas Vizcaíno, R. (2017). ¿Es el EXANI II un predictor o un factor de exclusión social? *Educación y Ciencia*, *6*(48), 43–52. http://www.educacionyciencia.org/index.php/educacionyciencia/article/view/444

Violato, C., & Donnon, T. (2005). Does the medical college admission test predict clinical reasoning skills? A longitudinal study employing the Medical Council of Canada clinical reasoning examination. *Academic medicine : journal of the Association of American Medical Colleges*, *80*(10), S14–S16. https://doi.org/80/10_suppl/S14 [pii]

Violato, C., Gauer, J. L., Violato, E. M., & Patel, D. (2020). A study of the validity of the new MCAT exam. *Academic Medicine*, *95*(3), 396–400. https://doi.org/10.1097/ACM.0000000000003064

Ware, J., & Vik, T. (2009). Quality assurance of item writing: During the introduction of multiple choice questions in medicine for high stakes examinations. *Medical Teacher*, *31*(3), 238–243. https://doi.org/10.1080/01421590802155597

Wong, S., Yang, L., Riecke, B., Cramer, E., & Neustaedter, C. (2017). Assessing the usability of smartwatches for academic cheating during exams. *Proceedings of the 19th International Conference on Human-Computer Interaction with Mobile Devices and Services, MobileHCI 2017*. https://doi.org/10.1145/3098279.3098568

Yela, M. (1996). Los tests y el análisis factorial. *Psicothema*, *8*(Supl), 73–88.

York, T. T., Gibson, C., & Rankin, S. (2015). Defining and measuring academic success. *Practical Assessment, Research and Evaluation*, *20*(5), 1–20.

Young, J. W. (2008). Ensuring valid content tests for English Language Learners. En *Educational Testing Service* (Número 8).

Young, M., St-Onge, C., Xiao, J., Vachon Lachiver, E., & 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

Yuni, J. A., & Urbano, C. A. (2014). Técnicas para investigar, recursos metodologicos para la preparacion de proyectos de investigacion. En *Recursos Metodológicos para la Preparación de Proyectos de Investigación* (Vol. 2). Editorial Brujas.

Zazueta Quirarte, E. R. (2003). 5. Despertar a la modernidad. Las ideas se traducen en acciones. En J. E. Noyola Bernal & E. R. Zazueta Quirarte (Eds.), *Historia de la Facultad de Medicina de la Universidad Autónoma de San Luis Potosí 1977-2002* (pp. 74–117). Editorial Universitaria Potosina.

9. ANEXOS TESIS

ANEXO A. DICTAMEN DE APROBACIÓN POR EL COMITÉ DE INVESTIGACIÓN DE LA FACULTAD DE MEDICINA DE LA UASLP.

UASLP
Universidad Autónoma de San Luis Potosí

FACULTAD DE MEDICINA

Dictamen del Comité investigación

San Luis Potosí, S.L.P. a 4 de diciembre del 2019

M.C. Blanca Ariadna Carrillo Ávalos
Cargo: Profesor de Tiempo Completo

Estimada Dra. Carrillo Ávalos:

Por medio de la presente le informo que el **Comité de Investigación (CI) de la Facultad de Medicina de la Universidad Autónoma de San Luis Potosí** ha revisado tanto el protocolo original titulado "*Una metodología para evaluar la validez de la interpretación de los resultados del proceso de admisión para ingresar a la licenciatura en médico cirujano: aplicación con datos de la UASLP*" presentado en la **sesión extraordinaria el día 3 de diciembre del año en curso.**

En conformidad con lo dispuesto en la Ley General de Salud en Materia de investigación para la Salud, en la NOM-012-SSA3-2012, en el Manual de Integración y Funcionamiento de los Comités de Investigación, y en los Lineamientos Internos del Comité de Investigación de la Facultad de Medicina de la UASLP, **este Comité ha decidido dictaminar como APROBADO** el protocolo arriba mencionado. Motivo lo anterior se otorga el siguiente número de registro: **CI-006-2019**. Se turna el protocolo y copia de este dictamen al Comité de Ética en Investigación.

De acuerdo a lo estipulado en Lineamientos Internos del Comité de Investigación de la Facultad de Medicina de la UASLP, es **responsabilidad del Investigador Principal presentar un informe anual sobre los avances del proyecto**. Así mismo, en cuanto este protocolo cuente con la aprobación del Comité de Ética en Investigación, podrá iniciar con la ejecución del proyecto.

Sin más por el momento quedo a sus órdenes.

Atentamente.

Dr. Andreu Comas García
Presidente del Comité Investigación

FACULTAD DE MEDICINA UASLP
COMITÉ DE INVESTIGACIÓN

www.uaslp.mx

Av. Venustiano Carranza 2405
CP 78210 • San Luis Potosí, S.L.P.
tel. (444) 826 2344 al 49
tel. Dirección (444) 826 2350
fax (444) 826 2352

ccp. Archivo
ccp. Comité de Ética en Investigación

ANEXO B. DICTAMEN DE APROBACIÓN POR EL COMITÉ DE ÉTICA EN INVESTIGACIÓN DE LA FACULTAD DE MEDICINA DE LA UASLP.

UASLP
Universidad Autónoma de San Luis Potosí

Dictamen del Comité de ética en investigación
Facultad de Medicina, UASLP

San Luis Potosi, S.L.P. a 18 de diciembre del 2019

M.C. Blanca Ariadna Carrillo Ávalos
Profesor de Tiempo Completo

Estimada Dra. Carrillo Ávalos:
Por medio de la presente le informo que el Comité de ética en investigación ha revisado y evaluado, de acuerdo con el reglamento vigente y los lineamientos de operación de CONBIOETICA, el protocolo de investigación titulado "Una metodología para evaluar la validez de la interpretación de los resultados del proceso de admisión para ingresar a la licenciatura en médico cirujano: aplicación con datos de la UASLP" presentado en la sesion ordinaria el dia 17 de diciembre del año en curso. Y se ha determinado por consenso:

Aprobarlo **(X)** No aprobarlo ( )

El número de Registro asignado es: **CEI-2019-004**

Con las siguientes consideraciones: Se recomienda asegurarse del anonimato de los datos de los participantes. Finalmente, se le solicita atentamente informar a este comité sobre cualquier inconveniente que se pudiera presentar durante el desarrollo del protocolo y se pide informar por escrito cuando el proyecto haya finalizado.
Sin más por el momento quedo a sus órdenes para cualquier duda y aclaración.

Atentamente,

Dra. Sofía Bernal Silva
Presidente del Comité de Ética en investigación

ccp. Dr. Andreu Comas Garcia. Presidente del comité de Investigación, Facultad de Medicina. UASLP.
ccp. Archivo

FACULTAD DE MEDICINA
Av. Venustiano Carranza 2405
CP 78210 • San Luis Potosí, S.L.P.
tel. (444) 826 2344 al 49
tel. Dirección (444) 826 2350
fax (444) 826 2352
www.uaslp.mx

Componentes del Examen de Admisión.

La selección de los aspirantes a ingresar a la Facultad de Medicina se hace con base en los resultados obtenidos en los exámenes:

(1) Psicométrico,

(2) Examen de Admisión para las Instituciones de Enseñanza Superior del Centro Nacional de Evaluación (CENEVAL), y

(3) Examen Tradicional.

Examen Psicométrico.

El examen psicométrico se realiza en el Centro de Salud Universitario en la fecha y hora que se le ha programado. Si aún no tiene cita para el mismo, diríjase a la brevedad posible al Departamento de Admisiones o al Centro de Salud Universitario localizado en la Zona Universitaria (Manuel Nava esquina con Salvador Nava, frente al monumento a Manuel José Othón). El resultado de este examen estará incluido en la calificación que obtendrá en el Examen de Admisión. Su ponderación es de un 15 % del total de la calificación. No puede ser admitido a presentar los siguientes dos exámenes si no ha presentado el examen psicométrico.

Examen de admisión para las instituciones de enseñanza superior del CENEVAL.

Esta prueba es obligatoria para todos los aspirantes y se llevará a cabo el 05 de julio de 2014, de las 16:00 a las 18:00 horas. Se le solicita al sustentante formarse a las 15:00 horas, ya que el ingreso al auditorio será a partir de las 15:30 horas. El examen será aplicado en las instalaciones del Centro Cultural Bicentenario, ubicado en avenida Sierra Leona y Camino a la Presa. Le hacemos notar que este examen tiene su propio instructivo. Si no se le ha proporcionado puede obtenerlo del sitio de internet del CENEVAL: www.ceneval.edu.mx. El resultado de este examen estará incluido en la calificación que obtendrá en el Examen de Admisión. Su ponderación es de un 40% del total de la calificación. De acuerdo a la normativa universitaria vigente tiene la obligación de presentarlo para poder ser considerado como aspirante a ingresar en la Universidad.

Examen Tradicional.

Será el sábado 05 de julio de 2014 de las 08:45 a las 12:45 horas. El examen será aplicado en las instalaciones del Centro Cultural Bicentenario, ubicado en avenida Sierra Leona y Camino a la Presa. Esta prueba es obligatoria para todos los aspirantes. El documento que ahora tiene en sus manos es el instructivo de esta sección. El resultado de este examen estará incluido en la calificación que obtendrá en el Examen de Selección. Su ponderación es de un 45% del total de la calificación. Los tópicos más importantes del examen tradicional de la Facultad de Medicina de la UASLP son los siguientes:

1. Física y matemáticas.

2. Biología.

3. Química.

4. Inglés.

5. Español.

En el área de Física y Matemáticas se exploran conocimientos básicos de álgebra, trigonometría, geometría analítica, relaciones y funciones, así como de lo básico sobre las principales leyes del movimiento, hidráulica, calor, temperatura, electricidad y magnetismo. En general se hacen preguntas conceptuales que pueden ser contestadas sin necesidad del uso de calculadora.

El área de Biología explora con mayor profundidad los conocimientos sobre el funcionamiento del cuerpo humano, los diferentes seres vivos que habitan el planeta, así como los principios de la Teoría de la Evolución y la Genética.

Dentro del área de Química se harán preguntas conceptuales sobre la composición de la materia y sus partículas.

El área de Inglés, que tiene especial importancia en este examen, es evaluado en dos formas diferentes: 1) Se hacen preguntas a partir de ciertas expresiones idiomáticas en inglés (por ejemplo refranes, analogías, etc.), donde se hace necesario no sólo una traducción literal sino una comprensión integral del significado de las palabras y las expresiones entre el español y el inglés. 2) Existe un cierto número de preguntas de otras áreas del conocimiento (Biología y Química) que serán redactadas en inglés.

En el área de Español, además de la comprensión de la lectura y el razonamiento, se explora la ortografía, signos de acentuación y semántica. También aquí se evalúa el conocimiento de las raíces latinas y griegas de diversas palabras del español.

Módulo Específico.

El examen incluye 160 preguntas que son contestadas en un sólo módulo y se desarrollará de las 08:45 horas a las 12:45 horas. El módulo consta de las siguientes secciones:

Física y Matemáticas 30 preguntas

Biología 40 preguntas

Química 40 preguntas

Inglés 40 preguntas

Español 30 preguntas

Las 5 secciones están constituidas en su totalidad por preguntas con cinco posibilidades de respuesta, de las que sólo una es correcta. El hecho de contestar erróneamente una pregunta no le resta puntos ni le somete a penalización alguna. Dicho de otro modo, no deje preguntas sin contestar.

Modalidades de Preguntas.

A continuación se presentan las categorías en que podrían estar redactadas las preguntas. El examen no se limita a estas formas; sin embargo, esta muestra resulta significativa. La mayoría de los ejemplos están tomados de exámenes ya aplicados.

Frases incompletas.

Una habilidad semejante a la usada para resolver series se explora en las preguntas donde se trata de que usted complete la frase. Esto ayuda a medir la capacidad para identificar las relaciones que guardan diferentes tipos de elementos. La lógica de la oración es, sin duda, el aspecto crucial en las preguntas donde hay que completar oraciones. En este formato de preguntas se muestra un texto en el que se han omitido una o más palabras. Lo que se pide es completarlo de tal manera que forme un todo armónico, coherente y, sobre todo, lógico. El completado de oraciones exige del aspirante algo más que la mera comprensión de lo que significan los términos de las opciones, requiere que el examinado tenga una idea de su uso dentro del contexto de la oración. Cada oración contiene la información y los indicadores gramaticales necesarios para que se pueda identificar la opción correcta.

Ejemplo: Una célula viva humana presenta siempre un _______ ya que el humano es un ejemplo de animal eucarionte.

Respuesta: núcleo

Analogías y relaciones.

Estas preguntas están basadas más directamente en el pensamiento analógico. Exigen entender los conceptos y las relaciones entre ellos e identificar las relaciones similares o paralelas. En matemática son de este formato, por ejemplo, las preguntas de razones y proporciones.

Ejemplo: Un eritrocito es al sistema circulatorio lo que...

Respuesta: ...las hojas de un libro son a sus portadas.

Construcción o reconstrucción de textos.

Una de las formas de medir la capacidad de razonamiento verbal es presentar un texto de forma desordenada y solicitar su reordenamiento.

Ejemplo: aparato el El es mide pH. potenciómetro que un

Respuesta: El potenciómetro es un aparato que mide el pH.

Clasificación y manejo de datos.

Otras habilidades necesarias para el trabajo escolar son las que nos permiten seleccionar, ordenar y clasificar datos. Como en los ejemplos anteriores, será necesario aguzar la observación de semejanzas y diferencias, regulares e irregularidades, todo y partes, enlaces o relaciones obvias.

Ejemplo: De las siguientes opciones descarte aquella que no tenga relación con las cuatro restantes: absorción, cartílago articular, digestión, peristaltismo, secreción hormonal.

Respuesta: cartílago articular.

Comprensión de datos.

El examen también le pedirá atención y dedicación a las preguntas de comprensión de textos. La comprensión de la lectura se relaciona con diversos procesos del pensamiento, entre losque destacan: el análisis y la síntesis, la interpretación de opiniones, principios o dichos; la generalización y la discriminación verbal. Los textos pueden pertenecer a diversos temas como la literatura, la ciencia, la sociología o la economía, y estar redactados, inclusive, en inglés. Cada pregunta se basa en el texto que la precede y en ese texto se contiene toda la información necesaria para contestar las preguntas.

Inferencias lógicas y silogísticas.

Dentro de las preguntas probablemente encontrará algunas en las que ha de decidir cuál de varias afirmaciones propuestas como opciones es la que está implicada o se sigue de la base; o aquéllas en las que directamente se le pide completar un silogismo sencillo u otro más complejo. Si usted cree que hay varias respuestas correctas o verdaderas escoja aquella que crea es la mejor.

Temario.

A continuación se proporciona el temario en el que se basarán las preguntas del examen. No habrá preguntas de temas que no estén anotados. Es importante que el aspirante se enfoque en los temas a continuación descritos. El nivel de profundidad puede variar. Asimismo se anota la bibliografía que será utilizada en la elaboración de las preguntas del examen de admisión de esta Facultad.

Área 1. Física y Matemáticas.

Álgebra.

- Números reales.
- Lenguaje algebraico.
- Polinomios de una variable.

• Ecuaciones de primer y segundo grado.

Geometría y trigonometría.

• Ángulos en el plano.

• Triángulos.

• Polígonos y circunferencia.

• Funciones trigonométricas para ángulos agudos y de cualquier magnitud.

• Ley de senos y cosenos.

Geometría analítica.

• Sistemas de ejes coordenados.

• La línea recta.

• La circunferencia.

• La parábola.

Relaciones y funciones.

• Nociones de relación y de función.

• Clasificación y transformación de funciones.

• Funciones polinomiales.

• Funciones racionales.

• Funciones exponenciales y logarítmicas.

Física.

• Impacto de la física en la ciencia y tecnología.

• Metodología de la física.

• Movimiento

• Leyes de Newton, trabajo, potencia y energía.

• Hidráulica.

• Calor y temperatura.

• Electricidad, magnetismo y electromagnetismo.

Área 2. Biología.

Características distintivas de los seres vivos.

• Composición química de los seres vivos.

• Teorías sobre el origen de la vida.

• Biología de la célula.

• Metabolismo celular.

• Respiración celular.

• Diversidad biológica (virus, bacterias y eucariontes).

• Reproducción y herencia.

• Teorías de la evolución.

• Genética y evolución.

• Procesos biológicos en los animales (digestión, respiración, excreción, secreción, circulación, reproducción, desarrollo y sistema nervioso).

Área 3. Química.

La materia y la energía.

• Estructura atómica y tabla periódica.

• Enlace químico: modelos de enlace e interacciones intermoleculares.

• Reacción química.

• Estequiometría.

• Disoluciones.

• Compuestos de carbono.

• Macromoléculas.

Área 4. Inglés.

Comprensión general del idioma inglés escrito.

Se harán dos tipos de preguntas:

• Expresiones idiomáticas en inglés (interpretación de frases,

analogías, etc.).

• Preguntas de otras áreas formuladas en inglés.

Área 5. Español.

Lectura y redacción.

• Léxico y semántica.

• Textos personales.

• Textos expositivos.

• Textos persuasivos.

Advertencias Importantes.

El aspirante deberá registrarse para el examen presentando su credencial debidamente autorizada y acompañada por su certificado o constancia de terminación de estudios de bachillerato.

Su constancia de terminación de estudios deberá indicar que cursó y aprobó la totalidad de las materias del bachillerato correspondiente. No pueden presentar examen de admisión aquellos alumnos que adeuden materias o que estén por presentarlas en esos días. En algunos casos el examen podrá presentarse en forma condicionada en espera de que sea validado en los siguientes días previos a la inscripción.

El sábado 05 de julio de 2014 deberá llegar al lugar indicado a las 06:45 horas y formarse en la fila de entrada. El ingreso será a las 07:15 horas. Para acceder será indispensable que cada aspirante se identifique con su credencial con fotografía expedida por el Departamento de Admisiones. Esta deberá portar los sellos correspondientes a los exámenes, médico y psicométrico.

En la estancia encontrarás unas mamparas con las listas de todos los sustentantes, busque su nombre de acuerdo a su primer apellido y ahí se le indicará en cuál aula se encuentra su lugar. Si no lo encuentra, solicite información a un miembro del equipo de apoyo. El sitio que usted deberá ocupar tendrá su nombre y su número de credencial de una manera visible.

Antes de dar inicio al examen de conocimientos, se contestará una encuesta relacionada con aptitudes y habilidades de estudio. Este cuestionario no tendrá ningún valor en la calificación del aspirante.

El Examen de Conocimientos dará inicio a las 08:45 horas en punto y terminará a las 12:45 horas. El material del examen de conocimientos (cuestionario y hoja de respuestas), está numerado y foliado, y deberá devolverse completo. La no devolución o mutilación del material será motivo de anulación automática del examen.

Para el uso de las hojas de respuestas tome en cuenta las recomendaciones siguientes:

- Solamente use lápiz No. 2.
- Marque la respuesta que considere correcta. Cada pregunta tiene una sola respuesta. El marcar más de una casilla en el mismo número de pregunta se tomará como una respuesta mala.
- No deje preguntas en blanco. Cuando tenga una duda escoja la que considere mejor opción.

- Verifique que el número de la pregunta que haya contestado corresponda con el número en la hoja de respuestas.
- Verifique que la opción que se seleccionó corresponda con la letra del inciso que va a marcar en la hoja de respuestas.
- No debería usted de borrar en su hoja de respuesta, pero en caso de error, borre completamente. Cambie la opción sin maltratar la hoja.
- Para anotaciones, operaciones aritméticas o cálculos use el reverso de las hojas del examen, nunca la hoja de respuestas.

La hoja de respuestas será calificada por una lectora de marcas ópticas y computadora. Cerciórese de que la clave única que escriba en la hoja de respuestas corresponda con la suya. La clave única es la que aparece en su credencial. Cuide su hoja de respuestas. No la manche. No la arrugue.

Es importante cerciorase que la casilla de respuesta que usted haya elegido como respuesta esté completamente llena. Dejar un punto en blanco puede provocar que la lectora no registre la respuesta. En estos casos, el error será imputable al aspirante.

A continuación se muestran unos ejemplos de preguntas que comprende el Examen de Selección (Tradicional) de la Facultad de Medicina de la Universidad Autónoma de San Luis Potosí.

<u>Ejemplos del Examen de Admisión.</u>

1. Usted tiene un examen a las 09:00 hrs. y quiere saber a qué hora entre las 08:00 y las 9:00, el minutero dista exactamente del horario 10 divisiones.

a) A las 8 con 20 y ¾ min.

b) A las 8 con 25 y 7/10 min.

c) A las 8 con 30 y 4/5 min.

d) A las 8 con 32 y 8/11 min.

e) A las 8 con 40 y 3/5 min.

2. ¿Cuántas señales distintas pueden hacerse con 9 banderas, izando 3 cada vez, sin importar el orden?

a) 27.

b) 81.

c) 243.

d) 504.

e) 729.

3. ¿Cuál es el 6° término de una progresión aritmética de 11 términos, si el 1er término es –2 y el último –52?

a) – 21.

b) – 23.

c) – 25.

d) – 27.

e) – 29.

4. Encuentre el par de valores de x y y que satisfaga el siguiente grupo de ecuaciones.

$2x + y = -1$, $x - 2y = -13$, $3x - 2y = -19$

a) $x = -1$, $y = 3$

b) $x = 1$, $y = 4$

c) $x = -2$, $y = -3$

d) $x = -3$, $y = 5$

e) $x = 2$, $y = 3$

5. La población de Rioverde, SLP ha aumentado en progresión geométrica de 59,049 habitantes que era en el año 2002 a 100,000 habitantes en 2007. ¿Cuál es la razón de crecimiento por año?

a) 4/3

b) 5/6

c) 6/5

d) 8/7

e) 10/9

6. La hemoglobina adquiere su poder de transportar oxígeno, gracias a la presencia en su molécula de:

a) Carbono.

b) Cobalto.

c) Hidrógeno.

d) Hierro.

e) Magnesio.

7. La versatilidad del carbono para formar compuestos orgánicos, es debida a que:

a) Forma compuestos funcionales carbonilos no–polares.

b) Los enlaces que forma se encuentran en el mismo plano geométrico.

c) Puede tener cuatro enlaces covalentes con compuestos distintos.

d) Se encuentra abundantemente en nuestro planeta.

e) Sus propiedades hidrofóbicas.

8. A mosquito that is not affected by a disease, but transmits it, is a:

a) Host.

b) Parasit.

c) Pathogen.

d) Transmisor.

e) Vector.

9. Seleccione el trío de palabras que completan la siguiente frase: De acuerdo a la Teoría Sintética de la Evolución, la ______________ genética es un fenómeno que puede ocurrir en todas las especies. En los humanos puede ser la causa de enfermedades, pero también puede relacionarse con cambios a largo plazo que gracias a la ___________ hacen posible una mejor ____________ de los individuos.

a) Amplificación, civilización, satisfacción.

b) Clonación, especiación, reproducción.

c) Diferenciación, convergencia, salud.

d) Morfogénesis, supervivencia, respuesta.

e) Mutación, selección, adaptación.

10. ¿Cuáles son elementos de carácter semimetálico?

a) Boro y silicio.

b) Bromo y argón.

c) Fósforo y azufre.

d) Potasio y calcio.

e) Titanio y magnesio.

11. Las proteínas:

a) Están formadas por 200 aminoácidos distintos.

b) Son básicamente energéticos.

c) Son básicamente estructurales.

d) Son cadenas ramificadas de α– aminoácidos unidos por

enlace peptídico.

e) Son insolubles cuando presentan estructura fibrilar.

12. Una reacción química es de primer orden:

a) Si al graficar la concentración del reactante frente al tiempo obtenemos una recta.

b) Si la concentración de reactante inicial no afecta la velocidad con la que ocurre la reacción.

c) Si la reacción requiere de dos reactantes.

d) Si la velocidad de la reacción varía linealmente con la concentración de dos o más reactantes.

e) Si la velocidad de la reacción varía linealmente con la concentración de un sólo reactante.

13. What is the meaning of the expression “Mind your own business”?

a) Debes de trabajar en los negocios.

b) Debes de ser propietario del negocio.

c) No te metas en lo que no te importa.

d) Piensa en tu propio autobús.

e) Ten en mente hacer un buen negocio.

14. Select the sentence that is correct according to the following statements and is written correctly:
John and Bob were born the same day. They share their father and mother.

a) John and Bob is brother.

b) John and Bob are cousins.

c) John and Bob are sister’s.

d) John and Bob are twins.

e) John and Bob share their bird–day

15. John was riding his bicycle last evening. He saw Ann crossing Main Street and waved at her. A few moments later, Ann was run over by a car. Although he was not responsible of the accident, he could not help feeling guilty.

a) Ann and John crashed last evening.

b) Ann was driving a car this morning.

c) Ann was crossing in the waves.

d) John feels he is to blame for the accident.

e) John had a bicycle accident.

16. Fill in the blanks.

John ________ a marathon yesterday morning.

a) Ran.

b) Run.

c) Runed.

d) Runned.

e) Runs.

17. Select the correct association:

a) Baby, man, mans.

b) Cub, wolf, wolves.

c) Fish, whale, whales.

d) Kitten, dog, dogs.

e) Puppy, cat, cats.

18. Bird is to airplane as:

a) Balloon is to videogame.

b) Helicopter is to train.

c) Horse is to car.

d) Parrot is to television.

e) Whale is to shark.

19. ¿Cuál de los siguientes pares de palabras comparte la raíz, pero son antónimos?

a) Malaria, Buenos Aires.

b) Microbio, macro–organismo.

c) Padecer, acostar.

d) Palabra, parábola.

e) Redondo, recto.

20. Las palabras “benéfico”, “confiar” y “efectivo”, tienen en común una raíz que significa:

a) Bueno.

b) Compañía.

c) Fe.

d) Fuerza.

e) Hacer.

21. La función fática de la lengua se muestra en este enunciado:

a) Bueno….Sí….Dime….Ajá…Adiós.

b) El E.P.R. saluda el debate amplio con el gobierno.

c) El verbo denota acción.

d) ¡Qué bueno es volver a México!

e) ¿Quieres ir a bailar?

22. La función metalinguística de la lengua se ejemplifica en lo siguiente:

a) ¡Entonces qué! ¿Qué onda eh?

b) Ha sido un placer platicar contigo.

c) La oración simple es un enunciado predicativo bimembre

que es sintácticamente autónomo.

d) Sus pupilas eran de fuego…

e) Volverán los paros magisteriales a Oaxaca.

23. Señale el enunciado con ortografía correcta:

a) Entonces se abrió un compas de espera.

b) Diles que no me máten.

c) ¿Te puedo ofrecer un té?

d) Yo sé que sé.

e) Yo no me refería a este.

24. De las siguientes, la característica más importante de un texto escrito es:

a) El tema.

b) La congruencia.

c) La intención.

d) La ortografía.

e) La presentación.

25. “Muerte sin Fin”, fue escrita por:

a) Carlos Fuentes.

b) José Gorostiza.

c) Manuel José Othón.

d) Octavio Paz.

e) Xavier Villaurrutia.

TOMA DE FOTOGRAFÍA Y EXAMEN PSICOMÉTRICO

Acudirás el día y hora marcados en tu Pase Examen Psicométrico y Fotografía, al Centro de Salud Universitario (ver ubicación en la página 110).

Es indispensable llevar lo siguiente:

1. Material:

- Lápiz del No. 2 ó 2 1⁄2.
- Una goma suave.

2. Documentos:

- Pase Examen Psicométrico y Fotografía.
- Cuestionario de Contexto (CENEVAL), debidamente contestado, ver página 17 para el correcto llenado de tu Clave Única en el mismo.

EVALUACIONES

Como parte de los requisitos del Proceso de Admisión a continuación se describen los distintos exámenes que deberás presentar:

• Examen Psicométrico

Son evaluaciones psicológicas estandarizadas que miden las aptitudes básicas para el estudio, esta evaluación se divide en tres pruebas:

a) Razonamiento verbal: Mide la capacidad para expresarse utilizando sinónimos y antónimos.
b) Retención y comprensión: Mide la habilidad para comprender y retener lo expuesto en el salón de clases.

c) Razonamiento abstracto: Mide la capacidad para razonar en forma lógica e inmediata ante problemas cotidianos.

El Examen Psicométrico y la toma de fotografía se aplicarán en el Centro de Salud Universitario ubicado dentro de la Zona Universitaria en Av. Manuel Nava Martínez s/n. La fecha y hora en que deberás acudir están impresos en el Pase Examen Psicométrico y Fotografía, de no ser puntual con tu asistencia perderás el derecho a continuar con los trámites para el Examen de Admisión. No olvides llevar tu cuestionario de contexto CENEVAL debidamente contestado.

• Exámenes de Conocimientos:

a) Uno elaborado por la Comisión de Admisión de la Entidad Académica a donde deseas ingresar.

b) El EXANI-II Admisión diseñado por el Centro Nacional de Evaluación para la Educación Superior, A.C. (CENEVAL).

Se aplicarán el día 11 de julio de 2015 en la Entidad Académica de tu elección, por lo que deberás ubicar oportunamente el aula donde los vas a presentar (ver páginas 105-113 de este instructivo). Es importante que llegues cuando menos media hora antes de cada evaluación, considera los puntos de la página 24.

<u>Examen elaborado por las entidades académicas a las 08:00 horas.</u>

Evalúa los conocimientos, las destrezas y las habilidades requeridas de los aspirantes a ingresar de acuerdo con el perfil del alumno pretendido en las licenciaturas. Para este examen podrás consultar la guía temática en la siguiente dirección: www.admisiones.uaslp.mx o en su caso puedes solicitarla en la entidad académica donde se ofrece la carrera que solicitaste.

<u>EXANI-II Admisión del CENEVAL a las 16:00 horas.</u>

Herramienta de apoyo en los procesos de admisión de candidatos a ingresar a un programa de educación superior. Explora competencias genéricas predictivas en las áreas de Pensamiento Matemático, Pensamiento Analítico y Competencias Comunicativas del Español. Su propósito es establecer el nivel de potencialidad de un individuo para lograr nuevos aprendizajes, por lo que todo sustentante debe responderlo. Ofrece a las instituciones usuarias información útil para la toma de decisiones sobre la admisión de los aspirantes. Para más información consulta la guía en el material que recibiste.

ENCUESTA DE OPINIÓN SOBRE EL PROCESO DE ADMISIÓN

Esta encuesta tiene como objetivo conocer su opinión acerca del proceso de admisión que llevó a cabo para ingresar a la licenciatura en médico cirujano; por favor conteste con respecto a lo que usted sabía y percibió cuando se presentó a estos exámenes. Sabemos que esta evaluación se llevó a cabo hace mucho tiempo, y posiblemente ya no recuerde muchos detalles; aún así, por favor intente responder lo mejor posible. Si realmente no se acuerda de lo que se le pregunta, puede marcar la opción correspondiente.

Los datos que se obtengan serán utilizados por la M en C Blanca Ariadna Carrillo Avalos en un estudio de validez relacionado con dicho proceso de admisión realizado en los años 2013 y 2014.

Si tiene alguna duda al respecto o desea conocer los resultados, por favor envíe un correo a ariadna.carrillo@uaslp.mx.

Sus respuestas son muy valiosas, ¡gracias por contestar esta encuesta!

I. Datos demográficos

- ¿Cuál es su edad? (Variable continua, numérica)
- ¿Cuántos años tenía cuando hizo el examen de admisión para ingresar a la licenciatura en médico cirujano? (Variable continua, numérica)
- ¿Cuál es su sexo? (Variable dicotómica: Hombre/mujer)

II. Momento previo a la evaluación.

Por favor indique su nivel de satisfacción acerca de los rubros siguientes, seleccionando una opción:

Rubro	Alto nivel de satisfacción		Bajo nivel de satisfacción			
	Muy satisfecho	Satisfecho	Insatisfecho	Muy insatisfecho	No me acuerdo	No aplica
1. El tiempo de emisión y vigencia de la convocatoria para el proceso de admisión.						
2. Los medios de difusión de la convocatoria del proceso de admisión.						
3. La utilidad del Instructivo para Aspirantes de Nuevo Ingreso para conocer el proceso en general						
4. La claridad con la que la UASLP resolvió las dudas sobre la convocatoria.						
5. La atención de la autoridad educativa para realizar el registro, recepción						

y revisión de la documentación.						
6. La utilidad del Instructivo para Aspirantes de Nuevo Ingreso para conocer cómo se va a evaluar.						
7. La utilidad de Guía temática del Examen de Conocimientos de la Facultad de Medicina para conocer qué se va a evaluar.						
8. El tiempo con el que contó para tener acceso a la bibliografía y guía de estudios.						
9. La relación de la guía de estudios y la bibliografía, con el contenido de los exámenes.						

III. Momento de aplicación de los exámenes de admisión.

Por favor indique su nivel de satisfacción acerca de los rubros siguientes, seleccionando una opción:

Rubro	Alto nivel de satisfacción		Bajo nivel de satisfacción		
	Muy satisfecho	Satisfecho	Insatisfecho	Muy insatisfecho	No me acuerdo
10. Los aspectos que se evalúan en los exámenes.					
11. La precisión de la redacción de los planteamientos en las preguntas.					
12. La cantidad total de preguntas del examen.					
13. La extensión de las preguntas del examen.					
14. La contextualización de las preguntas del examen.					
15. La localización de la sede.					
16. La accesibilidad de la sede.					
17. La comodidad del mobiliario de las aulas.					

18. La iluminación y la temperatura de las aulas.					
19. La precisión de las indicaciones brindadas por el aplicador durante el examen.					
20. La atención del aplicador ante las dudas de los sustentantes.					
21. El trato brindado a los sustentantes por el aplicador.					

IV. Conocimiento general del proceso .

Por favor seleccione una opción con referencia a qué tanto conoce acerca del rubro que se evalúa:

Rubro	Alto nivel de conocimiento	Bajo nivel de conocimiento
	Sí	No
22. Sabía el número de lugares que se concursan antes del examen.		
23. Sabía cómo se califican los exámenes antes de presentarlos.		
24. Cómo se conforman las listas de aceptados.		

DIMENSIONES

Las dimensiones abarcan los tres momentos: el momento previo a la evaluación (para conocer el grado de satisfacción de los encuestados en cuanto a los mecanismos e instrumentos con los que contaron para presentar los exámenes de admisión), el momento de aplicación de los exámenes de admisión (para conocer la percepción de los sustentantes en cuanto a la pertinencia de las características de los exámenes y las sedes) y el momento posterior a la evaluación (para conocer si los encuestados conocen cómo se califican los exámenes y cómo se distribuyen las los lugares para estudiar).

Momento previo a la evaluación.		
Dimensión	Preguntas	
Convocatoria	1.	El tiempo de emisión y vigencia de la convocatoria para el proceso de admisión.
	2.	Los medios de difusión de la convocatoria del proceso de admisión.
Atención de la UASLP	3.	La utilidad del Instructivo para Aspirantes de Nuevo Ingreso para conocer el proceso en general
	4.	La claridad con la que la UASLP resolvió las dudas sobre la convocatoria.
	5.	La atención de la autoridad educativa para realizar el registro, recepción y revisión de la documentación.
Utilidad de los documentos referentes a la evaluación	6.	La utilidad del Instructivo para Aspirantes de Nuevo Ingreso para conocer cómo se va a evaluar.
	7.	La utilidad de Guía temática del Examen de Conocimientos de la Facultad de Medicina para conocer qué se va a evaluar.
Bibliografía y guía de estudios	8.	El tiempo con el que contó para tener acceso a la bibliografía y guía de estudios.

	9.	La relación de la guía de estudios y la bibliografía, con el contenido de los exámenes.

Momento de aplicación de los exámenes de admisión.		
Dimensión	Preguntas	
Exámenes	10.	Los aspectos que se evalúan en los exámenes.
	11.	La precisión de la redacción de los planteamientos en las preguntas.
	12.	La cantidad total de preguntas del examen.
	13.	La extensión de las preguntas del examen.
	14.	La contextualización de las preguntas del examen.
Sede	15.	La localización de la sede.
	16.	La accesibilidad de la sede.
	17.	La comodidad del mobiliario de las aulas.
	18.	La iluminación y la temperatura de las aulas.
Aplicadores	19.	La precisión de las indicaciones brindadas por el aplicador durante el examen.
	20.	La atención del aplicador ante las dudas de los sustentantes
	21.	El trato brindado a los sustentantes por el aplicador.

Conocimiento general del proceso.		
Dimensión	Preguntas	
Información de resultados del proceso de admisión	22.	Sabía el número de lugares que se concursan antes del examen.

	23.	Sabía cómo se califican los exámenes antes de presentarlos.
	24.	Cómo se conforman las listas de aceptados.

Las escalas de Likert correspondientes fueron:

Momento previo a la evaluación (primera fase) y Momento de aplicación de los instrumentos (segunda fase)	
1. Muy satisfecho	Alto nivel de satisfacción
2. Satisfecho	
3. Insatisfecho	Bajo nivel de satisfacción
4. Muy insatisfecho	
Momento posterior a la evaluación (tercera fase)	
1. Sí	Nivel de conocimiento adecuado
2. No	Nivel de conocimiento inadecuado

10. ANEXO CON ARTÍCULOS DERIVADOS DE LA TESIS EN EXTENSO

El concepto moderno de validez y su uso en educación médica

Blanca Ariadna Carrillo Avalos[a,*], Melchor Sánchez Mendiola[b], Iwin Leenen[c]

Facultad de Medicina

Resumen

Para realizar inferencias apropiadas con base en los resultados obtenidos de las evaluaciones del aprendizaje en ciencias de la salud, es fundamental aportar evidencia de validez y así proveer el fundamento y la justificación de las decisiones que se tomen a partir de las evaluaciones. El concepto de validez es el más importante en evaluación educativa, pues aplica para todo tipo de uso de instrumentos de evaluación del aprendizaje, tanto sumativos como diagnósticos y formativos. En las últimas décadas han surgido nuevos marcos de referencia que modifican y enriquecen el concepto tradicional de validez. En este trabajo se exploran las perspectivas de Messick y Kane. Con respecto al primero se describen las fuentes de evidencia de validez y cómo obtenerlas, mientras que con relación al segundo se explican los pasos para llevar a cabo un argumento de usos que justifique las interpretaciones de los resultados de los exámenes. Con este panorama se presenta una perspectiva moderna de aproximación a la validez en evaluación educativa, de utilidad para los educadores en ciencias de la salud.

Palabras clave: *Validez; evaluación del aprendizaje; educación médica; México.*

[a]Departamento de Ciencias Morfológicas, Facultad de Medicina, Universidad Autónoma de San Luis Potosí, S. L. P., México.
[b]División de Estudios de Posgrado, Facultad de Medicina, Universidad Nacional Autónoma de México, Cd. Mx., México.
[c]División de Estudios de Posgrado, Facultad de Psicología, Universidad Nacional Autónoma de México, Cd. Mx., México.
Recibido: 16-octubre-2019. Aceptado: 2-diciembre-2019.
*Autora para correspondencia: Blanca Ariadna Carrillo Avalos. Av. Venustiano Carranza 2405, Col. Los Filtros, San Luis Potosí, S. L. P., México. CP 78210. Teléfono: 44 4826 2345, ext. 6635.
Correo electrónico: bariadna@gmail.com
La revisión por pares es responsabilidad de la Universidad Nacional Autónoma de México.

https://doi.org/10.22201/facmed.20075057e.2020.33.19216

Current concepts of validity and its use in medical education

Abstract

In order to articulate appropriate inferences based on the scores obtained from learning assessments in the health sciences, the collection of validity evidence to support decisions made on the basis of these assessments is of central importance. The concept of validity is key in educational assessment, since it is used in all kinds of learning evaluation strategies: summative, diagnostic, and formative. In the last decades, new frameworks which modify and enhance the traditional concept of validity have emerged. In this paper, we explore the perspectives of Messick and Kane. Regarding the first one, we describe the sources of validity evidence and how to obtain them; and in regard to Kane's arguments, we explain the steps needed to state an argument of use that justifies the interpretations of the scores obtained from the assessments. This overview describes the current perspective to approach validity in educative assessment, useful for health sciences educators.

Keywords: Validity; learning assessment; medical education; Mexico.

INTRODUCCIÓN

En una ocasión aplicamos un examen final de una materia de ciencias básicas que solo aprobó un pequeño porcentaje de alumnos; algunas personas comentaron que el examen no era válido, por lo que debíamos repetirlo. ¿Cómo podríamos comprobarlo? Primero, decir que un examen es válido o no, es un error de concepto frecuente que es importante despejar para contar con elementos que permitan elaborar y aplicar los exámenes de alto y bajo impacto, así como contar con resultados útiles.[1] Por otro lado, son numerosas las publicaciones que hablan acerca de aspectos de validez en evaluación en educación médica (como validez predictiva o validez de las preguntas del examen), cuyo análisis no menciona explícitamente el concepto actual de validez, y cómo se debe evaluar e interpretar.[2-4]

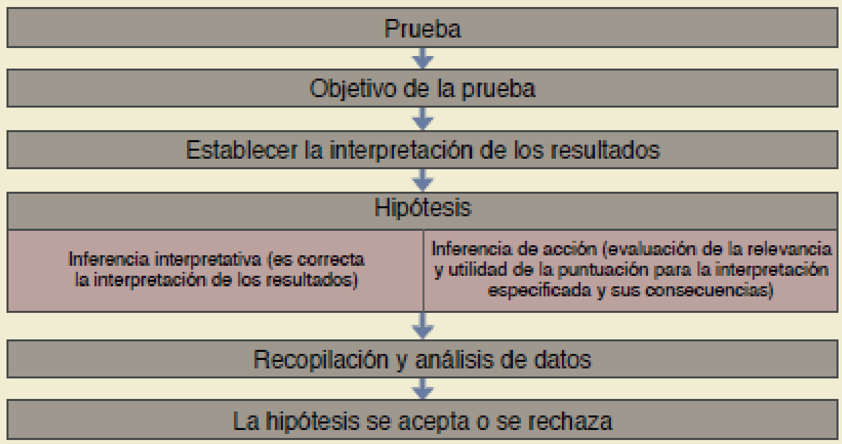
Al desarrollar y evaluar los exámenes, la validez, como el grado con que la evidencia empírica y las razones teóricas apoyan o refutan lo apropiado o adecuado de la interpretación o el uso que se da a los resultados de una evaluación, es la consideración más importante que debe hacerse.[5,6] Por otro lado, la característica o concepto que se mide en una evaluación específica es un constructo latente, y debe especificarse cuál es la interpretación que se va a dar acerca de éste con base en las puntuaciones obtenidas en la prueba. De esta manera son las inferencias que se hacen acerca de un constructo con base en la puntuación de una evaluación las que requieren evidencia de validez, mas no la evaluación por sí misma. Además, el análisis de la validez será en cuestión de grado y no de un enfoque dicotómico que certifique su existencia o inexistencia.[6-8]

El concepto de validez ha evolucionado desde la primera mitad del siglo XX, cuando se consideraba que un examen era válido cuando existía correlación con lo que pretendía medir.[9, citado por 10] Posteriormente se elaboró la teoría tradicional que identificaba tres tipos de validez: de contenido, de constructo y de criterio, esta última dividiéndose en validez concurrente y validez predictiva.[11] En las últimas décadas han surgido nuevos marcos de referencia que modifican y enriquecen el concepto tradicional de validez, de forma que actualmente existen dos de ellos que son considerados los más prominentes, por lo cual se estima necesario tomarlos en cuenta para evaluar la validez en evaluación educativa: el de Samuel Messick y el de Michael Kane. En este artículo presentamos una introducción general a estos marcos para la validez y sugerimos algunas ideas para su integración en educación en ciencias de la salud.

MARCO DE REFERENCIA DE MESSICK

El marco de referencia de Messick[8] considera que la validez de constructo es el único tipo que existe ya que las evaluaciones tienen como objetivo medir

Figura 1. Resumen del marco de referencia de Messick



Fuente: Elaboración propia.

constructos, es decir, las características o atributos de las personas que no pueden ser observados directamente (son latentes) y que se miden a través del examen diseñado.[6,11] Por ejemplo, el desempeño académico de un estudiante de medicina es una característica latente, por lo que se infiere a través de sus respuestas en los exámenes de cada asignatura, conformando un constructo susceptible de estudio.[7,12] A la luz de lo anterior, cualquier estudio de validez en el marco de Messick busca aportar, de forma directa o indirecta, evidencia para el constructo que subyace la evaluación.

Messick menciona que un análisis de la validez siempre parte de una hipótesis o inferencia acerca de la interpretación o el uso que se pretende dar a los resultados de la prueba. Posteriormente se deben recopilar y analizar los datos, enlazarlos a un marco teórico específico, y luego determinar la validez o invalidez de la hipótesis declarada para un momento particular en el tiempo, para una población específica **(figura 1)**.[7,8]

Así, este marco de referencia se enfoca en cinco fuentes de evidencia de validez. No es indispensable buscar todas estas fuentes en todos los análisis de resultados de exámenes. Las fuentes de evidencia de validez que se requieren dependen de los objetivos de la prueba y de sus consecuencias, entre otros aspectos[13], ya que éstas sirven para sustentar la interpretación que se haya determinado para la prueba previamente.[6,7] Por ejemplo, en el caso de pruebas de altas consecuencias como el examen de admisión a la escuela de medicina o el examen de titulación de enfermería, podría necesitarse mayor evidencia de validez que para una prueba utilizada con fines formativos.[6] A continuación, se discuten las cinco fuentes de evidencia de validez en el marco de Messick y algunos ejemplos de cómo documentarlas.

FUENTES DE EVIDENCIA DE VALIDEZ

1. Evidencia basada en el contenido de la prueba

El contenido de la prueba se refiere a los temas que evalúa; por ejemplo, en el caso de un examen de admisión abarcaría toda la información cuyo dominio debe demostrar un alumno antes de ingresar al nivel educativo que pretende. Este contenido también depende de las inferencias que se vayan a hacer a partir de las puntuaciones obtenidas en la prueba.[6]

Esta evidencia se puede obtener "a partir del análisis de la relación entre el contenido de la prueba y el constructo que pretende medir"[6], por ejemplo, se analiza la representatividad de la tabla de especificaciones con respecto al dominio del conocimiento que se examina, las especificaciones del examen, representatividad de los ítems con respecto al dominio del conocimiento examinado, coincidencia del contenido de los ítems con las especificaciones del examen y relación lógica o empírica del contenido evaluado con el dominio del conocimiento que se examina.

Para documentar esta fuente de evidencia también se evalúan procesos de alineación, que evalúan la correspondencia entre el contenido de la prueba y los resultados de aprendizaje del alumno, es decir, qué tanto se representa el dominio del conocimiento en la prueba con base en criterios como la complejidad cognitiva, el currículo y los métodos instruccionales. Esto se puede lograr de diferentes formas, una de ellas consiste en que expertos califiquen la semejanza entre pares de ítems en términos de las habilidades y el conocimiento evaluados por medio de escalas tipo Likert.[6,14]

2. Evidencia basada en los procesos de respuesta

En los *Standards for Educational and Psychological Testing*[6] esta fuente de evidencia se refiere a que se puede comprobar la relación entre el constructo que se pretende medir y los procesos cognitivos que intervienen en la resolución de la tarea o los ítems de la prueba. Esta evidencia puede obtenerse por medio de entrevistas cognitivas, herramientas que permiten conocer la comprensión de términos clave, así como entender el razonamiento utilizado para llegar a la respuesta correcta y así evitar falsos positivos (llegar a la respuesta correcta después de un razonamiento erróneo), de manera que el sustentante realmente esté aplicando lo necesario para resolver el problema propuesto y que así logre obtener resultados favorables en otros contextos.[15] También existen modelos matemáticos que relacionan la dificultad de los ítems o el tiempo de respuesta con los procesos cognitivos hipotéticos, mismos que permiten aportar evidencia de este tipo.[16]

Cabe mencionar que Downing[7] incluye para esta fuente de evidencia de validez también un análisis de aspectos asociados con la administración del examen, por ejemplo, la familiaridad de los sustentantes con el formato del examen, que sepan llenar adecuadamente las hojas de respuesta, la claridad de las instrucciones, etc. Sin embargo, es importante aclarar que esta interpretación de Downing[7] se encuentra algo desalineada con la visión del mismo Messick y de los psicómetras prominentes en esta área, como Kane y Embretson, entre otros.

3. Evidencia basada en la estructura interna

La estructura interna es el grado en que las relaciones de los ítems de la prueba están alineadas con la teoría detrás del constructo que se mide.[6] Evidencia de este tipo se puede obtener analizando las características psicométricas de las preguntas del examen, las características de la escala, y el modelo psicométrico que se utilizó para establecer la escala y calificar el examen.[7] El análisis de datos para obtener evidencia de validez de este tipo suele recurrir a análisis factorial (exploratorio o confirmatorio) o análisis en el marco de la teoría de respuesta al ítem; ambos permiten investigar las relaciones entre las respuestas en los ítems y el constructo subyacente a la prueba.[17,18]

El análisis de la estructura interna también atañe a la confiabilidad; en general, es importante documentar que las puntuaciones pudieran ser reproducibles si se aplicara nuevamente la prueba. De lo contrario, la interpretación de los resultados de este examen se puede ver comprometida.[7,18,19]

4. Evidencia basada en las relaciones con otras variables

Este tipo de evidencia se basa en el análisis de la relación de los resultados de la prueba con los resultados de otras pruebas que midan o no el mismo constructo u otras variables externas a la prueba. Proporciona información acerca del grado en que estas relaciones son coherentes con el constructo en el que se basan las interpretaciones de los resultados de la prueba.[6] Se puede buscar evidencia por esta fuente con base en relaciones convergentes (cuando se evalúan las relaciones entre las puntuaciones y medidas del mismo constructo) y/o discriminantes (cuando se evalúan las relaciones entre las puntuaciones y medidas de constructos diferentes).[7] Una manera de investigar ambos tipos de relaciones es a través de una matriz multirasgo-multimétodo, que es una matriz de correlaciones entre distintas pruebas que, en conjunto, miden dos o más constructos a través de dos o más métodos.[20]

Se consideran dos diseños para la evidencia de validez de este tipo:[6]

- Estudio predictivo. Evalúa el grado de la relación entre las puntuaciones de la prueba y las puntuaciones del criterio que se obtiene en un tiempo posterior. Por ejemplo, estudios que evalúan exámenes de admisión académica y que investigan la relación con el desempeño académico subsecuente.
- Estudio concurrente. Evalúa el grado de la relación entre las puntuaciones de la prueba y las puntuaciones del criterio que se obtiene al mismo tiempo. En este tipo de estudios se evitan los cambios temporales y pueden ser útiles para buscar formas alternas de medición del constructo en cuestión, por ejemplo, analizar la correlación de los puntajes de una variante corta de una prueba con los de una variante original más larga, que mide el mismo constructo, pero ya cuenta con evidencia de validez.

La generalización de los resultados que aporta el estudio de esta fuente de validez depende de que las condiciones en la nueva situación sean iguales a las presentes en el análisis original. Los resúmenes estadísticos de los estudios de validación anteriores en condiciones semejantes, como en un meta-análisis, pueden ser útiles para estimar las nuevas relaciones, pero dependen del tamaño de la muestra y de la cantidad de estudios realizados a lo largo del tiempo.[6,21]

5. Evidencia basada en las consecuencias de la prueba

Generalmente, la interpretación y el uso de los resultados de la prueba tienen impacto o consecuencia de diferentes grados o tipos sobre los sustentantes. Por ejemplo, en el caso de las evaluaciones de admisión para una licenciatura, esta evidencia lleva a reflexionar sobre las posibles equivocaciones en la interpretación de los resultados de la prueba con respecto a falsos positivos y falsos negativos, así como tomar en cuenta estas consecuencias negativas para que se lleve a cabo una evaluación de qué tan grave es un falso positivo y qué tan grave un falso negativo y que se considere al ponderar las consecuencias diferenciales de ambos tipos de errores.

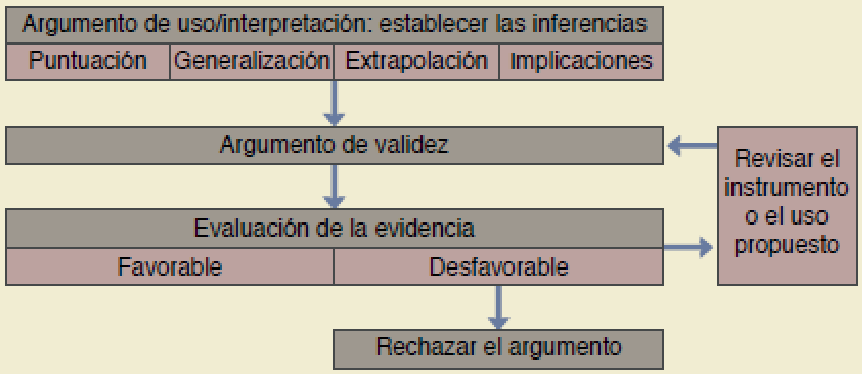
Esta fuente de validez requiere analizar el impacto de los resultados de la prueba en los estudiantes y la sociedad, el balance entre las consecuencias positivas y las negativas involuntarias, lo razonable del punto de corte de aprobado/reprobado o admitido/no admitido, las consecuencias de aprobar o reprobar, de los falsos positivos y falsos negativos, y las consecuencias institucionales y del estudiante.[6,7] Este análisis puede realizarse por medio de entrevistas y grupos focales, así como la teoría de acción para identificar los componentes críticos de los programas académicos y sus puntos de impacto.[22]

Como ejemplo, considérese el Examen Nacional para Aspirantes a Residencias Médicas, que "es un instrumento de medición de conocimientos en el contexto del ejercicio de la medicina general, objetivo y consensuado, que constituye la primera etapa del proceso para ingresar al Sistema Nacional de Residencias Médicas."[23] A pesar del objetivo establecido por los desarrolladores de esta evaluación, algunas instituciones utilizan sus resultados como una forma de determinar cual es "la mejor escuela de medicina" en nuestro país, produciendo consecuencias no intencionadas e indeseables. Analizar estas consecuencias y hacer lo necesario para evitarlas en la medida de lo posible constituye un ejemplo de este tipo de evidencia de validez.

MARCO DE REFERENCIA DE KANE

Kane consideró que, aunque la visión de Messick acerca de la validez de constructo es importante, no es fácil de evaluar, ya que no provee de guías para iniciar el procedimiento, y no es muy práctica[24]; por ello desarrolló su propio marco de referencia que se enfoca en el proceso de recolección de evidencia de validez mediante cuatro inferencias para desarrollar un argumento de validez.[25] El planear un examen considerando las fuentes de validez marca el camino para partir de la evaluación de una sola observación (inferencia de puntuación) hacia la puntuación general del examen (generalización) y de ahí a establecer las implicaciones de la puntuación en el desempeño en la vida real (extrapolación), llegando finalmente a la interpretación de esta información y a la toma de decisiones (implicaciones).[26] Una ventaja de este

Figura 2. Marco de referencia de Kane



Basado en Cook et al., 2015.

acercamiento a la validez es que es factible para quienes no poseen experiencia amplia en psicometría, además de que propone pasos muy claros.[27]

En general, los pasos que propone son dos: el primero es establecer el argumento de uso o interpretación (AUI) y el segundo es desarrollar el argumento de validez; este último es facilitado al considerar los cuatro tipos de inferencias (**figura 2**).

1. Establecer el argumento de uso o interpretación (AUI)

La interpretación de los resultados de la prueba implica explicar el significado de la puntuación, mientras que el uso de las puntuaciones se refiere a las decisiones que se toman con base en los resultados de la prueba. Kane considera que ambos términos (interpretación y usos) incluyen todas las suposiciones que se pueden hacer al respecto de las puntuaciones de una prueba, por lo que se debe establecer la validez de la interpretación o el uso de las puntuaciones en términos de lo creíble y apropiado que tengan en un punto del tiempo. Tener claro lo que se quiere evaluar permite elaborar un plan de evaluación preciso, por lo que el AUI puede conformar una red de inferencias y suposiciones que van desde el desempeño en las pruebas hasta las conclusiones que se obtienen, y las decisiones que se toman con base en estas conclusiones.[28,29]

Kane sugiere las siguientes inferencias que se encuentran presentes en la mayoría de los AUI, aunque también menciona que no es indispensable evaluarlas todas:[29,30]

- Inferencia de puntuación. Es la suposición acerca de lo apropiado de los criterios de la puntuación y las reglas para combinar las puntuaciones, mismas que conforman un estimado acerca de un atributo y son la base para la toma de decisiones.
- Inferencia de generalización. Si la prueba contiene una muestra de posibles escenarios o posibles ítems, esta inferencia supone que el sustentante va a obtener puntuaciones semejantes al presentar otra prueba con ítems diferentes extraídos del mismo universo de ítems, de manera que las puntuaciones observadas son representativas de todo el universo de puntuaciones posibles. Esta inferencia puede utilizar evidencia empírica en el marco de la teoría de la generalizabilidad,[31] debido a la importancia de puntuaciones reproducibles y generalizables.
- Inferencia de extrapolación. Por medio de este tipo de suposiciones se podría extender la interpretación a otros dominios de desempeño y predecir cuál será el resultado del sustentante en contextos diferentes al del examen. Un ejemplo de este tipo de inferencia sería que si la puntua-

ción observada tiene un valor particular (examen de admisión), entonces se espera un valor específico del criterio (desempeño académico durante la carrera); las herramientas analíticas para evaluar inferencias de este tipo suelen utilizar modelos de regresión.

- Inferencia de implicaciones. Se refiere al impacto que tiene la interpretación de los resultados de la prueba en el sustentante, en su familia y en la sociedad. Kane considera que, si las consecuencias de la interpretación de los resultados de una prueba son negativas, entonces la prueba no debería utilizarse.

2. Establecer el argumento de validez

Una vez que se han establecido las inferencias concernientes a las puntuaciones de la prueba en cuestión, se deben evaluar las garantías o métodos de comprobación de estas inferencias. Por ejemplo, la garantía de una inferencia de extrapolación con interés predictivo sería una ecuación de regresión, cuyo soporte estaría conformado por un análisis empírico acerca de la relación entre la puntuación de la prueba y los resultados del criterio seleccionado. El calificador de la garantía es el término que expresa la fuerza de la relación que se está analizando, y puede expresarse de manera numérica y con palabras (como coeficientes de correlación).[29]

Con estas consideraciones, el primer paso será realizar un análisis conceptual del AUI y verificar que sea coherente y que todas las inferencias importantes se encuentren presentes. Posteriormente, se deberán evaluar las inferencias presentadas. En la **tabla 1** se resumen las inferencias que propone Kane, así como los procedimientos que se deben definir y la manera de evaluarlos.

El argumento de validez debe ser claro para poder ser reproducible por cualquier investigador, conteniendo detalles específicos y presentando información coherente, de manera que las conclusiones sean lógicas. Por lo anterior, el argumento también debe estar completo y ser verificable.[32]

CONCLUSIONES

Se han revisado brevemente los marcos de referencia modernos y prominentes de validez a considerar cuando se interpretan y utilizan los resultados de las pruebas evaluativas en medicina; esta información es importante ya que su conocimiento y aplicación permitirá iniciar la elaboración de evaluaciones mejor planeadas y con objetivos más claros, además de que los resultados serán realmente útiles y su interpretación tendrá mayor grado de validez. No todas las fuentes de evidencia de validez se encontrarán presentes en todos los exámenes; sin embargo, son indispensables las que sustenten la interpretación descrita al inicio de la planeación.

Por otro lado, mientras que el marco de referencia de Messick deja claras las fuentes de evidencia de validez, Kane propone los pasos para que, a partir de inferencias bien definidas, podamos analizar estas fuentes. Al realizar cualquier análisis de validez es importante hacer referencia al marco que se está utilizando y explicar la justificación de las fuentes de evidencia propuestas, las que deben estar alineadas al uso e interpretaciones establecidos. Ambos marcos de referencia toman en cuenta aspectos semejantes de las evaluaciones, por lo que una posible línea de investigación sería considerar las fuentes de evidencia de validez de Messick como pruebas o garantías de las inferencias que se hacen a partir del método de Kane, obteniendo así las fuentes de evidencia de validez de manera sistematizada.

REFERENCIAS

1. Sánchez-Mendiola M. «Mi instrumento es más válido que el tuyo»: ¿Por qué seguimos usando ideas obsoletas? Inv Ed Med. 2016;5(19):133-5.
2. Roméu Escobar MR, Díaz Quiñones JA. Valoración metodológica de la confección de temarios de exámenes finales de Medicina y Estomatología. Rev Cuba Educ Med Super. 2015;29(3):522-31.
3. Salvatori P. Reliability and Validity of Admissions Tools Used to Select Students for the Health Professions. Adv Heal Sci Educ. 2001;6(2):159-75.
4. Baladrón J, Curbelo J, Sánchez-Lasheras F, Romeo-Ladrero JM, Villacampa T, Fernández-Somoano A. El examen al examen MIR 2015. Aproximación a la validez estructural a través de la teoría clásica de los tests. FEM. 2016;19(4):217.
5. Shepard LA. Evaluating test validity: reprise and progress. Assess Educ. 2016;23(2):268-80. Disponible en: http://dx.doi.org/10.1080/0969594X.2016.1141168
6. American Educational Research Association, American Psychological Association, National Council on Measurement in Education. STANDARDS for Educational and Psychological Testing. 6th ed. American Educational Research Association. Washington, D. C.: American Educational Research

Tabla 1. Las inferencias y sus fuentes de evidencia correspondientes para establecer el argumento de validez

Inferencia	Consiste en	Procedimientos a definir, establecer o seleccionar	Evaluación empírica de:
Puntuación	Suposición acerca de lo apropiado de los criterios de la puntuación y las reglas para combinar las puntuaciones.	• Ítems y opciones de respuesta (preguntas de opción múltiple, falso/ verdadero) • Formato de la observación • Estandarización entre formatos y ocasiones • Rúbrica o criterio de puntuación, procedimientos de implementación, estándar de aprobado/no aprobado • Selección y entrenamiento de los evaluadores (p ej., ECOE) • Reglas para combinar los elementos relacionados con la prueba a partir de fuentes diferentes o para separar elementos no relacionados de la misma fuente • Seguridad de los datos y control de calidad	• Desempeño de ítems y de opciones de respuesta • Formato de observación • Estandarización • Rúbrica o criterio de puntuación • Selección y entrenamiento de los evaluadores, confiabilidad y precisión de los evaluadores (p ej. en evaluación de desempeño – ECOE) • Seguridad de los datos y control de calidad
Generalización	Los ítems de la prueba conforman una muestra del universo de ítems posibles. Esta inferencia supone que se puede generalizar hacia todo el universo de ítems posibles. Se relaciona con la confiabilidad.	• Estrategia de muestreo de los ítems • Tamaño de la muestra (número de preguntas)	• Confiabilidad o generalizabilidad por medio de la teoría de la generalizabilidad • Teoría de respuesta del ítem
Extrapolación	Se podría extender la interpretación a otros dominios de desempeño y predecir cuál será el resultado del sustentante en contextos diferentes al del examen o tareas diferentes en contextos diferentes.	• Alcance de la prueba • Autenticidad del contexto de la prueba • Autenticidad del ítem/escenario • Análisis que demuestren la relación entre el desempeño en la prueba y los dominios o contextos diferentes a los que se desea extrapolar	• Análisis para definir el alcance/ objetivos • Acuerdo entre el proceso y el constructo • Relevancia y autenticidad • Correlación con otra medida que presente la misma relación esperada (con referencia al criterio o convergente; concurrente o predictiva) • Discriminación • Sensibilidad al cambio después de la intervención • Perfil del constructo • Funcionamiento diferencial del ítem
Implicación	Acerca del impacto de la interpretación de los resultados de la prueba sobre el sustentante, otros interesados y la sociedad.	• Estándar de aprobado/no aprobado • Acciones planeadas con base en los resultados de la prueba • Consecuencias voluntarias o involuntarias de las decisiones que se toman a partir de los resultados de la prueba	• Estándar de aprobado/no aprobado • Efectividad de las acciones basadas en los resultados de la prueba • Consecuencias voluntarias o involuntarias de la prueba • Funcionamiento diferencial del ítem

Fuente: Cook et al., 2015; Kane, 2013; Schuwirth & van der Vleuten, 2012.

Association, American Psychological Association & National Council on Measurement in Education; 2014. 243 p.
7. Downing SM. Validity: On the meaningful interpretation of assessment data. Med Educ. 2003;37(9):830-7.
8. Messick S. Validity. 1987. Disponible en: https://onlinelibrary.wiley.com/doi/abs/10.1002/j.2330-8516.1987.tb00244.x
9. Guilford JP. New Standards For Test Evaluation. Educ Psychol Meas. 1946;6(4):427-39.
10. Shepard LA. Evaluating Test Validity." En: Darling-Hammon L, editor. Review of Research in Education. Washington, DC.: AERA; 1993. p. 405-50.
11. Cronbach LJ, Meehl PE. Construct validity in psychological tests. Psychol Bull. 1955;52(4):281-302.
12. York TT, Gibson C, Rankin S. Defining and measuring academic success. PARE. 2015;20(5):1-20.
13. Cook DA, Hatala R. Validation of educational assessments: a primer for simulation and beyond. Adv Simul. 2016;1(1):1-12. Disponible en: http://dx.doi.org/10.1186/s41077-016-0033-y
14. Sireci S, Faulkner-Bond M. Evidencia de validez basada en el contenido del test. Psicothema. 2014;26(1):100-7.
15. Padilla JL, Benítez I. Evidencia de validez basada en los procesos de respuesta. Psicothema. 2014;26(1):136-44.
16. Embretson SE. A Cognitive Design System Approach to Generating Valid Tests: Application to Abstract Reasoning. Psychol Methods. 1998;3(3):380-96.
17. Leenen I. Virtudes y limitaciones de la teoría de respuesta al ítem para la evaluación educativa en las ciencias médicas. Inv Ed Med. 2014;3(9):40-55.
18. Rios J, Wells C. Evidencia de validez basada en la estructura interna. Psicothema. 2014;26(1):108-16.
19. Downing SM. Reliability: on the reproducibility of assessment data. Med Educ. 2004;38:1006-12.
20. Campbell D, Fiske D. Convergent and discriminant validation by the multitrait-multimethod matrix. Psychol Bull. 1959;56(2):81-105.
21. Coates H. Establishing the criterion validity of the Graduate Medical School Admissions Test (GAMSAT). Med Educ. 2008;42(10):999-1006.
22. Lane S. Evidencia de validez basada en las consecuencias del uso del test. Psicothema. 2014;26(1):127-35.
23. Secretaría de Salud, Secretaría de Educación Pública, Comisión Interinstitucional para la Formación de Recursos Humanos para la Salud. XLIII Examen Nacional para Aspirantes a Residencias Médicas. Convocatoria 2019. Ciudad de México, México.; 2019. Disponible en: http://www.cifrhs.salud.gob.mx/site1/enarm/docs/2019/E43_convo_2019.pdf
24. Kane M. Validating score interpretations and uses: Messick Lecture, Language Testing Research Colloquium, Cambridge, April 2010. Lang Test. 2011;29(1):3-17.
25. Cook DA, Kuper A, Hatala R, Ginsburg S. When assessment data are words: Validity evidence for qualitative educational assessments. Acad Med. 2016;91(10):1359-69.
26. Cook DA, Brydges R, Ginsburg S, Hatala R. A contemporary approach to validity arguments: A practical guide to Kane's framework. Med Educ. 2015;49(6):560-75.
27. Brennan R. Commentary on "Validating the Interpretations and Uses of Test Scores." J Educ Meas. 2013;50(1):74-83.
28. Kane MT. An argument-based approach to validity in evaluation. Psychol Bull. 1992;112(3):527-35.
29. Kane MT. Validating the Interpretations and Uses of Test Scores. J Educ Meas. 2013;50(1):1-73.
30. Chalhoub-Deville M. Validity theory: Reform policies, accountability testing, and consequences. Lang Test. 2016;33(4):453-72.
31. Brennan R. Generalizability Theory. New York: Springer-Verlag New York; 2001. XX, 538.
32. Schuwirth LWT, van der Vleuten CPM. Programmatic assessment and Kane's validity perspective. Med Educ. 2012;46(1):38-48.

Amenazas a la validez en evaluación: implicaciones en educación médica

Blanca Ariadna Carrillo Avalos[a,*], Melchor Sánchez Mendiola[b], Iwin Leenen[c]

Facultad de Medicina

Resumen

Las amenazas a la validez en evaluación educativa son elementos que interfieren con la interpretación propuesta de los resultados de una prueba, pueden ocurrir tanto en exámenes escritos como en pruebas de desempeño y evaluación de competencias clínicas. Estas amenazas se suelen agrupar en dos clases principales: subrepresentación del constructo y varianza irrelevante al constructo. La primera se refiere a que en la prueba no haya suficientes ítems, casos u observaciones para generalizar apropiadamente al dominio completo que se pretende evaluar. La segunda tiene que ver con la presencia de sesgos que interfieren de manera sistemática con la interpretación de los resultados de una prueba, como pueden ser la calidad de los ítems y errores sistemáticos de los evaluadores, entre otros factores que pueden influir sobre la puntuación obtenida. En este artículo se describen las características de las amenazas principales, su importancia y algunas recomendaciones para evitarlas al elaborar y aplicar instrumentos de evaluación en ciencias de la salud. La comprensión de estas amenazas es útil para desarrollar pruebas cuyos resultados tengan niveles aceptables de validez que nos permitan conocer mejor el desempeño de los estudiantes.

Palabras clave: *Amenazas a la validez; evaluación del aprendizaje; educación médica; México.*

[a]Departamento de Ciencias Morfológicas, Facultad de Medicina, Universidad Autónoma de San Luis Potosí, S. L. P., México.
[b]División de Estudios de Posgrado, Facultad de Medicina, Universidad Nacional Autónoma de México, Cd. Mx., México.
[c]División de Estudios de Posgrado, Facultad de Psicología, Universidad Nacional Autónoma de México, Cd. Mx., México.
Recibido: 10-diciembre-2019. Aceptado: 17-febrero-2020.
*Autora para correspondencia: Blanca Ariadna Carrillo Avalos. Av. Venustiano Carranza 2405, Col. Los Filtros, San Luis Potosí, San Luis Potosí, México. CP 78210. Teléfono: 4448 2623 45, ext.: 6635. Correo electrónico: bariadna@gmail.com
La revisión por pares es responsabilidad de la Universidad Nacional Autónoma de México.

https://doi.org/10.22201/facmed.20075057e.2020.34.221

Threats to validity in assessment: implications in medical education
Abstract
Validity threats in educational assessment are elements that interfere with the proposed interpretation of a test score. They can occur in written tests as well as in performance and clinical competency assessments. They are usually grouped in two major categories: construct underrepresentation and construct-irrelevant variance. The former refers to tests with insufficient items, cases, or observations to make a proper generalization towards the full to-be-assessed domain. The latter is related to the presence of biases that can interfere systematically with the interpretation of a test score, such as item quality and raters' systematic errors, among other factors that may have an effect on the obtained score. In this paper we describe the characteristics of some of these threats, their importance, and some recommendations to avoid them during the development of assessment instruments in health sciences education. The insights offered can be useful to devise tests and assessment instruments that allow us to draw more valid inferences about students' knowledge and abilities.

Keywords: *Validity; validity threats; learning assessment; medical education; Mexico.*

INTRODUCCIÓN

El análisis de la validez de los usos e interpretaciones de las puntuaciones de una prueba nos informará sobre el grado en que son apropiados estos usos e interpretaciones para los fines de la evaluación. Sin embargo, la tarea de validación no termina aquí, ya que es necesario descartar otras hipótesis que puedan explicar resultados que no concuerden con la hipótesis original, e identificar elementos que puedan interferir con la interpretación apropiada de los resultados[1-3]. Estas hipótesis apuntan a posibles amenazas a la validez y considerarlas dará mayor fortaleza a las decisiones que se tomen con base en las puntuaciones del examen que estamos valorando. Este análisis cobra mayor relevancia mientras mayor sea el escrutinio al que esté sometido el proceso de evaluación, y mayores sean las potenciales consecuencias del uso de los resultados en los sustentantes, los docentes y las instituciones educativas.

En otro artículo revisamos el concepto moderno de validez en evaluación educativa y su relevancia en educación médica[4]. En este trabajo describiremos las principales amenazas a la validez que existen en evaluación educativa, sus implicaciones en educación en ciencias de la salud y algunas recomendaciones para evitarlas.

Las amenazas a la validez son factores que interfieren con la interpretación del significado de la puntuación obtenida en la evaluación[2,3]. Pueden encontrarse en cualquier tipo de evaluación, ya sea de conocimientos teóricos o prácticos, diagnóstica, formativa o sumativa[3]. En muchas ocasiones los exámenes que se aplican en las escuelas y facultades de medicina, enfermería y otras ciencias de la salud se hacen por medio de preguntas de opción múltiple (POM)[5,6], en este artículo nos enfocaremos principalmente en este tipo de pruebas, aunque las amenazas a la validez se pueden presentar –y deben considerarse– también en evaluaciones prácticas como el examen clínico objetivo estructurado (ECOE). Con respecto a las evaluaciones con POM, se han publicado varios estudios que documentan que la calidad de los reactivos o ítems es limitada[7-9], ya que con frecuencia no se elaboran con el profesionalismo necesario ni siguiendo los lineamientos técnicos para ello[6].

Aunque se mencionan varios tipos de amenazas (por ejemplo, Crooks, Kane y Cohen consideran al menos 23, relacionadas con ocho inferencias)[10], en general se agrupan en dos clases principales: la subrepresentación del constructo (SC) y la varianza irrelevante al constructo (VIC)[11]. A continuación explicamos estos dos conceptos.

Según la teoría clásica de los test (TCT), la puntuación observada *(X)* es una combinación de la puntuación verdadera *(true = T)*, más un componente de error aleatorio *(random error = E_r)*:[12,13]

$$X = T + E_r.$$

En esta fórmula, la puntuación verdadera T resulta de todos los factores que tienen un efecto sistemático sobre la puntuación observada X, incluyendo tanto el constructo de interés como otros factores sistemáticos que no son el objetivo de la medición (por ejemplo, gran severidad de un examinador en un ECOE que cause disminución sistemática de las puntuaciones). Por otro lado, el error aleatorio (E_r) recoge el efecto de todas las circunstancias que afectan la puntuación observada de manera no sistemática, es decir factores que varían cada vez que se aplica la prueba, como el cansancio o estrés del alumno[14]. Tanto la puntuación verdadera como el error aleatorio son constructos hipotéticos y desconocidos, pero por medio de métodos de la TCT se pueden hacer conclusiones a partir de una muestra[15].

La discusión anterior indica que la puntuación verdadera puede descomponerse en dos partes: la puntuación en el constructo de interés (Θ) más la puntuación que se debe a otros factores sistemáticos. Como la segunda parte incluye efectos de factores no intencionados, Haladyna y Downing[14] la denominan el error sistemático *(E_s)* y obtienen la siguiente fórmula:

$$X = \Theta + E_s + E_r. \quad (1)$$

A partir de esta fórmula, se definen los conceptos de SC y VIC. Por un lado, existe una amenaza a la validez cuando la medición de Θ es a través de ítems que no son representativos del dominio completo a evaluar; es decir, cuando los ítems de la prueba evalúan *de manera incompleta* el constructo que se desea medir. Este caso se considera SC. Por otro lado, la VIC está asociada con el error sistemático E_s, el cual es causado por la medición involuntaria de constructos irrelevantes –cuya medición no es el objetivo del examen–, por lo que interfieren con la medición del constructo original y por lo tanto con la validez de la interpretación de la puntuación[2,11,14].

Mención aparte merece el componente E_r de la fórmula (1). Por definición, este componente no produce SC ni VIC, ya que su efecto no es sistemático. Sin embargo, la varianza debido a E_r no es deseable y también constituye una amenaza a la validez. En el marco de la TCT, los factores reunidos en E_r conllevan una baja confiabilidad (y un error estándar de medición grande)[2,9,16]. En este sentido, la fórmula (1) permite ilustrar la diferencia entre validez y confiabilidad. Por un lado, tanto E_r y E_s se refieren a errores a la medición del constructo y, por lo tanto, ambos constituyen amenazas a la validez; por otro lado, solo E_r causa varianza no sistemática y, por lo tanto, solo este factor está asociado con la (baja) confiabilidad. Esto aclara por qué confiabilidad se considera un prerrequisito para validez. En el resto de este artículo solo se considerarán amenazas a la validez relacionadas con factores sistemáticos: SC y VIC.

SUBREPRESENTACIÓN DEL CONSTRUCTO (SC)

En el caso de una prueba escrita, la SC se refiere a que, considerando el universo de ítems o preguntas posibles relevantes al dominio explorado, la prueba esté integrada por una muestra de ítems que puede:

- Tener muy pocos ítems y ser insuficiente para evaluar el dominio del conocimiento correspondiente,
- Estar sesgada hacia un área del tema a evaluar, convirtiéndose en una muestra no representativa,
- Evaluar contenido trivial o factual al nivel más bajo de la pirámide de Miller[2,9,17].

La SC es una amenaza particularmente importante para la inferencia de extrapolación, ya que la interpretación de las puntuaciones es más limitada si los resultados no son representativos del constructo que se supone que la prueba evalúa[18].

Utilizaremos para ilustrar las distintas amenazas a la validez un ejemplo de ciencias básicas: el tema de anatomía de la cabeza. Este tema, sin neuroanatomía, abarca 160 páginas del libro de "Anatomía con orientación clínica de Moore"[19], uno de los libros más utilizados para la enseñanza de anatomía humana en México. Si aplicáramos el examen de la **tabla 1** con el objetivo de evaluar los conocimientos de anatomía representados en el libro de Moore, las amenazas a la validez con respecto a la SC serían las siguientes:

tuación observada *(X)* es una combinación de la puntuación verdadera *(true = T)*, más un componente de error aleatorio *(random error =* E_r*)*:[12,13]

$$X = T + E_r.$$

En esta fórmula, la puntuación verdadera T resulta de todos los factores que tienen un efecto sistemático sobre la puntuación observada X, incluyendo tanto el constructo de interés como otros factores sistemáticos que no son el objetivo de la medición (por ejemplo, gran severidad de un examinador en un ECOE que cause disminución sistemática de las puntuaciones). Por otro lado, el error aleatorio (E_r) recoge el efecto de todas las circunstancias que afectan la puntuación observada de manera no sistemática, es decir factores que varían cada vez que se aplica la prueba, como el cansancio o estrés del alumno[14]. Tanto la puntuación verdadera como el error aleatorio son constructos hipotéticos y desconocidos, pero por medio de métodos de la TCT se pueden hacer conclusiones a partir de una muestra[15].

La discusión anterior indica que la puntuación verdadera puede descomponerse en dos partes: la puntuación en el constructo de interés (θ) más la puntuación que se debe a otros factores sistemáticos. Como la segunda parte incluye efectos de factores no intencionados, Haladyna y Downing[14] la denominan el error sistemático (E_s) y obtienen la siguiente fórmula:

$$X = \theta + E_s + E_r. \quad (1)$$

A partir de esta fórmula, se definen los conceptos de SC y VIC. Por un lado, existe una amenaza a la validez cuando la medición de θ es a través de ítems que no son representativos del dominio completo a evaluar; es decir, cuando los ítems de la prueba evalúan *de manera incompleta* el constructo que se desea medir. Este caso se considera SC. Por otro lado, la VIC está asociada con el error sistemático E_s, el cual es causado por la medición involuntaria de constructos irrelevantes –cuya medición no es el objetivo del examen–, por lo que interfieren con la medición del constructo original y por lo tanto con la validez de la interpretación de la puntuación[2,11,14].

Mención aparte merece el componente E_r de la fórmula (1). Por definición, este componente no produce SC ni VIC, ya que su efecto no es sistemático. Sin embargo, la varianza debido a E_r no es deseable y también constituye una amenaza a la validez. En el marco de la TCT, los factores reunidos en E_r conllevan una baja confiabilidad (y un error estándar de medición grande)[2,9,16]. En este sentido, la fórmula (1) permite ilustrar la diferencia entre validez y confiabilidad. Por un lado, tanto E_r y E_s se refieren a errores a la medición del constructo y, por lo tanto, ambos constituyen amenazas a la validez; por otro lado, solo E_r causa varianza no sistemática y, por lo tanto, solo este factor está asociado con la (baja) confiabilidad. Esto aclara por qué confiabilidad se considera un prerrequisito para validez. En el resto de este artículo solo se considerarán amenazas a la validez relacionadas con factores sistemáticos: SC y VIC.

SUBREPRESENTACIÓN DEL CONSTRUCTO (SC)

En el caso de una prueba escrita, la SC se refiere a que, considerando el universo de ítems o preguntas posibles relevantes al dominio explorado, la prueba esté integrada por una muestra de ítems que puede:

- Tener muy pocos ítems y ser insuficiente para evaluar el dominio del conocimiento correspondiente,
- Estar sesgada hacia un área del tema a evaluar, convirtiéndose en una muestra no representativa,
- Evaluar contenido trivial o factual al nivel más bajo de la pirámide de Miller[2,9,17].

La SC es una amenaza particularmente importante para la inferencia de extrapolación, ya que la interpretación de las puntuaciones es más limitada si los resultados no son representativos del constructo que se supone que la prueba evalúa[18].

Utilizaremos para ilustrar las distintas amenazas a la validez un ejemplo de ciencias básicas: el tema de anatomía de la cabeza. Este tema, sin neuroanatomía, abarca 160 páginas del libro de "Anatomía con orientación clínica de Moore"[19], uno de los libros más utilizados para la enseñanza de anatomía humana en México. Si aplicáramos el examen de la **tabla 1** con el objetivo de evaluar los conocimientos de anatomía representados en el libro de Moore, las amenazas a la validez con respecto a la SC serían las siguientes:

Tabla 1. Ejemplos de preguntas de un examen de anatomía de la cabeza

Pregunta	Opciones de respuesta
1. ¿Cuántos huesos conforman el viscerocráneo?	a. 11 b. 12 c. 13 d. 14 e. 15*
2. La siguiente estructura generalmente está inervada por el nervio laríngeo interno:	a. Aritenoides oblicuo b. Cricoaritenoideo posterior c. Cricotiroideo d. Mucosa infralaríngea e. Mucosa supralaríngea*
3. En la coroides no ocurre lo siguiente:	a. Contiene ramas de la arteria central de la retina* b. La lámina coroidocapilar es la más interna c. Produce el reflejo rojo del fondo de ojo d. Se encuentra entre la esclera y la retina e. Sus venas drenan en una vena vorticosa
4. Una mujer joven se golpea la cabeza con el cuadro de mandos del automóvil durante una colisión frontal. A continuación, sufre un desgarro de la parte frontal del cuero cabelludo con sangrado abundante. La herida se lava con suero fisiológico y se cubre con una venda estéril. Cuando la mujer llega al hospital tiene los dos ojos morados. En la exploración posterior no se aprecia ninguna lesión ocular[19]. ¿Cuál es la arteria que más probablemente se lesionó en este caso?	a. Auricular posterior b. Facial, porción cervical c. Mentoniana d. Supraorbitaria* e. Temporal superficial
5. ¿Cuál es la acción principal del músculo recto inferior? I. Abducir el globo ocular II. Aducir el globo ocular II. Descender el globo ocular IV. Rotar lateralmente el globo ocular V. Rotar medialmente el globo ocular	a. I, II y III b. II, III y IV* c. III, IV y V d. I, III y V e. I y IV
6. Which bone does NOT contribute to the orbit?	a. Frontal bone b. Maxilla c. Palate bone d. Sphenoid bone e. Temporal bone*
7. Un boxeador recibió un golpe en la cara lateral de la nariz, quedando deformada y con los huesos nasales desplazados. Asimismo, presentaba una rotura de los cartílagos de la nariz, epistaxis y obstrucción de la vía respiratoria nasal. ¿Cuál es la arteria en donde se origina la epistaxis?	a. Etmoidal anterior b. Nasal lateral* c. Supraorbitaria d. Supratroclear e. Transversa de la cara
8. ¿Cuál de los siguientes es un músculo de la cara?	a. Bíceps braquial b. Dorsal ancho c. Esternocleidomastoideo d. Frontal* e. Psoas mayor
* Respuesta correcta	

- *Número de preguntas insuficiente.* Un examen que consta de 8 ítems no será adecuado a la luz del amplio universo de ítems de anatomía de la cabeza que se pueden considerar, con base en la extensión de los temas que comprende esta unidad y los objetivos de aprendizaje que se hayan establecido en el currículo. Downing y Haladyna[2] sugieren un mínimo de 30 preguntas en general, mientras que en el manual del *National Board of Medical Examiners* sugieren 100 preguntas para obtener resultados reproducibles[20], aunque estos autores no especifican el tipo de prueba al que van dirigidas estas recomendaciones. En general, para determinar la cantidad adecuada de ítems

Figura 1. Pirámide de Miller[17]

se sugiere considerar los resultados de aprendizaje establecidos en la tabla de especificaciones y factores como el tiempo real que tienen los alumnos para contestar el examen, así como ponderar la importancia de cada uno de los temas a examinar. También es relevante si la prueba es una evaluación sumativa o formativa, así como la exactitud necesaria de las puntuaciones[21,22].

- *Sesgo.* Esta amenaza puede presentarse en caso de que los ítems solo examinen una parte de los temas establecidos en la tabla de especificaciones de la evaluación, sin incluir otras porciones importantes de dicha tabla[2,9].
- *Nivel de evaluación con base en la pirámide de Miller.* Un marco de referencia utilizado en educación médica es la pirámide de Miller (**figura 1**), en la que se proponen los niveles de desarrollo académico y profesional a evaluar, así como una estructura de evaluación y planeación de actividades de aprendizaje[23,24]. Esta amenaza se refiere a que, en el caso de que los objetivos de aprendizaje y evaluación contemplaran niveles de competencia, desempeño o ejecución en la pirámide de Miller, las preguntas fueran mayoritariamente acerca de hechos memorizables (como las preguntas 1, 6 y 8 de la **tabla 1**), y que no evaluaran niveles superiores como la integración entre estos conocimientos y otros previamente adquiridos, ni su relación con la aplicación clínica o con los contenidos de otras asignaturas cuyos temas estén relacionados con las estructuras estudiadas. En una ciencia básica como anatomía, no es fácil elaborar ítems que vayan más allá de conocimientos factuales; sin embargo, es posible lograrlo mientras se tengan claros los objetivos de aprendizaje y los de evaluación en la tabla de especificaciones, así como los usos e interpretaciones de los resultados de la prueba[25].

VARIANZA IRRELEVANTE AL CONSTRUCTO (VIC)

Como ya se mencionó, la VIC se origina del error sistemático debido a una variable irrelevante al constructo que se pretende medir[14]. A continuación, discutimos algunas características de un examen que suelen ocasionar VIC y las ilustramos con el mismo ejemplo del examen de 8 preguntas en la **tabla 1**:

- *Ítems mal elaborados*. Es importante conocer las características de una POM de calidad, descritas en varios documentos[20,26], para evitar ítems defectuosos que puedan causar mayor dificultad para contestarlos o que incluso presenten pistas basadas en aspectos formales para determinar la respuesta correcta[27]. Por ejemplo, en la pregunta 2 de la **tabla 1** no sabemos qué significa "generalmente" ni a qué tipo de estructuras se refiere (¿músculos?). Además, la respuesta correcta es la única estructura que parece no ser un músculo. Otros defectos consisten en elaborar preguntas con opciones que incluyan "todas las anteriores" o "ninguna de las anteriores"[28]. Otro tema ampliamente estudiado en la elaboración de POM es la cantidad de opciones[29-31].
- *Lenguaje*. Por su nivel, dificultades o ambigüedad en la redacción, los formatos muy complicados o extensos (como la pregunta 4 de la **tabla 1**), hacen que el sustentante pase más tiempo leyendo que determinando la respuesta correcta, y esto debe considerarse con respecto al tiempo real que se tiene para presentar la prueba[20]. Un defecto común es elaborar preguntas que pueden confundir al alumno; es decir, preguntas que, aunque sí conoce la respuesta, podría contestar mal: por ejemplo, al contestar la pregunta 5 de la **tabla 1**, el alumno que podría saber las funciones del músculo referido, tendrá que pasar tiempo relacionando los números romanos con la opción correcta. Además, primero debe saber los números romanos[20]. La estructura de las oraciones debe ser lo suficientemente clara y evitar el uso de jerga para que no sea causa de respuestas equivocadas. Un ejemplo es la pregunta 7 de la **tabla 1**, que contiene la palabra "epistaxis" que puede ser confusa para un estudiante de primer año, pues todavía no conoce los términos clínicos[32].
- *Formato en negativo*. Chiavaroli[33] explica que deben evitarse preguntas que incluyen negaciones como las preguntas 3 y 6 de la **tabla 1**. Esto es debido a que existe un doble negativo (en el sentido de que la pregunta incluye una negación e identificar las opciones incorrectas implica negarlas –decir que *no* son correctas–), por lo que existe el riesgo de que el alumno no identifique la parte negativa de la pregunta (aunque la palabra "excepto" o "no" se encuentre en negritas), y que la forma de contestar no se lleva a cabo mediante el proceso de respuesta deseado, afectando así esta evidencia de validez. En el caso de POM en asignaturas de ciencias clínicas, se puede evitar la negación utilizando términos como "cuál es la contraindicación o el riesgo".
- *Funcionamiento diferencial de ítem (differential item functioning; DIF).* El DIF significa que los sustentantes con características o antecedentes distintos (por ejemplo, de diferente sexo o nivel socioeconómico) no tienen la misma probabilidad de responder de manera correcta *a pesar de poseer el mismo nivel en el constructo que se desea medir*. Además de diferencias de género o nivel socioeconómico, un análisis DIF puede comparar grupos diferentes con respecto a características demográficas, religiosas, culturales o lingüísticas[34]. Un ejemplo de esta amenaza se presenta en la pregunta 6 de la **tabla 1**: está en otro idioma (además de que su formato es negativo), de tal manera que los alumnos que no sepan inglés, aunque tengan el conocimiento evaluado por esta pregunta (y suponiendo que el inglés no es parte del constructo que se desea medir), podrían responder incorrectamente[35]. Otro ejemplo es cuando preguntamos las manifestaciones de la lesión del nervio mediano con el término "mano de predicador"; los estudiantes de algunas religiones pueden no entender a qué se refiere.
- *Discordancia con el dominio*. Si entre los objetivos de aprendizaje no se establece el estudio de un tema en particular, sería equivocado evaluarlo, ya que esto causaría que los ítems no correspondieran al dominio de contenido que se pretende evaluar[2].
- *Hacer trampa*. Hay muchas formas de hacer trampa en los exámenes: copiar al compañero de junto, usar un "acordeón" o algo similar, tener acceso a las preguntas de manera previa a la presentación del examen, y hasta el uso de los *smart watches*[36]. Estos comportamientos pueden generar falsos positivos y en este sentido introducir varianza sistemática no deseada en las puntuaciones de las pruebas[9].
- *Enseñar a la prueba ("teaching to the test")*. Se refiere a que los alumnos reciban entrenamiento

para contestar los ítems de una prueba en particular, incluso practicando con las preguntas que aparecerán en el examen real. Esta práctica es una amenaza para la validez porque los alumnos están aprendiendo las respuestas de memoria sin adquirir el conocimiento que están evaluando las preguntas; de esta manera no es posible generalizar el resultado hacia el resto del universo de ítems posibles que evalúan el constructo deseado[37].

- *Testwiseness.* Con base en la gran cantidad de exámenes de opción múltiple que contestan durante su vida académica, se considera que muchos estudiantes de medicina son *test wise*[28]. Quiere decir que han desarrollado estrategias para contestar exámenes deduciendo cuál es la respuesta correcta con base en la estructura gramatical y de redacción: opciones más largas, opciones con más detalles, etc. El dominio de dichas estrategias es irrelevante al constructo, ya que causa que las respuestas no reflejen lo que los estudiantes saben realmente[38]. Es importante distinguir el *testwiseness* de otros conceptos como el *educated guessing*[39], de manera que con el primero se pueden conseguir respuestas correctas, aun sin tener conocimiento; mientras que con el segundo los alumnos logran eliminar opciones con base en el rasgo latente que se desea medir por medio de la evaluación, pero no consiguen identificar por completo la respuesta correcta, por lo que terminan adivinando.

CONCLUSIONES

Las amenazas a la validez resultan aspectos importantes a tomar en cuenta durante la planeación y desarrollo de una prueba, ya que su presencia disminuye la validez de sus resultados, confunde la interpretación propuesta de los mismos y lleva a conclusiones e inferencias erróneas.

Cuando planeamos y desarrollamos pruebas para evaluar eficazmente el constructo deseado, es necesario que capacitemos y motivemos a los elaboradores de preguntas de nuestras escuelas para que tengan "la voluntad de invertir bastante tiempo y esfuerzo en crear preguntas de opción múltiple efectivas"[9]. Tomar en cuenta las amenazas a validez descritas permite afrontarlas y corregirlas antes de que ocurran y afecten las interpretaciones de las puntuaciones de la prueba. Debemos adoptar una actitud más proactiva hacia la prevención de estas amenazas, incluyendo su descripción y efectos en las actividades de formación docente.

Con respecto a las amenazas por subrepresentación del constructo, una recomendación fundamental es establecer claramente, desde la tabla de especificaciones, los objetivos de aprendizaje y el dominio explorado, así como la importancia y la proporción de preguntas que deberán asociarse a cada subtema. Por otro lado, la varianza irrelevante de constructo puede disminuirse significativamente al desarrollar habilidades para la elaboración correcta de ítems de opción múltiple.

Debemos impartir talleres de elaboración de preguntas, tanto para ciencias básicas como para ciencias clínicas; un comité evaluador con experiencia en la elaboración correcta de preguntas debe revisar de forma colegiada el instrumento de evaluación antes y después de su aplicación. Asimismo, sería recomendable incluir en la prueba preguntas que consideren varios niveles de la pirámide de Miller, para ampliar y profundizar el abanico de evaluación de los profesionales de la salud.

REFERENCIAS

1. Cronbach LJ. Five perspectives on validity argument. En: Wainer H, Braun HI, editores. Test validity [Internet]. New York: Routledge; 1988. p. 3-17. Disponible en: https://doi.org/10.4324/9780203056905
2. Downing SM, Haladyna TM. Validity threats: Overcoming interference with proposed interpretations of assessment data. Med Educ. 2004;38(3):327-33.
3. Downing SM, Yudkowski R, editores. Assessment in health professions education. New York and London: Routledge; 2009. 317 p.
4. Carrillo BA, Sánchez M, Leenen I. El concepto moderno de validez y su uso en educación médica. Inv Ed Med. 2020; 9(33):98-106.
5. Norman G, van der Vleuten C, Newble D. International Handbook of Research in Medical Education. Norman G, van der Vleuten C, Newble D, editores. Springer; 2002. 1106 p.
6. Jozefowicz RF, Koeppen BM, Case S, Galbraith R, Swanson D, Glew RH. The quality of in-house medical school examinations. Acad Med. 2002;77(2):156-61.
7. Ware J, Vik T. Quality assurance of item writing: During the introduction of multiple choice questions in medicine for high stakes examinations. Med Teach. 2009;31(3):238-43.
8. Tarrant M, Knierim A, Hayes SK, Ware J. The frequency of item writing flaws in multiple-choice questions used in

high stakes nursing assessments. Nurse Educ Today. 2006;26(8):662-71.
9. Downing SM. Threats to the validity of locally developed multiple-choice tests in medical education: Construct-irrelevant variance and construct underrepresentation. Adv Heal Sci Educ. 2002;7(3):235-41.
10. Crooks TJ, Kane MT, Cohen AS. Threats to the valid use of assessments. Assess Educ Princ Policy Pract. 1996;3(3):265-85.
11. Messick S. Validity. En: Linn RL, editor. Educational Measurement [Internet]. New York: Macmillan; 1989. p. 13-103. Disponible en: https://onlinelibrary.wiley.com/doi/abs/10.1002/j.2330-8516.1987.tb00244.x
12. Schuwirth LWT, Van Der Vleuten CPM. General overview of the theories used in assessment: AMEE Guide No. 57. Med Teach. 2011;33(10):783-97.
13. De Champlain AF. A primer on classical test theory and item response theory for assessments in medical education. Med Educ. 2010;44(1):109-17.
14. Haladyna TM, Downing SM. Construct-Irrelevant Variance in High-Stakes Testing. Educ Meas Issues Pract [Internet]. 2004;23(1):17-27. Disponible en: https://onlinelibrary.wiley.com/doi/abs/10.1111/j.1745-3992.2004.tb00149.x
15. Leenen I. Virtudes y limitaciones de la teoría de respuesta al ítem para la evaluación educativa en las ciencias médicas. Inv Ed Med. 2014;3(9):40-55.
16. Downing SM. Reliability : on the reproducibility of assessment data. Med Educ. 2004;38:1006-12.
17. Miller GE. The assessment of clinical skills/competence/performance. Acad Med. 1990;65(9):S63-7.
18. Hawkins RE, Margolis MJ, Durning SJ, Norcini JJ. Constructing a validity argument for the mini-clinical evaluation exercise: A review of the research. Acad Med. 2010;85(9):1453-61.
19. Moore K, Dailey A, Agur A. Anatomía con orientación clínica. 7a ed. Philadelphia: Wolters Kluwer Health, S.A., Lippincot Williams & Wilkins; 2013.
20. National Board of Medical Examiners. Cómo elaborar preguntas para evaluaciones escritas en el área de ciencias básicas y clínicas. 4th ed. Paniagua MA, Swygert KA, editores. Philadelphia, PA: National Board of Medical Examiners; 2016. 100 p.
21. Moreno R, Martínez RJ, Muñiz J. Directrices para la construcción de ítems de elección múltiple. Psicothema [Internet]. 2004;16(3):490-7. Disponible en: https://www.redalyc.org/articulo.oa?id=72716324
22. American Educational Research Association, American Psychological Association, National Council on Measurement in Education. STANDARDS for Educational and Psychological Testing. 6th ed. American Educational Research Association. Washington, D. C.: American Educational Research Association, American Psychological Association & National Council on Measurement in Education; 2014. 243 p.
23. Williams BW, Byrne PD, Welindt D, Williams M V. Miller's pyramid and core competency assessment: A study in relationship construct validity. J Contin Educ Health Prof. 2016;36(4):295-9.
24. Pangaro L, Ten Cate O. Frameworks for learner assessment in medicine: AMEE Guide No. 78. Med Teach. 2013;35:e1197-e1210.
25. Hadie SNH. The Application of Learning Taxonomy in Anatomy Assessment in Medical School. Educ Med J. 2018;10(1):13-23.
26. Haladyna TM, Downing SM, Rodriguez MC. A Review of Multiple-Choice Item-Writing Guidelines for Classroom Assessment. Appl Meas Educ. 2002;15(3):309-34.
27. Downing SM. Construct-irrelevant variance and flawed test questions: Do multiple-choice item-writing principles make any difference? Acad Med. 2002;77(10 SUPPL.):103-4.
28. Downing SM. The effects of violating standard item writing principles on tests and students: The consequences of using flawed test items on achievement examinations in medical education. Adv Heal Sci Educ. 2005;10(2):133-43.
29. Abad FJ, Olea J, Ponsoda V. Analysis of the optimum number alternatives from the Item Response Theory. Psicothema. 2001;13(1):152-8.
30. Rodriguez MC. Three options are optimal for multiple-choice items: A meta-analysis of 80 years of research. Educ Meas Issues Pract. 2005;24(2):3-13.
31. Haladyna TM, Rodriguez MC, Stevens C. Are Multiple-choice Items Too Fat? Appl Meas Educ [Internet]. 2019;32(4):350-64. Disponible en: https://doi.org/10.1080/08957347.2019.1660348
32. Hicks NA. Guidelines for identifying and revising culturally biased multiple-choice nursing examination items. Nurse Educ. 2011;36(6):266-70.
33. Chiavaroli N. Negatively-worded multiple choice questions: An avoidable threat to validity. Pract Assessment, Res Eval. 2017;22(3):1-14.
34. Gómez-Benito J, Sireci S, Padilla JL, Dolores Hidalgo M, Benítez I. Differential item functioning: Beyond validity evidence based on internal structure. Psicothema. 2018;30(1):104-9.
35. Young JW. Ensuring valid content tests for English Language Learners. Educational Testing Service. 2008.
36. Wong S, Yang L, Riecke B, Cramer E, Neustaedter C. Assessing the usability of smartwatches for academic cheating during exams. En: Proceedings of the 19th International Conference on Human-Computer Interaction with Mobile Devices and Services, MobileHCI 2017. Association for Computing Machinery; 2017.
37. Bond L. Teaching to the Test: Coaching or Corruption. New Educ. 2008;4(3):216-23.
38. Lane S, Raymond M, Haladyna T. Handbook of Test Development [Internet]. 2nd ed. Lane S, Raymond M, Haladyna T, editores. International Journal of Testing. New York: Routledge; 2016. 676 p. Disponible en: http://www.tandfonline.com/doi/abs/10.1080/15305050701813433
39. Jurado A, Leenen I. Reflexiones sobre adivinar en preguntas de opción múltiple y cómo afecta el resultado del examen. Inv Ed Med. 2016;5(17):55-63.

Facultad de Medicina

Cartas

Letters

Neuromitos del aprendizaje en un programa de posgrado de educación en ciencias de la salud

Learning neuromyths in a posgraduate in health sciences education program

ESTIMADO SR. EDITOR:

Los profesores en las escuelas y facultades de ciencias de la salud nos preocupamos continuamente por mejorar nuestra forma de enseñanza. Esto ha llevado en diversas ocasiones a cometer equivocaciones o errores de implementación, como en el caso de los neuromitos. Los neuromitos son falsas creencias que se desarrollan, ya sea por una mala interpretación o afirmaciones fuera de contextos de hechos científicamente establecidos. Como lo menciona Hernández Espinosa, 2020, "A diferencia de los mitos en otros ámbitos en la sociedad, los mitos sobre el funcionamiento del cerebro repercuten directamente sobre el ámbito educativo"[1].

Los autores de esta carta somos alumnos en un posgrado en educación en ciencias de la salud, y quisimos darnos cuenta del grado de conocimiento de nuestros compañeros al respecto de los neuromitos. Macdonald y colaboradores aplicaron un instrumento para conocer la prevalencia de los neuromitos y compararla entre educadores, no educadores y personas con conocimientos en neurociencia. Decidimos utilizar el mismo instrumento con nuestros compañeros y recibimos 13 participaciones[2]. La pregunta que tuvo cero respuestas correctas fue con relación a la integración de la función de hemisferios derecho e izquierdo del cerebro a través de sesiones breves de coordinación motriz. Otras cuestiones que tuvieron muy pocas respuestas correctas fueron con respecto al tamaño del cerebro de los niños y las niñas, los signos de dislexia, y la mejoría de habilidades en lectura a través de ejercicios para practicar la coordinación de habilidades de percepción motriz.

Llama mucho la atención que 6 de los 13 participantes (46.2%) estuvieron de acuerdo con que los individuos aprenden mejor cuando reciben información en su estilo de aprendizaje preferido. Estos dos puntos en particular son interesantes en nuestra pequeña muestra porque la mayoría son médicos que están cursando un posgrado en educación en ciencias de la salud, por lo que se esperaría que estuvieran enterados de la gran *leyenda urbana* que constituyen los estilos de aprendizaje. Sin embargo, los tipos de aprendizaje son un neuromito con tal aceptación y penetración en el medio educativo, que lo más común es que muchos de nosotros los considere no solo reales, sino útiles e indispensables de tomar en cuenta para su aplicación en nuestras clases, a pesar de que se ha demostrado en varias ocasiones que no existe evidencia que apoye esta creencia[3].

CONTRIBUCIÓN INDIVIDUAL

Los dos autores contribuyeron en partes iguales al desarrollo del manuscrito.

AGRADECIMIENTOS

A los compañeros del posgrado en Ciencias Sociomédicas con especialidad en Educación en Ciencias de la Salud.

PRESENTACIONES PREVIAS
Ninguna.

FINANCIAMIENTO
Ninguno.

CONFLICTO DE INTERESES
Ninguno.

REFERENCIAS

1. Hernández-Espinosa DR. Mitos y hechos del cerebro que aprende: Las neurociencias en la docencia. Mens Bioquim. 2020;44:65-71.
2. Macdonald K, Germine L, Anderson A, Christodoulou J, McGrath LM. Dispelling the myth: Training in education or neuroscience decreases but does not eliminate beliefs in neuromyths. Frontiers in Psychology. 2017;8(AUG):1-16.
3. Kirschner PA, van Merriënboer JJG. Do Learners Really Know Best? Urban Legends in Education. Educational Psychologist. 2013;48(3):169-83.

Blanca Ariadna Carrillo-Avalos[a,‡], Kevin David Laguna-Maldonado[b,§]
[a] Facultad de Medicina, Universidad Autónoma de San Luis Potosí, México.
[b] Departamento de Bioquímica, Facultad de Medicina, Cd. Mx., México.
ORCID ID:
[a,‡] https://orcid.org/0000-0003-4111-4795
[b,§] https://orcid.org/0000-0002-8428-739X
Recibido: 8-septiembre-2021. Aceptado: 14-septiembre-2021.
Autor para correspondencia: Kevin David Laguna Maldonado. Av. Universidad 3000 Colonia Universidad Nacional Autónoma de México, Ciudad Universitaria, Alcaldía de Coyoacán, C.P. 04510, Ciudad de México. Correo electrónico: d_laguna@hotmail.com.
https://doi.org/10.22201/fm.20075057e.2022.41.21401

Ingeniería biomédica en ciencias de la salud: una necesidad lectiva que surge ante la COVID-19

Biomedical engineering in health sciences: a teaching need that arises from COVID-19

SR. EDITOR:

La ingeniería biomédica, se ocupa de la implementación, funcionamiento y uso de los equipos biomédicos (EB) en el ámbito hospitalario u otros entornos clínicos. Hasta antes del inicio de la pandemia el uso y manejo de los EB, necesarios para el manejo de la terapia intensiva respiratoria, estuvo limitado al personal de áreas críticas; quizá, esto podría haber restringido el conocimiento de su uso y manejo por un mayor número de personal sanitario. Durante la pandemia, la cantidad necesaria de EB y sobre todo personal sanitario capacitado en su manejo, se han convertido en elementos necesarios para enfrentarla. En consecuencia, la necesidad de contar con mayor y mejor personal capacitado, debido al aumento súbito de su demanda, resultarían siendo un factor decisivo en el control de la pandemia.

A pesar que la ingeniería biomédica existe desde hace casi cincuenta años, cuando los ingenieros buscaron adaptar sus conocimientos a la medicina; aún hay una brecha importante entre los beneficios de los conocimientos de ingeniería en medicina y su aplicación en el manejo de los EB por el personal sanitario[1]. Su escasa enseñanza alrededor del mundo[2], podrían justificar el resultado fatal reportado en el 10% del total de pacientes que ingresan con efectos adversos debido al mal uso de los EB en los Estados Unidos[3]. Así, urge la necesidad de que, durante la formación del futuro personal de salud, se incrementen los conocimientos de ingeniería biomédica, permitiendo un mejor manejo y uso de la tecnología médica, desde pregrado, ante cambios inesperados como la COVID-19. Lo anterior plantea la necesidad de contar con docentes, entornos clínicos hospitalarios y autoridades comprometidas en la necesidad de profundizar la enseñanza de la ingeniería biomédica en ciencias de la salud, permitiendo la mejora continua de las habilidades del futuro personal de salud.

En muchos países, principalmente los países subdesarrollados, existen limitaciones en la adquisición y mantenimiento de los EB, debido, quizá, a la poca oferta de personal capacitado. En Perú, solo 5 universidades tienen la inclusión lectiva de ingeniería biomédica, en comparación a Estados Unidos que cuenta con aproximadamente 118 programas acreditados[1]. Tal escasez de profesionales sanitarios capacitados podría limitar la eficiencia en la compra y mantenimiento de los EB, decisivos en tiempos de pandemia.

Evaluación y aprendizaje en educación universitaria: estrategias e instrumentos

EDITORES

Melchor Sánchez Mendiola
Adrián Martínez González

CUAIEED
Coordinación de Universidad Abierta
Innovación Educativa y Educación a Distancia

Primera edición, junio 2022.

Evaluación y aprendizaje en educación universitaria: estrategias e instrumentos

Primera edición
UNAM, Coordinación de Universidad Abierta, Innovación Educativa y Educación a Distancia, 2022

Editores
Sánchez Mendiola, Melchor
Martínez González, Adrián
Evaluación y aprendizaje en educación universitaria: estrategias e instrumentos / Sánchez Mendiola, Melchor, Martínez González, Adrián. —1ª ed. — Ciudad de México, UNAM, 2022.
p. 774
ISBN 978-607-30-6076-9
1. Educación.

La presente obra fue sometida a consideración del Comité Editorial de la Coordinación de Universidad Abierta, Innovación Educativa y Educación a Distancia de la Universidad Nacional Autónoma de México, y fue aprobada por un dictamen académico a través de arbitraje por pares.

Editores
Melchor Sánchez Mendiola
Adrián Martínez González

ISBN: 978-607-30-6071-4
Impreso y hecho en México.

ÍNDICE

Capítulo 2

VALIDEZ, CONFIABILIDAD Y AMENAZAS A LA VALIDEZ

Blanca Ariadna Carrillo Ávalos, Melchor Sánchez Mendiola

"La validez es simple. La validación puede ser difícil."
MICHAEL KANE, 2009

INTRODUCCIÓN

A lo largo de la vida como docentes realizamos muchas evaluaciones para intentar conocer el nivel de conocimiento o desempeño de nuestros estudiantes. Este proceso implica la elaboración, aplicación e interpretación de exámenes de diferentes tipos: diagnósticos, formativos y sumativos. Independientemente de su finalidad, la meta de cualquier evaluación incluye la identificación del nivel de algún constructo, como conocimiento sobre química orgánica, habilidad para resolver problemas de trigonometría, o competencia de comunicación escrita. Los resultados de las evaluaciones idealmente deben reflejar de una manera precisa y reproducible lo que se pretende evaluar, para poder interpretar de forma racional los resultados de la misma y estar en capacidad de realizar inferencias y tomar decisiones con fundamentos sólidos. Cuando evaluamos a nuestros estudiantes sobre un tema particular, deseamos identificar el proceso y resultados del aprendizaje que permitan inferir el nivel de desempeño en los constructos de interés. Después de aplicar las evaluaciones, obtenemos resultados en forma de puntuaciones que ayudan a tomar decisiones, que conllevan las siguientes interrogantes: ¿estamos evaluando exactamente lo que deseamos evaluar?, ¿qué implican los resultados con respecto al avance académico del alumno?, si se trata de una evaluación sumativa, ¿cuál es la calificación mínima para aprobar el curso?, ¿qué tan reproducible es la medición?, entre muchas otras.

La evaluación en educación es una disciplina cada vez más sofisticada y sustentada en investigación, que requiere incorporar conceptos académicos fundamentales para llevarse a cabo con profesionalismo y solidez metodológica (Instituto Nacional para la Evaluación de la Educación, 2017). El pilar conceptual más importante de la evaluación en educación es la validez, tema del que trata este capítulo. Actualmente el concepto de **validez** ha evolucionado del tradicional "medir lo que se pretende medir", a un modelo más amplio y profundo, en el que *"se refiere al grado en que la evidencia y la teoría respaldan las interpretaciones de los puntajes de una prueba para usos propuestos de las pruebas"* (AERA, APA y NCME, 2018). Es un conjunto de acciones que se ubican a lo largo del proceso evaluativo, para fundamentar la interpretación de los resultados y así generar inferencias. El análisis de validez, o **validación**, es el proceso mediante el que evaluamos las evidencias que se presentan para determinar cuál es el grado de validez (Cook y Hatala, 2016). Se puede realizar para los diferentes tipos de exámenes, diagnósticos, formativos y sumativos, aunque es particularmente relevante para las evaluaciones sumativas de alto impacto.

Tradicionalmente la validez en educación se clasificaba como "las 3 Cs": validez de contenido, de criterio y de constructo (Cronbach y Meehl, 1955). En la definición actual esta distinción desapareció, ya que el modelo vigente propone diversas fuentes de evidencia que arrojan luz sobre distintos aspectos de la validez, no es que reflejen diferentes tipos de la misma. La validez es un concepto unitario, por lo que se considera que toda la validez es validez de constructo. La palabra **constructo** significa colecciones de conceptos abstractos y principios, inferidos de la conducta y explicados por una teoría educativa o psicológica, es decir, atributos o características que no pueden observarse directamente, por ejemplo: inteligencia emocional, extroversión, conocimientos sobre matemáticas (AERA, APA y NCME, 2018).

La validez es un juicio valorativo holístico e integrador que requiere múltiples fuentes de evidencia para su interpretación, e intenta responder a la pregunta: *"¿qué inferencias pueden hacerse sobre la persona basándose en los resultados del examen?"* (Downing, 2003). Validez no es una propiedad intrínseca del examen, sino del significado de los resultados en un entorno educativo específico y las inferencias que pueden hacerse a partir de los mismos (Carrillo-Ávalos et al., 2020; Downing, 2003). Por ejemplo, los resultados de los exámenes de admisión a las universidades no deben interpretarse categóricamente como evidencia predictiva de que la persona vaya a tener éxito en la vida, ya que el examen no está diseñado con ese propósito. Quienes desarrollan pruebas e interpretan sus resultados, deben poseer nociones básicas de los conceptos de validez y confiabilidad, para incorporarlos de manera apropiada en el proceso de enseñanza y aprendizaje. Este capítulo presenta un panorama básico de dichos conceptos, además de describir las principales amenazas a la validez.

VALIDEZ

Anteriormente se hablaba de tres tipos de validez: de contenido, de constructo y de criterio; la de criterio se dividía en validez concurrente y validez predictiva (Cronbach y Meehl, 1955). Posteriormente, a finales del siglo XX, un nuevo marco de referencia de validez fue propuesto y aceptado por las principales organizaciones de evaluación educativa y pruebas psicológicas (American Educational Research Association et al., 2018), incorporando el concepto holístico de validez de constructo. Este modelo establece que, para determinar el grado de validez de los usos e interpretaciones de los resultados de una evaluación, se deben proveer diversos elementos que lo demuestren (Downing, 2003). Este esquema propone los siguientes elementos como **cinco fuentes de evidencia** de validez (Downing, 2003; Messick, 1989):

1) **Evidencia basada en el contenido de la prueba.** El contenido de la prueba alude a los conocimientos que evalúa; por ejemplo, en el caso de un examen final de un curso de biología evolutiva, son los relacionados con la historia de la teoría de la evolución y sus evidencias, sus bases genéticas, los procesos del cambio evolutivo, el origen de las especies, macro-evolución y la evolución de los homínidos. Es decir, todos los temas que debe saber el alumnado cuando termina de estudiar esta asignatura.

   Para demostrar que los usos e interpretaciones de un examen de esta naturaleza son adecuados, podemos buscar evidencia de que, en efecto, las preguntas del examen versan sobre los temas mencionados. Estas evidencias descansan en cuatro elementos (Sireci y Faulkner-Bond, 2014):

   - *Definición del dominio.* Se trata de la descripción detallada de las áreas del contenido y las habilidades cognitivas que se desean evaluar del constructo definido en el currículo, o de los resultados de la actividad de aprendizaje. En la tabla de especificaciones de la prueba se deben enlistar las subáreas del contenido y los niveles cognitivos que estaremos midiendo (conocimiento, comprensión, aplicación, análisis, síntesis o evaluación), así como los estándares del contenido y los objetivos curriculares.
   - *Representación del dominio*. Con frecuencia se hacen demasiadas preguntas sobre un tema y se dejan de lado otros. Para decidir cuántos ítems, preguntas o reactivos corresponden a cada tema, podemos establecer una tabla de ponderaciones en la que se establece la relevancia de cada uno, a los de mayor importancia se asignan más preguntas en el examen. También se debe buscar la alineación entre el contenido del examen y los estándares para ese conocimiento. Por ejemplo, cada asignatura que cursa un alumno tiene una carta descriptiva en la que se han establecido los objetivos o metas de aprendizaje, lo que se debe saber al terminar de cursar esa materia. En un examen sumativo, las preguntas deberán contemplar tales objetivos y coincidir con la tabla de especificaciones. Otra manera de demostrar alineación es que expertos en el tema opinen sobre el examen, cómo los ítems evalúan los diferentes aspectos del constructo de interés.
   - *Relevancia del dominio*. Se refiere a qué tan importantes son los ítems con respecto al aspecto del constructo que se está midiendo, que se pregunten conceptos importantes y no datos triviales.
   - *Procedimientos apropiados de diseño de la prueba.* Los procedimientos que se llevan a cabo al diseñar la prueba deben servir para asegurarse de que su contenido evalúa fielmente y representa por completo al constructo de interés. Esto puede comprobarse al implementar controles de calidad durante el desarrollo del examen, como revisión de ítems por expertos en contenido para asegurarse de su veracidad técnica,

revisión por expertos en evaluación para verificar que estén bien elaborados, probar los ítems por medio de estudios piloto.

- *Credenciales de los creadores del examen, elaboradores de reactivos y expertos en contenido*. Es importante documentar que las personas que intervienen en el proceso de diseño del examen, definición del constructo, elaboración de los reactivos y análisis de los resultados, tengan las credenciales correspondientes para dar certidumbre a todo el proceso. Si el examen evalúa habilidades de traducción del idioma alemán, deben participar expertos en contenido que dominen dicho constructo; si se trata de elaborar un examen práctico de habilidades de comunicación verbal, deben participar expertos en el tema y profesionales del diseño de este tipo de exámenes.

2) **Evidencia basada en los procesos de respuesta.** Los procesos de respuesta son los procesos mentales que lleva a cabo el sustentante cuando contesta las preguntas de una prueba. Se esperaría que responda cada ítem integrando los conocimientos que se indagan y que no esté tratando de adivinar la respuesta correcta. Si el examen contiene preguntas de opción múltiple (POM) o abiertas, cuando el alumno termine de contestar se puede indagar cómo llegó a la respuesta a través de una entrevista cognitiva, así se puede saber si comprendió los conceptos clave, si hay errores de redacción en las preguntas, cuál fue el razonamiento que utilizó, y si hay posibilidad de que existan falsos positivos (que el alumno haya utilizado un razonamiento erróneo para llegar a la respuesta). Al final, lo que se busca es que el estudiante aplique en verdad los conocimientos adquiridos para resolver los problemas que se proponen en el examen y que luego los pueda aplicar en la vida real. Otra manera de aportar a esta evidencia es a través de modelos matemáticos que evalúen la dificultad y el tiempo que tardan en contestar cada ítem (Padilla y Benítez, 2014).

   Algunos autores agregan en este apartado la familiaridad de los sustentantes con el formato del examen (por ejemplo, evaluación asistida por computadora), la validación de la hoja de respuestas correctas, el control de calidad del reporte de los resultados, entre otros (Downing, 2003).
3) **Evidencia basada en la estructura interna.** La estructura interna presenta tres características básicas: dimensionalidad, funcionamiento diferencial y confiabilidad (Rios y Wells, 2014). Al diseñar la prueba, se debe determinar cuáles dimensiones se desean evaluar sobre el constructo de interés, y esta información se describe en la tabla de especificaciones del examen. Siguiendo el ejemplo de un examen de biología evolutiva, una dimensión sería el proceso cognitivo que siguen los alumnos para resolver problemas de ese tema. Sin embargo, en ocasiones los exámenes contienen ítems que evalúan diferentes dimensiones del mismo constructo, por ejemplo, procesos cognitivos y valores. Para saber cuántas dimensiones posee el examen, se pueden hacer diversos análisis, entre ellos un análisis factorial confirmatorio, el que también permite identificar cuáles ítems valoran la misma dimensión buscando la relación existente entre ellos. Con el resultado de este análisis podemos justificar, por ejemplo, dar el mismo valor a todos los ítems, porque todos evalúan la misma dimensión del mismo constructo.

Las pruebas también deben aportar resultados imparciales, por lo que es útil buscar información con respecto al funcionamiento diferencial de los ítems (DIF, por sus siglas en inglés) (Leenen, 2014; Rios y Wells, 2014). Se trata de que las características de los ítems de la prueba son comparables entre diferentes grupos; es decir, que no habrá diferencia entre los resultados de hombres y mujeres, o por edades, por ejemplo. Para conocer el grado de invarianza de los resultados entre grupos, se pueden realizar pruebas basadas en el funcionamiento diferencial del ítem, como un análisis factorial confirmatorio de grupos múltiples.

Este apartado de evidencia de validez también incluye el análisis estadístico de los reactivos de la prueba (Downing, 2003), para documentar diversos indicadores como grado de dificultad, índice de discriminación, confiabilidad, error estándar de medición, curvas características del ítem, entre muchos otros, para lo cual existen diversos modelos y métodos matemáticos que se describen en otros capítulos de esta obra.

4) **Evidencia basada en las relaciones con otras variables.** En el ejemplo mencionado, los alumnos que hicieron el examen de la unidad de las bases genéticas de la evolución en la asignatura de biología evolutiva, también contestaron uno de genética, en donde se les hicieron preguntas acerca de genética de poblaciones. Ambas pruebas son semejantes en cuanto al constructo evaluado. La validación de la interpretación de los resultados del examen de biología evolutiva podría incluir el análisis de la relación convergente con los resultados del examen de genética, si ambas pruebas valoran constructos similares. Por otro lado, también puede existir una relación divergente cuando se evalúan constructos diferentes. El análisis para establecer ambas correlaciones se puede realizar por medio de la matriz multirasgo-multimétodo (MTMM) de Campbell (Campbell y Fiske, 1959), en el que se establece una matriz de correlaciones entre las dos pruebas en donde se representan al menos dos características y cada una de ellas debe ser medida por lo menos por dos métodos.

   Esta fuente de evidencia proporciona información acerca del grado en que la relación divergente o convergente es coherente con el constructo cuya medición es la base de la interpretación de los resultados de la prueba. Otra evidencia que aporta a las relaciones con otras variables es la relación entre la prueba y el criterio, la que se puede establecer por medio de uno de estos diseños:

   - *Estudio predictivo*. Para conocer el grado de relación entre el resultado de la prueba y el resultado del criterio que se evalúa posteriormente. Por ejemplo, si en el examen de admisión a la universidad se evalúa biología general, podríamos tratar de responder si las calificaciones de biología general en el examen de admisión predicen las de biología evolutiva. En este caso, el criterio sería el resultado de biología evolutiva. Por otro lado, también se pueden hacer estudios para hacer predicciones diferenciales por grupo de edad, sexo, antecedentes académicos, entre otras.
   - *Estudio concurrente*. Se mencionó en el ejemplo que los alumnos llevan las asignaturas de genética y biología evolutiva al mismo tiempo, por lo que las evaluaciones

correspondientes también ocurrieron en fechas cercanas. Si se evalúa un constructo al mismo tiempo, podemos estimar la relación entre las puntuaciones de la prueba y del criterio.

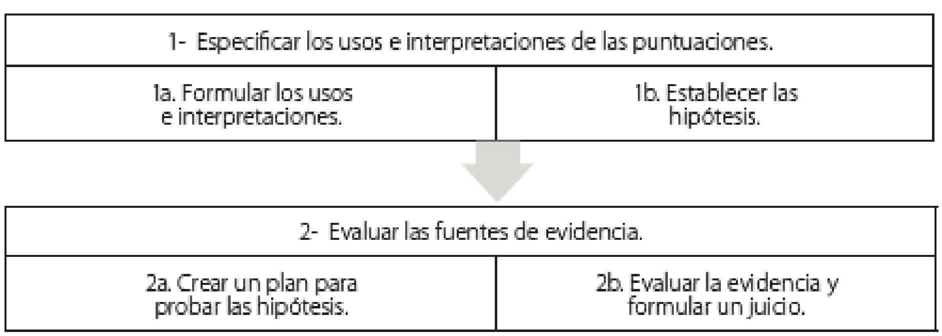
5) **Evidencia basada en las consecuencias de la prueba.** Los resultados de las pruebas, sobre todo las sumativas y de alto impacto como los exámenes de admisión o de titulación, tienen grandes consecuencias para los sustentantes. En el caso de los exámenes de admisión a las universidades, uno de los objetivos puede ser seleccionar a los sustentantes con mayor probabilidad de tener éxito académico (Downing, 2003). Las consecuencias de pruebas sumativas de impacto moderado son importantes también, ya que se van agregando para generar un promedio final que permita el avance en las trayectorias escolares.

   Las fuentes de evidencia de consecuencias de las pruebas se buscan al analizar cuáles son los impactos de los resultados en el avance académico del estudiante, en sus familias, y en la sociedad. En las pruebas cuyos resultados se reportan con base en criterio, la decisión del punto de corte mínimo debe fundamentarse; en el caso de pruebas con referencia a norma también se debe sustentar la decisión de, por ejemplo, por qué admitir a los 100 sustentantes con las puntuaciones más altas. Otras fuentes de evidencia en este rubro pueden ser las consecuencias de aprobar o reprobar, de falsos positivos y falsos negativos, y las consecuencias institucionales. La manera de obtener esta información puede ser a través de entrevistas, grupos focales, cuestionarios, para conocer cuáles son los componentes más importantes de los programas académicos y sus puntos de mayor impacto (Lane, 2014).

VALIDACIÓN

La validación es un proceso que se debe planear al mismo tiempo que se diseña la prueba, para asegurarse de contar con las fuentes de evidencia necesarias para obtener el mayor grado posible de validez de la interpretación de sus resultados. Una manera de realizar este proceso se sugiere a continuación (Figura 1).

Figura 1. Esquema de pasos generales para el proceso de validación de una prueba (elaboración propia)



1. Especificar los usos e interpretaciones de las puntuaciones

1a. Formular los usos e interpretaciones. Los usos y las interpretaciones de las puntuaciones que se obtienen en una prueba son conceptos diferentes y ambos se deben aclarar desde que inicia el diseño de la prueba. La justificación del uso de las puntuaciones se puede conocer respondiendo a preguntas como: ¿debemos utilizar estas puntuaciones para tomar decisiones sobre quiénes pueden ingresar a un programa de posgrado? Para ello, se deben conocer las características de los usuarios principales, quiénes son las personas que presentan dichas evaluaciones; además, también son de interés las instituciones que las desarrollan, profesores, personal administrativo.

En cuanto a las interpretaciones, se pueden determinar al responder preguntas como: ¿las calificaciones del examen de graduación reflejan las competencias que debe poseer un egresado de este programa de posgrado?, ¿los resultados de esta prueba miden el nivel de desempeño de los estudiantes en esta asignatura?

Los datos que se deben evaluar en este paso se describen en la Tabla 1.

Tabla 1. Formulación de los usos e interpretaciones de una prueba o examen

Dato necesario	Pregunta a contestar	Ejemplo
El objetivo de la prueba .	¿Para qué se aplica la prueba?	Establecer el nivel de desempeño en los conocimientos acerca de Biología Evolutiva.
Los usuarios propuestos .	¿Quiénes utilizarán los resultados de la prueba?	La Facultad de Ciencias Biológicas de la universidad.
Los usos.	¿Para qué se utilizarán las puntuaciones?	Las calificaciones serán utilizadas para determinar cuáles alumnos demuestran un desempeño satisfactorio en la asignatura.
El constructo medido.	¿Cuál es constructo que se evalúa con esta prueba?	Los conocimientos en Biología Evolutiva.
Las interpretaciones de los resultados.	¿Los resultados serán interpretados con referencia a norma o a criterio? ¿Los resultados de esta prueba miden el nivel de desempeño de los estudiantes en esta asignatura?	Los resultados del examen final de la asignatura se interpretan con referencia a criterio, por lo que quienes obtengan una puntuación mayor a 6.0 tendrán un desempeño satisfactorio.
La población examinada.	¿Quiénes contestarán el examen?	Los alumnos de 7° semestre de la licenciatura en Biologla. Porcentaje de hombres y mujeres, rango de edades.

1b. Establecer las hipótesis. Las hipótesis son preguntas que nos podemos hacer acerca de la evaluación que se está elaborando. Deben probarse por medio de las fuentes de evidencia mencionadas. Algunos ejemplos de hipótesis se describen en la Tabla 2.

Tabla 2. Ejemplos de hipótesis que pueden buscar comprobarse durante el proceso de validación

Hipótesis	Fuente de evidencia
Los contenidos evaluados coinciden con los objetivos de aprendizaje. Las preguntas utilizadas conforman una muestra adecuada del posible universo de preguntas sobre los mismos temas. La ponderación de cada tema evaluado es correcta. El nivel de complejidad de los ítems utilizados es adecuado con respecto al nivel de estudios.	Evidencia basada en el contenido de la prueba.
Los procesos mentales que llevaron a cabo los estudiantes son los esperados.	Evidencia basada en los procesos de respuesta.
El examen es unidimensional (si eso era lo planeado). Si es multidimensional, la calificación de cada dimensión es consistente con el constructo evaluado. La confiabilidad de los resultados es adecuada. Si se utilizan las puntuaciones de diferentes partes de la prueba como sub-puntuaciones, se especifica y se justifica cómo se combinan dichas sub-puntuaciones.	Evidencia basada en la estructura interna.
Existe relación entre los resultados de otras evaluaciones (que también poseen un nivel de confiabilidad adecuado) y la evaluación sumativa que examinan el mismo tema. Se toman en cuenta las variables confusoras que pueden sesgar los resultados para grupos específicos. Se demuestra que el desempeño en la prueba predice el desempeño en el criterio evaluado.	Evidencia basada en las relaciones con otras variables.
Las consecuencias negativas involuntarias no son graves y son menores a las consecuencias positivas.	Evidencia basada en las consecuencias de la prueba.

2. Evaluar las fuentes de evidencia

2a. Crear un plan para probar las hipótesis. Con base en las hipótesis seleccionadas, se buscan las fuentes de evidencia y se reúne la información correspondiente.

2b. Evaluar la evidencia y formular un juicio. En este último paso se evalúan todas las evidencias en orden y se establece el grado de validez de la interpretación de las puntuaciones de la prueba evaluada. Este grado dependerá de la calidad de las evidencias presentadas y también de las evidencias más importantes, según la prueba.

AMENAZAS A LA VALIDEZ

Además de analizar las fuentes de evidencia de validez, se sugiere identificar elementos que puedan afectar el grado de validez de los resultados de la evaluación. Este paso es importante porque da fortaleza a las decisiones que se toman con base en los resultados de la prueba. Los elementos que reducen el grado de validez se les denomina amenazas a la validez; se les llama así porque interfieren con la correcta interpretación de las puntuaciones (Carrillo-Ávalos et al., 2020; Downing y Haladyna, 2004). Estas amenazas pueden estar pren sentes en cualquier tipo de evaluación. En general, se reconocen dos tipos de amenazas a la validez: subrepresentación del constructo (SC) y varianza irrelevante al constructo (VIC) (Downing, 2003; Messick, 1989). Para explicar estos conceptos, partiremos de la teoría clásica de los tests (TCT). De acuerdo con esta, la puntuación que se obtiene a partir de las respuestas de un examen (puntuación observada o total = X) está compuesta por la suma de la puntuación verdadera (V) más el error aleatorio (E_a) (de Champlain, 2010; Schuwirth y van der Vleuten, 2011):

$$X = V + E_a$$

De esta manera, los factores que tienen un efecto sistemático sobre la puntuación observada "X" se agregan a la puntuación verdadera "V". Estos factores incluyen tanto al constructo de interés, como a otros factores sistemáticos que se presentan en todas las versiones del examen, pero cuya medición no es el objetivo; por ejemplo, que un reactivo de opción múltiple no tenga una respuesta correcta entre las opciones, por lo que todos los alumnos la contestarán mal. Por otra parte, el error aleatorio (E_a) está conformado por todas las condiciones que aportan variabilidad a la puntuación verdadera "V" de manera no sistemática, ya que son diferentes en cada ocasión que se aplica el examen y para cada persona, como el cansancio, el desvelo y el estrés. En realidad, no conocemos el valor del error aleatorio ni el de la puntuación verdadera, aunque con diversos métodos estadísticos se pueden estimar (Haladyna y Downing, 2004).

Con base en lo anterior, la puntuación verdadera se puede dividir en dos partes: la puntuación del constructo de interés (θ), más la puntuación causada por factores sistemáticos (E_s). Por lo tanto, la fórmula quedaría así (Haladyna y Downing, 2004):

$$X = \theta + E_s + E_a$$

A partir de esta fórmula podemos definir los dos tipos de amenazas a la validez. Una de ellas se presenta cuando se mide θ a través de ítems que representan *de forma incompleta* el dominio del conocimiento que queremos evaluar: esto es la subrepresentación del constructo. Por otro lado, la varianza irrelevante al constructo se asocia con el error sistemático E_s, que es causado por la medición involuntaria de elementos que no son el objetivo de la medición, e interfieren con la medición del constructo de interés y, por lo tanto, con la validez de la interpretación de los resultados (Downing y Haladyna, 2004; Haladyna y Downing, 2004; Messick, 1989). La varianza indeseable de X debida a E_a también aporta elementos para conformar una amenaza a la validez; sin embargo, sus factores conllevan a una confiabilidad baja, de lo que hablaremos en la siguiente sección de este capítulo.

A continuación, se describen ejemplos de los tipos de amenazas a la validez:

- **Subrepresentación del constructo (SC).** En el ejemplo del examen sumativo de biología evolutiva mencionado, supongamos que consta de 10 preguntas, que constituyen una muestra del universo de todas las preguntas posibles sobre el mismo tema y que miden el mismo constructo. Se considera que 10 ítems son pocos para evaluar a cabalidad los temas involucrados, por lo que no son suficientes para representar adecuadamente al constructo de interés ni a las dimensiones que deseamos evaluar. Otros problemas que podríamos encontrar en esta muestra es que se estén evaluando contenidos triviales, contenidos diferentes a los estudiados, que solo pregunten aspectos de un solo tema o que la confiabilidad sea insuficiente. Es evidente que estos problemas están relacionados con la evidencia de validez de contenido. Si aplicáramos este examen de 10 reactivos, las amenazas a la validez pueden ser: número insuficiente de preguntas, sesgo, concordancia del dominio, baja confiabilidad (Moreno et al., 2004).
- **Varianza irrelevante al constructo.** Esta amenaza tiene origen en el error sistemático E_s causado por una variable irrelevante al constructo de interés. Por ejemplo:

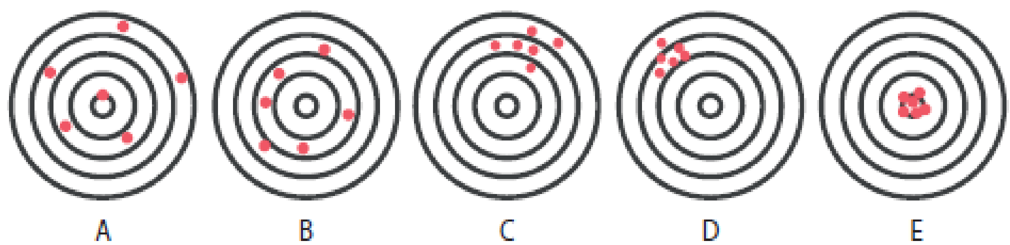
  - *Ítems defectuosos.* La escritura de un reactivo de opción múltiple de calidad debe cumplir con los estándares establecidos por expertos para evitar que aporten varianza irrelevante al constructo (Haladyna et al., 2002; National Board of Medical Examiners, 2016). La baja calidad de una pregunta la hace más difícil de contestar, convirtiéndola en un ítem defectuoso. Otros defectos posibles son que entre las opciones de respuesta se encuentre "todas las anteriores" o "ninguna de las anteriores", o el número de opciones (varios expertos consideran que el número óptimo es de tres a cuatro) (Abad et al., 2001; Haladyna et al., 2019; Rodríguez, 2005).
  - *Formato en negativo.* Varios expertos consideran que escribir preguntas en formato negativo debe evitarse porque se corre el riesgo de escribir un doble negativo (la pregunta y una o más de las opciones contienen términos negativos, por lo que identificar las opciones incorrectas implica negarlas, o sea, indicar que no son correctas) (Chiavaroli, 2017). También existe el riesgo de que el alumno no reconozca la negación en la pregunta, aunque se destaque en negritas "no" o "excepto", y que el proceso de respuesta llevado a cabo para responder no sea el esperado.
  - *Lenguaje.* Una pregunta que presenta una viñeta demasiado extensa o con explicaciones innecesarias ocasiona que el estudiante pierda tiempo leyéndola y en realidad puede evaluar la velocidad de lectura en lugar de medir el constructo deseado (Hicks, 2011; National Board of Medical Examiners, 2016). Por otro lado, las oraciones deben estar bien estructuradas, de manera que la pregunta quede clara.

- *Discordancia con el dominio*. Cuando no hay relación entre lo que se ha estudiado y lo que se está evaluando.
- *Ítems muy fáciles o difíciles*. Reactivos que seguramente todos los alumnos contestarán correctamente, lo que aumentará el promedio del grupo y el individual de manera artificial, lo que interfiere con la validez. Por otro lado, no discrimina entre estudiantes de bajo y alto desempeño.
- *Hacer trampa*. Este es uno de los retos principales de los profesores al aplicar exámenes, pues los estudiantes pueden hacer trampa de muchas formas y obtener ventaja de manera deshonesta: voltear a ver el examen de un compañero, utilizar un "acordeón", conocer las preguntas antes de hacer el examen, e incluso utilizar dispositivos digitales. La presencia de estas conductas genera falsos positivos, e incrementa la varianza sistemática en las puntuaciones.
- *Decisión indefendible de puntuación aprobatoria*. En general, en México los exámenes tienen una calificación mínima aprobatoria de 6.0. Si estos exámenes no son reproducibles o la distribución de las calificaciones es demasiado amplia (calificaciones demasiado altas y demasiado bajas), es difícil justificar por qué el 6.0 es la calificación de pase (Norcini, 2003).
- *Enseñar a la prueba ("teaching to the test")*. Cuando los alumnos han recibido preparación para contestar preguntas de un examen específico, como los exámenes de admisión a las licenciaturas o incluso los exámenes que elabora un profesor en particular, ello añade ventaja a algunos estudiantes y distorsiona el verdadero propósito de la enseñanza (Bond, 2008). Se ha visto que en ocasiones los estudiantes acceden a bancos de preguntas y prefieren aprenderlas de memoria en lugar de aprender los contenidos de los temas estudiados. Este factor es una amenaza a la fuente de evidencia de validez de relación con otras variables, porque no será posible generalizar los resultados de la prueba con el resto de los ítems del universo que miden el constructo deseado.
- *Habilidad para responder exámenes (testwiseness)*. Los alumnos que han presentado una gran cantidad de exámenes escritos a lo largo de su vida, se convierten en expertos para identificar la respuesta correcta en un ítem defectuoso. Por ejemplo, el hecho de que con frecuencia la opción correcta suele ser la más larga, la que tiene mayor detalle y está mejor escrita. Los alumnos que desarrollan estas habilidades como estrategias para responder pruebas, ocasionan que las respuestas no demuestren con veracidad el constructo de interés. Otro concepto relacionado es la "adivinanza educada" (*educated guessing)*, que consiste en que contestan por eliminación de las opciones menos probables, mas no porque realmente sepan la respuesta (Jurado-Núñez y Leenen, 2016).

CONFIABILIDAD

La confiabilidad es la característica de las evaluaciones que se refiere a que los puntajes sean consistentes de persona a persona, de instrumento a instrumento y de un conjunto de ítems a otro dentro del mismo universo de ítems (Cizek, 2009). Va de la mano de la validez en cuanto a que depende de la interpretación de las puntuaciones de la prueba, y forma parte de la fuente de evidencia de validez basada en la estructura interna (Tavakol y Dennick, 2011). Sin embargo, una prueba puede ser confiable, pero tener validez limitada. Tomemos como ejemplo un tiro al blanco con las marcas de los dardos, haciendo una analogía con las puntuaciones de una prueba de persona a persona, de instrumento a instrumento o de un conjunto de ítems a otro dentro del mismo universo de ítems (Figura 2).

Figura 2. Esquema de "tiro al blanco" para visualizar los conceptos de validez y confiabilidad



A. Prueba no confiable, sin validez.

B. Confiabilidad regular, validez regular.

C. Prueba confiable, pero sin validez.

D. Prueba muy confiable, pero sin validez.

E. Prueba confiable y con validez.

En el inciso A las marcas están muy dispersas en una prueba que no es confiable, porque no se concentran cerca del mismo sitio cada vez, ni dan en el centro (que es lo que se acerca más a lo que se desea evaluar). En el inciso B las marcas demuestran regular confiabilidad, pues no están tan dispersas, y validez regular porque, aunque algunas sí están en el centro, otras no. El inciso C demuestra una prueba confiable, pero con poca validez, ya que las marcas se concentran alrededor del mismo sitio, pero no en el centro, al igual que en el inciso D, aunque en este último están más cercanas entre sí. Finalmente, el inciso E demuestra lo deseable en cuanto a confiabilidad y validez: todas las marcas están concentradas en el centro.

¿Cómo se mide la confiabilidad? Los coeficientes de confiabilidad expresan la relación entre los puntajes que obtiene el mismo estudiante cuando presenta el mismo examen en dos ocasiones diferentes o en dos partes del mismo examen (Downing, 2004; Fraenkel et al., 2019). Esta relación da una idea de cuánta variación se puede esperar entre las diferentes ocasiones, de persona a persona o de muestra de ítems a muestra de ítems. A continuación se describen algunas fórmulas para obtenerlos y adquirir información acerca de la consistencia interna de la prueba:

- *Kuder-Richardson*. Se utiliza para pruebas con variables dicotómicas (correcto/incorrecto), y otorga información sobre qué tan bien la prueba mide el constructo de interés (Kuder y Richardson, 1937). para conocer el valor de este coeficiente se pueden utilizar dos fórmulas:
  - KR20 – para pruebas con ítems con dificultad variable.

$$KR20 = \left(\frac{n}{n-1}\right)\left(\frac{1-\sum p \cdot q}{var}\right)$$

n= número de ítems

p= proporción de personas que aprueban el ítem

q= proporción de personas que reprueban el ítem

var= varianza para la prueba

  - KR21 – para pruebas con ítems con la misma dificultad.

$$KR21 = \left(\frac{n}{n-1}\right)\left(1-\frac{M(n-M)}{n \cdot var}\right)$$

n= número de ítems

M= media de la puntuación para la prueba

var= varianza de la prueba.

- *Alfa de Cronbach*. Es conocido como el coeficiente alfa, y se utiliza para pruebas con ítems con más de dos respuestas correctas, por ejemplo, preguntas con respuestas en forma de escala de valoración tipo Likert (Tavakol y Dennick, 2011).

$$\alpha = \frac{N\bar{c}}{\bar{v}+(N-1)\bar{c}}$$

$\propto$= alfa de Cronbach

N= número de ítems

$\bar{c}$= covarianza promedio inter-ítem

$\bar{v}$= varianza promedio

Los valores de los coeficientes de confiabilidad van de 0.00 a 1.00, donde 0 corresponde a una ausencia completa de relación y 1.00 el máximo posible de la relación (Fraenkel et al., 2019).

En el caso de pruebas de mediano impacto, si el valor es menor a 0.5, la consistencia interna es no aceptable; si es de 0.5 a 0.59, es pobre; si es de 0.6 a 0.69, es cuestionable; si es de 0.7 a 0.79, es aceptable; si es de 0.8 a 0.89, es buena; y mayor a 0.9 es excelente. Si la prueba es de alto impacto, el valor mínimo deseado es de 0.8 (Bland y Altman, 1997; Downing, 2004).

CONCLUSIONES

La validez es uno de los conceptos más importantes en evaluación del y para el aprendizaje. El modelo actual de validez es el resultado de una gran cantidad de investigaciones y discusiones entre expertos en el tema, y se le considera como un concepto holístico que se alimenta de diversas fuentes (contenido, proceso de respuesta, estructura interna, relación con otras variables y consecuencias). Validez se refiere a las inferencias que pueden hacerse con los resultados, más que a las pruebas o exámenes *per se*. Durante el proceso de evaluación deben cuidarse las amenazas a la validez, tanto de subrepresentación como de varianza irrelevante al constructo. La confiabilidad o reproducibilidad de los resultados, se integra como un elemento de fuente de evidencia de validez en el rubro de estructura interna.

REFERENCIAS

Abad, F. J., Olea, J., y Ponsoda, V. (2001). Analysis of the optimum number alternatives from the Item Response Theory. *Psicothema*, *13*(1), 152–158.

American Educational Research Association, American Psychological Association, & National Council on Measurement in Education. (2018). *Estándares para pruebas educativas y psicológicas* (Original w). American Educational Research Association.

Bland, J., y Altman, D. (1997). Statistics notes: Cronbach's alpha.pdf. *British Medical Journal*, *314*, 572.

Bond, L. (2008). Teaching to the Test: Coaching or Corruption. *New Educator*, *4*(3), 216–223. https://doi.org/10.1080/15476880802234482

Campbell, D., & Fiske, D. (1959). Convergent and discriminant validation by the multitrait-multimethod matrix. *Psychological Bulletin*, *56*(2), 81–105.

Carrillo-Avalos, B. A., Sánchez-Mendiola, M. y Leenen, I. (2020). El concepto moderno de validez y su uso en educación médica. *Revista de Investigación en Educación Médica*, *9*(33), 98–106. https://doi.org/https://doi.org/10.22201/facmed.20075057e.2020.33.19216

Carrillo-Avalos, B. A., Sánchez-Mendiola, M., y Leenen, I. (2020). Amenazas a la validez en evaluación: implicaciones en educación médica. *Investigación en Educación Médica*, *9*(34), 100–107. https://doi.org/10.22201/facmed.20075057e.2020.34.221

Chiavaroli, N. (2017). Negatively-worded multiple choice questions: An avoidable threat to validity. *Practical Assessment, Research and Evaluation*, *22*(3), 1–14.

Cizek, G. J. (2009). Reliability and validity of information about student achievement: Comparing large-scale and classroom testing contexts. *Theory into Practice*, *48*(1), 63–71. https://doi.org/10.1080/00405840802577627

Cook, D. A., y Hatala, R. (2016). Validation of educational assessments: a primer for simulation and beyond. *Advances in Simulation*, *1*(1), 1–12. https://doi.org/10.1186/s41077-016-0033-y

Cronbach, L. J., y Meehl, P. E. (1955). Construct validity in psychological tests. *Psychological Bulletin*, *52*(4), 281–302. https://doi.org/10.1037/h0040957

de Champlain, A. F. (2010). A primer on classical test theory and item response theory for assessments in medical education. *Medical Education*, *44*(1), 109–117. https://doi.org/10.1111/j.1365-2923.2009.03425.x

Downing, S. M. (2003). Validity: On the meaningful interpretation of assessment data. *Medical Education*, *37*(9), 830–837. https://doi.org/10.1046/j.1365-2923.2003.01594.x

Downing, S. M. (2004). Reliability: on the reproducibility of assessment data. Medical education, *38*(9), 1006–1012. https://doi.org/10.1111/j.1365-2929.2004.01932.x

Downing, S. M., y Haladyna, T. M. (2004). Validity threats: Overcoming interference with proposed interpretations of assessment data. *Medical Education*, *38*(3), 327–333. https://doi.org/10.1046/j.1365-2923.2004.01777.x

Fraenkel, J. R., Wallen, N. E., & Hyun, H. H. (2019). *How to Design and Evaluate Research in Education* (10th ed.). Mc Graw-Hill Education.

Haladyna, T. M., y Downing, S. M. (2004). Construct-Irrelevant Variance in High-Stakes Testing. *Educational Measurement: Issues and Practice*, *23*(1), 17–27. https://onlinelibrary.wiley.com/doi/abs/10.1111/j.1745-3992.2004.tb00149.x

Haladyna, T. M., Downing, S. M., y Rodriguez, M. C. (2002). A Review of Multiple-Choice Item-Writing Guidelines for Classroom Assessment. *Applied Measurement in Education*, *15*(3), 309–334. https://doi.org/10.1207/s15324818ame1503_5

Haladyna, T. M., Rodriguez, M. C., & Stevens, C. (2019). Are Multiple-choice Items Too Fat? *Applied Measurement in Education*, *32*(4), 350–364. https://doi.org/10.1080/08957347.2019.1660348

Hicks, N. A. (2011). Guidelines for identifying and revising culturally biased multiple-choice nursing examination items. *Nurse Educator*, *36*(6), 266–270. https://doi.org/10.1097/NNE.0b013e3182333fd2

Instituto Nacional para la Evaluación de la Educación. (2017). Criterios técnicos para el desarrollo, uso y mantenimiento de instrumentos de evaluación. En *Diario Oficial de la Federación*. https://www.inee.edu.mx/wp-content/uploads/2019/04/P1E104.pdf

Jurado-Núñez, A., y Leenen, I. (2016). Reflexiones sobre adivinar en preguntas de opción múltiple y cómo afecta el resultado del examen. *Investigación en Educación Médica*, *5*(17), 55–63. https://doi.org/10.1016/j.riem.2015.07.004

Kane, M. T. (2009). Validating the interpretations and uses of test scores. In R. W. Lissitz (Ed.), *Validity: Revisions, New Directions and Applications* (pp. 39–64). Information Age Publishing, Inc.

Kuder, G. F., & Richardson, M. W. (1937). The theory of the estimation of test reliability. *Psychometrika*, *2*(3), 151–160.

Lane, S. (2014). Evidencia de validez basada en las consecuencias del uso del test. *Psicothema*, *26*(1), 127–135. https://doi.org/10.7334/psicothema2013.258

Leenen, I. (2014). Virtudes y limitaciones de la teoría de respuesta al ítem para la evaluación educativa en las ciencias médicas. *Investigación en Educación Médica*, *3*(9), 40–55.

Messick, S. (1989). Validity. In R. L. Linn (Ed.), *Educational Measurement* (pp. 13–103). MacMillan. https://doi.org/10.1002/j.2330-8516.1987.tb00244.x

Moreno, R., Martínez, R. J., y Muñiz, J. (2004). Directrices para la construcción de ítems de elección múltiple. *Psicothema*, *16*(3), 490–497. https://www.redalyc.org/articulo.oa?id=72716324

National Board of Medical Examiners. (2016). *Cómo elaborar preguntas para evaluaciones escritas en las áreas de ciencias básicas y clínicas*. *3*, 1–98.

Norcini, J. J. (2003). Setting standards on educational tests. *Medical Education*, *37*(5), 464–469. https://doi.org/10.1046/j.1365-2923.2003.01495.x

Padilla, J. L., & Benítez, I. (2014). Validity evidence based on response processes. *Psicothema*, *26*(1), 136–144. https://doi.org/10.7334/psicothema2013.259

Rios, J., y Wells, C. (2014). Evidencia de validez basada en la estructura interna. *Psicothema*, *26*(1), 108–116. https://doi.org/10.7334/psicothema2013.260

Rodriguez, M. C. (2005). Three options are optimal for multiple-choice items: A meta-analysis of 80 years of research. *Educational Measurement: Issues and Practice*, *24*(2), 3–13. https://doi.org/10.1111/j.1745-3992.2005.00006.x

Schuwirth, L. W. T., & van der Vleuten, C. P. M. (2011). General overview of the theories used in assessment: AMEE Guide No. 57. *Medical Teacher*, *33*(10), 783–797. https://doi.org/10.3109/0142159X.2011.611022

Sireci, S., & Faulkner-Bond, M. (2014). Validity evidence based on test content. *Psicothema*, *26*(1), 100–107. https://doi.org/10.7334/psicothema2013.256

Tavakol, M., & Dennick, R. (2011). Making sense of Cronbach's alpha. *International Journal of Medical Education*, *2*, 53–55. https://doi.org/10.5116/ijme.4dfb.8dfd

Routledge
Taylor & Francis Group

Teaching and Learning in Medicine
An International Journal

ISSN: (Print) (Online) Journal homepage: https://www.tandfonline.com/loi/htlm20

Bridging Validity Frameworks in Assessment: Beyond Traditional Approaches in Health Professions Education

Blanca Ariadna Carrillo-Avalos, Iwin Leenen, Juan Andrés Trejo-Mejía & Melchor Sánchez-Mendiola

To cite this article: Blanca Ariadna Carrillo-Avalos, Iwin Leenen, Juan Andrés Trejo-Mejía & Melchor Sánchez-Mendiola (18 Dec 2023): Bridging Validity Frameworks in Assessment: Beyond Traditional Approaches in Health Professions Education, Teaching and Learning in Medicine, DOI: 10.1080/10401334.2023.2293871

To link to this article: https://doi.org/10.1080/10401334.2023.2293871

Published online: 18 Dec 2023.

TEACHING AND LEARNING IN MEDICINE
https://doi.org/10.1080/10401334.2023.2293871

VALIDATION

Bridging Validity Frameworks in Assessment: Beyond Traditional Approaches in Health Professions Education

Blanca Ariadna Carrillo-Avalos[a], Iwin Leenen[b], Juan Andrés Trejo-Mejía[c] and Melchor Sánchez-Mendiola[c,d]

[a]Faculty of Medicine, Autonomous University of San Luis Potosí (UASLP), San Luis Potosí, Mexico; [b]Faculty of Psychology, National Autonomous University of Mexico (UNAM), Mexico City, Mexico; [c]Faculty of Medicine, UNAM, Mexico City, Mexico; [d]Educational Innovation and Distance Education, UNAM, Coordination of Open University, Mexico City, Mexico

ABSTRACT
Construct: High-stakes assessments measure several constructs, such as knowledge, competencies, and skills. In this case, validity evidence for test scores' uses and interpretations is of utmost importance, because of the consequences for everyone involved in their development and implementation. ***Background:*** Educational assessment requires an appropriate understanding and use of validity frameworks; however, health professions educators still struggle with the conceptual challenges of validity, and frequently validity analyses have a narrow focus. Important obstacles are the plurality of validity frameworks and the difficulty of grounding these abstract concepts in practice. ***Approach:*** We reviewed the validity frameworks literature to identify the main elements of frequently used models (Messick and Kane's) and proposed linking frameworks including Russell's recent overarching proposal. Examples are provided with commonly used assessment instruments in health professions education. ***Findings:*** Several elements in these frameworks can be integrated into a common approach, matching and aligning Messick's sources of validity with Kane's four inference types. ***Conclusions:*** This proposal to contribute evidence for assessment inferences may provide guidance to understanding the use of validity evidence in applied settings. The evolving field of validity research provides opportunities for its integration and practical use in health professions education.

ARTICLE HISTORY
Received 25 June 2023
Revised 12 November 2023
Accepted 28 November 2023

KEYWORDS
Construct validity; health professions education; educational measurement

Introduction

High-stakes assessments in health professions education are used to deliver results with important consequences for examinees, teachers, and institutions, requiring a complex collaboration of everyone involved. Examples of these types of examinations are tests included in admission or certification processes, like the Medical College Admission Test, the United States Medical Licensing Exam, Board Certification exams, and their equivalents in other countries. Validity studies involving admission and licensing exams for health professions education primarily focus on its short and long-term predictive features. Despite their relevance, frequently they have shortcomings as other important aspects of validity are not considered, such as consequential validity. Often, this oversight arises from not using an integrative framework that comprehensively organizes these key validity aspects.[1–10]

In general, scores interpretation of summative assessments needs support (i.e., validation) for the intended use of the test, which is a joint responsibility of the test developer and the test user;[11] the higher the stakes, the more validity evidence is required, in quality and quantity. Validation creates trust in learners and organizations since it provides evidence that the conception, design, and development of the instrument have been carried out adequately. This includes considering the measured construct, the appropriate application of the instrument, and documentation that test results accurately reflect the candidate's performance level and, consequently, that decision-making will be fair.[12] Health professions educators (HPEs) vary greatly in their use of the term "validity", which leads to discrepancies in validity practices and processes. According to the Standards for Educational and Psychological Testing, validity is a property of the test that reflects whether

CONTACT Melchor Sánchez-Mendiola melchor_sanchez@cuaieed.unam.mx, melchorsm@unam.mx Faculty of Medicine, CUAIEED, UNAM, Anexo Centro Ciencias de la Complejidad. 1er Piso. Circuito Cultural S/N, Ciudad Universitaria. Del. Coyoacán. Ciudad de México, CdMx 04510. México.

interpretations of test scores are supported by evidence.[11] Understanding validation and its uses and interpretations requires making this process more accessible for all stakeholders (researchers, administrators, students, and their parents).[13,14]

The lack of routine use of a clear validity framework prevents a precise understanding of how to improve high-stakes tests and whether interpretations of exam results are correct, and also how to defend the decisions based on those interpretations. Our goal in this paper is to continue the discussion about validity and its frameworks: their importance, commonalities, and suggestions about how to use them. We briefly describe Messick's and Kane's frameworks as well as the more recent integrated approach proposed by Russell. Then, we match Messick and Kane's frameworks, followed by the incorporation of Russell's approach. To provide scaffolding for the frameworks' description, we will use two high-stakes examples as a continuous thread throughout the paper: a multiple-choice question (MCQ) admission exam, and a licensing OSCE test, providing examples of validity evidence for each framework.

Current approaches to validity

Arguably, the most widely known contemporary validity frameworks are Messick's and Kane's; this is likely related to the fact that the Standards for Educational and Psychological Testing,[11] a leading reference jointly published by the American Educational Research Association, the American Psychological Association, and the National Council on Measurement in Education, heavily relies on these frameworks. Messick's and Kane's frameworks emerged from the traditional view on validity, which comprised content, construct, and criterion validity (with the latter including concurrent and predictive validity).[15]

Messick's framework

Arguing that assessments aim at measuring constructs, in 1989 Messick[16] concluded that all kinds of validity fall under the umbrella of construct validity. A construct is a latent (i.e., not directly observable) attribute,[11,15] such as anatomy knowledge, clinical ability, and teamwork competence, about which inferences can be made by measuring responses to challenges *via* specifically designed instruments.[17] This standpoint is thoroughly described in the Standards for Education and Psychological Testing.[11] Here we present a brief explanation.

As a first step for validation through Messick's framework, the interpretation or intended use of the test results should be specified. Subsequently, empirical data as well as theoretical arguments may provide support or evidence for the interpretations or uses at a specific time in a specific population.[16,17] Messick considers five sources of validity evidence. The sources that are explored in a particular case depend on the test's goals and proposed uses.[11,17,18] For instance, the validity evidence required for high-stakes evaluations, like admission or certification exams, differs in type and quality from the evidence required for a formative quiz.[11]

The first source of validity evidence in Messick's framework is content-based evidence, which is related to the extent of agreement between the testing purposes and the content of the instrument. This source of evidence has four elements: domain definition (the intended construct to be measured), domain representation or alignment (the degree to which the instrument represents and measures the defined domain), domain relevance (the degree to which each item is relevant for the measured domain), and the appropriateness of instrument development (the processes through which the instrument is built to ensure that the content fully represents the intended construct and that all content is relevant). For an MCQ test, supporting evidence may be obtained from an analysis of the test blueprint and test specifications, the item representativity, the matching of items' content with the test specifications, and the relationship of test content with the tested knowledge (see Table 1).[11,19] For an OSCE, there is need for experts to compare the curriculum objectives and test specifications with the test blueprint, and to verify the level of competence to be assessed (e.g., "show how" on Miller's pyramid). It is also important to ensure the test authors' expertise, clarity of instructions, and fidelity of standardized patients (SP) portrayals.[20–22]

Evidence based on response processes, the second source of evidence in this framework, focuses on the theoretical assumptions—as implied by the construct the test aims to measure—about the cognitive processes involved in solving the task or item. Ideally, cognitive theory should guide item generation, selection, and parameters during test development, as cognitive models specify the main processes used in item solving.[23] Empirical data may provide evidence of a match between the processes that are actually performed during item solving, and the processes theoretically delineated in test specifications. Often a distinction is made between direct and indirect evidence. Examples of the former for an MCQ test include data from cognitive interviews, interviews with aloud-thinking protocols, and focus groups (Table 1). Indirect evidence may rely on an analysis of response times (often by using advanced psychometric models that formalize the relationship between the number, type, and complexity of the cognitive processes involved in solving the item and the time needed to respond) or data from eye movement tracking during item solving.[24] Downing[17] also includes in this category the test-taker's familiarity with the test format and the response sheet, key validation, quality control, and clarity of instructions.[25,26] For an OSCE, it is important to ensure the applicants are familiar with the test format through detailed orientation and that the data entry is accurate. A description and evaluation of score calculation and report methods may also be necessary.[20,21]

The third source of evidence is based on internal structure, it refers to the degree to which item relationships are aligned with the theory behind the measured construct and usually includes the following three aspects: dimensionality, measurement invariance, and reliability.[11] Dimensionality refers to the number and nature of latent dimensions or subconstructs that are being measured; in particular, how many dimensions are being measured, inter-dimension relationships, and the relationship between dimensions and items. Tools to study dimensionality in an MCQ test include confirmatory factor analysis (CFA) and analysis based on item response theory (IRT) (Table 1). Measurement invariance assumes that scores are expected to be the same when compared between groups with the same level of the underlying construct, although they differ on other characteristics, such as race, age, or sex.[27–29] Measurement invariance is usually examined by tests of differential item functioning.[30] Reliability refers to the reproducibility of the test results, frequently measured with Cronbach's alpha.[31,32] On the other hand, OSCE's internal structure may also be explored through an analysis of interrater reliability, interstation correlation, and, if interpersonal skills are assessed, the correlation of ratings between raters and SP. High reliability and generalizability coefficients are needed too (>0.80 for both).[20,21]

The fourth source of validity evidence studies the relationship between test results and external variables and is often organized along two axes: the first distinguishes convergent versus divergent relations, which boils down to showing high correlations with other tests that measure the same or similar constructs versus showing low or zero correlations with tests that measure a different construct. Second, the distinction between concurrent versus predictive designs focuses on outcomes that are simultaneously available with the test results (e.g., existing expert judgments as validity evidence for a diagnostic test) versus future outcomes (e.g., academic success for an MCQ admission test). As we mentioned, many publications about admission tools are focused on their predictive validity over academic performance. Conversely, this source of evidence for an OSCE as licensing test could come from correlation with concurrent assessments that measure the same or similar level of competence, or with the performance of licensed health professionals.[20,21,33] A tool to study simultaneously the relations with multiple other variables is the multitrait-multimethod matrix.[30]

Messick's fifth source of validity evidence, based on test consequences, evaluates the impact that the uses and interpretation of test results have on applicants, institutions, society, and other stakeholders. It is related to possible errors in the interpretation of test results, the likelihood that the results may give rise to false positive or false negative decisions, as well as whether positive consequences exceed involuntary negative consequences. For instance, a positive consequence of an admission exam can be the accurate selection of successful students, and a negative consequence could be an increase in mental health issues in applicants. This type of evidence is often collected retrospectively or longitudinally. Evidence for MCQ tests and OSCE can be obtained through focus groups and interviews for applicants and professors, as well as action theory, to identify critical components of academic programs and their impact points.[20,21,34]

Kane's framework

Although Messick's framework has been widely used, Kane and other authors[35] consider Messick's framework difficult to execute, as it lacks clear guidelines on how to begin and how to prioritize among evidence sources or types of assessments. To address this issue, Kane proposed an argument-based approach to validation.[36,37] His approach implies two steps. The first step requires establishing the interpretation/use as an argument (IUA), where *interpretation* refers to the explanation of the test score meaning, whereas *use* refers to the decisions taken based on the score. Based on a clear understanding of the assessed construct, the IUA establishes a network of inferences where, based on test scores' use and interpretations for a specific population and context, a set of inferences is developed. These include inferences about

(a) the observed test score (scoring inference), (b) the generalized test score (generalization inference), (c) the meaning of the test score for real-life performance (extrapolation inference), and (d) its implications for decision making (implications inference).[35,38–40]

The meaning of the four inferences is as follows: the scoring inference refers to the appropriateness of score criteria and, particularly, the rules to convert responses into one or more final scores. The generalization inference assumes that the components (e.g., items or stations) of a test are an adequate (i.e., representative) sample from the universe of possible components and, hence, that the observed score provides a good estimate of the individual's universe score which they would have obtained if all relevant items were administered.[41] If score interpretations can spread toward other domains and develop a prediction of the applicant's test result in a different context, this is evidence for the extrapolation inference. Finally, the impact of score interpretations on the applicant, their family, society, and other stakeholders, are considered within the implications inference. Assumptions or claims for each inference are stated in the IUA and can be expressed in the form of hypotheses for which support is sought (Table 1).

The second step, after establishing the IUA, is to establish the validity argument, where evidence is collected to prove the hypotheses for each inference. For example, evidence for the scoring inference for an OSCE can be found through experts' judgments about the scoring criteria and proving that score rules or rubrics were correctly applied and unbiased (e.g., by evaluating interrater reliability and training of raters), training SP, analyzing items and reliability within each station, and ensuring security and quality.[42] For an MCQ admission test, the following are useful: items and response options' performance, scoring rubric, data security, and quality control.[35]

Generalizability theory is commonly used to provide support for the generalization inference, as it allows researchers to quantify the generalizability of the observed scores to the universe scores. In the case of an OSCE, it is important to ensure an appropriate sample of the domain through the use of a blueprint and to perform tests on internal consistency and interrater reliability.[18,42] Reliability, test blueprint, sample size, and item response theory are some of the sources of evidence for an MCQ test.[35]

The extrapolation inference could rely on a regression analysis that provides evidence of the predictive power of the test score with some criterion of interest. The comparison of experts to a novice group (e.g., licensing students vs first years), demonstration of correlations with other tests that measure the same construct, qualifications of stations authors to develop authentic cases, and the relevance to real-life tasks, are sources of evidence for an OSCE.[42] In the case of an MCQ, previous domain specification, construct definition, correlation with another assessment that measures the same or a similar construct, and differential item functioning, are sources to be analyzed.[35]

The implications inference should be addressed by ensuring that the positive consequences outweigh the negatives. For this inference, an OSCE, as well as an MCQ test, rely on a standard setting method and process (e.g., for a virtual format, it must be ensured that it will continue to be a high-quality assessment and that it can achieve the original interpretations and uses), analysis of pass-fail consequences, and exploration of how it influences curriculum and learning.[42,43]

Validity of interpretation or use of test scores is established as being credible and appropriate at a certain point in time and for a specific context, so, if the use or the interpretation were to change, a new validation effort is required.[39] Cook's review is recommended to readers interested in a deeper analysis of Kane's framework in health professions education.[35]

Russell's integrative approach

Other authors have published different viewpoints regarding validity that contrast with Messick's and Kane's framework. For instance, Borsboom et al.[44] agree with Kane's view about the uses and interpretations of tests, although they argue that validity is a test property that measures an attribute and requires that variations in the attribute cause variations in the test results. This is why in their view there is no need for a unified validity concept. They also consider that most of the validation efforts should be made during the instrument design phase: knowing exactly what it is we want to measure allows us to know how to measure it. Cizek[45] emphasizes the difference between uses and interpretations of test scores and argues that validity evidence based on consequences should be used to justify its uses, whereas the other four sources of evidence should confirm the results' interpretation. Sireci[46] adds that scores always have both an interpretation and a use, which he exemplified by asking: "Why would a doctor order an X-ray, make an interpretation of the image and then do nothing about it?" Validity evidence is required for use as well as for interpretation of test scores.

Recently, Russell[47] proposed a novel framework with three stages, according to the time the evidence is produced during the instrument's development and utilization. The model is consistent with the work of Messick, Kane, Sireci, Borsboom, and Cizek. He considers the sequence of actions while implementing an assessment activity. These begin with the *purpose of the instrument*, which initiates with a test that aims to measure a construct; this measurement generates scores that support an inference about the construct. Afterwards, the *use of scores* depends on the interpretation (informed by the inference about the measured construct) and the subsequent decision(s) that lead to consequences. Based on this sequence, he recommends to split validation into three stages: instrument purpose validity, verification of interpretation and decision, and utility of actions.

Russell's first stage involves questioning whether the instrument fulfills its purpose by supporting the inferences about the construct it intends to measure. To address this question, evidence should be gathered about the psychometric properties of the scores as well as about the correct representation of the construct, which implies avoiding construct underrepresentation (CU) and construct irrelevant variance (CIV). Also, there should be evidence that demonstrates that the scores possess the quantitative properties that a measure should have, especially additivity. This stage should be explored by the instrument developer through field testing.

The second stage, verification of interpretation and decision, entails an analysis of whether the interpretation based on the scores is appropriate and, if so, whether this interpretation can truly inform a decision about the subsequent action(s). This stage of validation can be performed after the instrument is applied by the score user. It implies an analysis of whether scores are accurate and can be used to sort test takers into categories, and if the decisions are appropriate for each category. The degree of correlation between the decisions about each group and their abilities or needs should be verified. It is important to note that this stage does not include actions or their consequences.

The third and last stage involves evaluating the utility of the actions, i.e., their consequences, voluntary or involuntary, and if they are positive or negative. Explicit efforts should be made to understand the origin and implications of negative consequences and to avoid them as much as possible. This stage should be assessed by the score users (Table 1).

Bridging validity approaches

Integrating Messick and Kane

An issue in Messick's and Kane's frameworks is that they provide limited guidance on how to validate interpretations and use of inferences for practitioners. We looked for correspondences between these frameworks, not only in terms of evidence sources and inferences but also in terms of the theory underlying each framework. In particular, we looked for a match of Messick's sources of evidence with Kane's inferences and their warrants. We visualized how a specific source of evidence can contribute to multiple types of inferences, and created a plan to establish the links among sources and inferences. Now, we describe how the sources of validity evidence in Messick's framework can contribute to a validity argument for inferences in Kane's framework.

The scoring inference depends on appropriate scoring procedures, free of bias and performing as intended. Three sources of evidence may support this inference: (1) Content, through its elements of domain definition, domain relevance, and the appropriateness of instrument development; (2) Response processes, to demonstrate that scores reflect the appropriate cognitive process; and (3) Internal structure, through dimensionality and measurement invariance (Table 1).

A generalization inference involves making claims about whether the test scores are useful as estimates of performance in a larger domain, that is, on different occasions and varying conditions. Messick's sources of evidence that may support this inference are: (1) Content, through domain representation as alignment, which is concerned with the relationship between the learning goals and the test content; (2) Internal structure, through generalizability theory, which allows for generalization to the rest of the population; and (3) Relationship with other variables, through analysis of convergent or divergent correlations with other variables, supporting or not generalization to a larger domain.

The extrapolation inference requires evidence that from the universe of scores, the interpretation may be extrapolated to real-world situations. In Messick's framework this evidence may be based on (1) Content, as domain relevance implies that the instrument accurately covers the content described in the test blueprint, as well as the subcategories and subclassifications of the content; and (2) Relationship with other variables, considering test-criterion relationships, such as a related real-world assessment.

TABLE 1. Validation process in Russell's stages using Messick's sources of evidence as Kane's inferences warrants, with a medical education assessment example (MCQ test and an OSCE).

Russell	Kane	Messick			
Stage	Inference and hypotheses	Source of evidence	Element	Source of evidence for an MCQ test	Source of evidence for an OSCE
I. Instrument purpose validity (before using the instrument)	**Scores** • Rule of scoring is appropriate. • Good quality of wording and format of items/stations.	Content	Domain definition	• Test blueprint. • Course's learning objectives, construct definition, scoring criteria. • Description of the assessed content, its subcategories, and subclassifications, the proportion of questions and cognitive level, qualifications of item developers.	• Test blueprint. • Specification table based on the curriculum, and the epidemiology of the level of care. • Scoring rubric. • Observers training.
			Instrument development process	• Items properties. • Best practices for item writing and cultural revisions.	• Stations' properties. • Accuracy and authenticity of cases. • Cases of the core areas with patient scripts, evaluator guide, and student instructions.
			Domain relevance	• Items and dominion review. • Experts review the content to be assessed.	• Review validation by an expert committee.
	Generalizability • Items appropriately sample the dominion.	Content	Domain representation	• Prove that the number of items is an appropriate sample of the dominion. • Sample strategy. • Sample size.	• Prove that the number of stations is an appropriate sample of the dominion. • Sample strategy. • Sample size.
	Extrapolation • The test reflects fully a real-life task.	Content	Domain relevance	• Test blueprint (content and assessed cognitive level).	• The authenticity of the station.
II. Interpretation and decision verification (when/after using the instrument)	**Scores** • Scores reflect the appropriate cognitive process. • Instrument measured the planned dimension(s). • Scores are free from bias.	Response processes		• Cognitive interviews, and think-aloud testing.	• Use of holistic rubrics for each competency with four performance levels and detailed descriptors. • Observer accuracy and reliability.
		Internal structure	Dimensionality	• CFA • IRT	• CFA
			Measurement invariance	• EFA • DIF	• Identification of factors and indicators that explain the variance.
	Generalizability • Sample size was adequate. • There are no differences amongst groups. • Reliability. • Observed score relates to instruments that measure the same construct/does not relate to instruments that measure a different construct.	Internal structure	Measurement invariance Reliability	• G theory, linear and nonlinear models • Cronbach's alpha	• Discrimination analysis of the items. • Cronbach's alpha. • G coefficient.
		Relationship with other variables	Convergent-divergent	• Correlation with similar or different instruments: that measure a similar or a different construct.	• Correlation with an instrument that measures another construct (knowledge). A theoretical exam with multiple choice items in which competendes in the core areas are evaluated.
	Extrapolation • Observed score relates to a relevant criterion.	Relationship with other variables	Relation test-criterion	• Correlation of instrument with criterion (scores from another instrument, in a concurrent or a predictive manner).	• Regression model. • Transferability.
			Validity generalization	• Meta-analysis.	• Meta-analysis.
III. Utility of actions	**Implications** • Pass or fail standard.	Consequences		• Pass or fail standard. • Long-term follow-up. • Qualitative studies (satisfaction polls).	• Pass or fail standard. • Long-term follow-up. • Qualitative studies (satisfaction polls).

MCQ: Multiple choice question; OSCE: Objective Structured Clinical Examination; CFA: Confirmatory Factorial Analysis; IRT: Item response theory; EFA: Exploratory Factorial Analysis; DIF: Differential Item Functioning.

Serious intended and unintended consequences have significant effects on individuals. These arise from the decision rules, differential impact on groups, and systemic effects, positive and negative. In this sense, Messick's source of validity evidence based on its consequences relates naturally to Kane's implications' inference.

The use of each framework separately does not explicitly provide this overall picture, which can be helpful to grasp the panorama of the assessment process.

Incorporation of Russell's model

We describe a way to contribute evidence for each inference, incorporating Russell's proposal to segment the validation process into three stages and provide examples of how this process can be implemented for an MCQ test and an OSCE (Table 1).

For the first step, we build upon Kane's argument-based framework, which requires a precise description of the uses and interpretations to be obtained from the test scores (i.e., the interpretation/use argument, IUA). This step should be followed by the development of assumptions for the validity argument and the creation of a testing plan as the scores are produced, reported, and interpreted. Finally, the analysis of the evidence provided leads to a judgment of the degree of validity. These three steps are described as follows:

I. *Use and interpretation*. At the beginning of the instrument development process, the following should be defined in detail: test objective, stakeholders, intended construct and its representation, score interpretations, and uses. Features of the tested population should be described.

II. *Identify and describe claims or assumptions and create a testing plan*. Messick[16] recommends that every validation process should begin by stating the hypotheses to be proven through different sources of evidence. We propose in this second step to outline these hypotheses based on Kane's inferences. In Table 1 we provide examples of these assumptions for each inference and different sources of evidence. We consider that the validity argument can be delineated based on Russell's three stages:

   a. ***Instrument purpose validity.*** Messick's content source should provide evidence through Kane's inferences of Scores, Generalizability, and Extrapolation, as described in Table 1, thus contributing information about the intended domain to be measured, specified in the instrument blueprint.

   b. ***Verification of interpretation and decision.*** Kane's inferences about Scoring, Generalizability, and Extrapolation are useful, and they can be verified with Internal Structure, Response processes, and Relationship with other variables sources (Table 1). In this step, we can explore whether the scores reflect the expected cognitive process, determine the dimensionality, measurement invariance, and reliability of the instrument, and look for correlations within convergent-divergent or test-criterion relationships.

   c. ***Utility of actions.*** Kane's implications inference, through Messick's Consequences source, can be useful in this stage to look for unintended consequences and make sure that positive consequences outweigh the negative.

   As we have hypotheses to be proven and sources of evidence, a plan needs to be developed to gather this information and analyze it. The strategy will depend on the amount of evidence needed for each type of instrument and its stakes: the higher the stakes, the higher the required quality of evidence.

III. *Evaluate the evidence to formulate a judgment.* Finally, the results of the analysis must be studied to produce a judgment about the degree of validity of the test scores' uses and interpretations. The validation process will have to begin again if there is a change in the interpretations or uses.

Importantly, throughout this process, possible threats to validity should be considered and actively avoided. Examples include CU and CIV. The former may be due to bias toward a specific domain area creating a non-representative sample that neglects or underrepresents other relevant areas, assessment of trivial or factual content at the lowest levels of Miller's pyramid, or mismatch of the intended content domain. Some causes of CIV are defective or biased items/stations, inappropriate items' reading level, cheating, indefensible pass score, low reliability, and teaching to the test.[25]

Discussion and conclusions

There is not a global consensus about several aspects of validity, including its definition.[44,46,48] Academic organizations that study educational assessment and evaluation have reached an agreement that encompasses a framework to prove the validity of uses and interpretations of test scores.[11,49] In several academic and testing circles, the lack of consensus causes differences of opinion and confusion in the understanding and application of validity frameworks.[50] Therefore, the process of validation usually becomes a major conceptual and organizational enterprise, which can be challenging for HPEs who do not have an assessment background or an appropriate testing support infrastructure.

We propose a strategy to arrange and visualize assessment data and information in a way that can make sense to HPEs and help them comprehend the results and that is congruent with the need to validate both the score uses and the score meaning separately, following Cizek[51] and Sireci.[46] Matching frameworks with the goal of enriching and potentiating them amplifies the vision and understanding of the interested parties about validity. Furthermore, as validity cannot be oversimplified because of its complexity, the contribution of this paper is not only to attempt to build connections among the most important validity frameworks through the validity evidence sources' elements, but to incorporate a model that considers the different stages of instrument development making it easier to understand where we stand at each step of test development in terms of validity.

There are other efforts in this area, for instance, Kinnear et al.[52] proposed a validity map in which they described their views on how to merge Messick's and Kane's frameworks. We consider that Kane's inferences align with Messick's sources of evidence somewhat different than Kinnear, as we took into account the different elements of each source of evidence and how they match with Kane's inferences and its own evidence sources.

On the other hand, Cook and Hatala[18] suggested a practical approach to validation which requires ten key points to consider. We agree with their recommendations, for instance performing validation on an extant instrument rather than creating a new one from scratch and using a framework. We can situate some of these recommendations in Russell's three stages: in stage one, define the construct and proposed interpretation, and make explicit the intended decision; for the second stage, develop the argument; for every stage, prioritize needed validity evidence. We also are in accordance with Kane regarding the importance of a separate IUA,[53] and consider that the specification of inferences and assumptions should be part of every stage of our proposed step II.

A common question during validation is how much evidence is needed for each instrument and how to structure its argumentation. When the time comes to judge the evidence and determine the degree of validity of scores uses and interpretations, there are no clear or absolute guidelines. From Step III it follows that the degree of validity is as strong as its weakest link, so our suggestion is to recognize the most important sources of evidence for the particular test, to make sure they are in the best possible shape, and to assign gathering of information to specific persons. For example, test designers should provide evidence about the quality of high-stakes assessments because their consequences are of utmost importance for stakeholders.[54–56]

In many cases, the test consequences are considered last in the validation process, some authors may think it is not critical to do a thorough analysis,[57,58] as few reports consider this evidence source.[33] However, Haertel[59] emphasizes that unintended consequences should be studied, illustrating this example with college admissions tests: if their specific use is to lead to better admissions decisions, then validation should include, but not be limited to, a study of predictive validity. Moreover, it is important to acknowledge that laying out the evidence is not the same as argumentation, so we should use an explicit validity argumentation structure and direct it toward a specific audience, as Kinnear et al. suggest.[60]

One limitation of this proposal is that validation preferably should be performed prospectively; in scenarios where instruments have already been applied and their scores interpreted and used, its usefulness may be limited. In these cases, an adaptation of this model could be to organize the information as if it were to be prospective, with the awareness that no changes can be made to the testing process. An advantage of Russell's three stages is that it favors clear identification of gaps and weakest links in evidence through each moment of instrument development. The value of this proposal is to continue the discussion on validity and validation, and contribute to its understanding and application in HPE, as every effort put into validation will contribute to fairer assessments.

Disclosure statement

Ethical review, consent to participate or publish, and data availability are not applicable. The authors do not have financial or competing interests or benefits to declare. No external funding was provided.

Funding

The author(s) reported there is no funding associated with the work featured in this article.

ORCID

Blanca Ariadna Carrillo-Avalos http://orcid.org/0000-0003-4111-4795
Iwin Leenen http://orcid.org/0000-0003-4807-540X
Juan Andrés Trejo-Mejía http://orcid.org/0000-0002-0680-6836
Melchor Sánchez-Mendiola http://orcid.org/0000-0002-9664-3208

References

1. Violato C, Gauer JL, Violato EM, Patel D. A study of the validity of the new MCAT exam. *Acad Med*. 2020;95(3):396–400. doi:10.1097/ACM.0000000000003064.
2. Paton LW, McManus IC, Cheung KYF, Smith DT, Tiffin PA. Can achievement at medical admission tests predict future performance in postgraduate clinical assessments? A UK-based national cohort study. *BMJ Open*. 2022;12(2):e056129. doi:10.1136/bmjopen-2021-056129.
3. Simpson PL, Scicluna HA, Jones PD, et al. Predictive validity of a new integrated selection process for medical school admission. *BMC Med Educ*. 2014;14(1):86. doi:10.1186/1472-6920-14-86.
4. Bala L, Pedder S, Sam AH, Brown C. Assessing the predictive validity of the UCAT – A systematic review and narrative synthesis. *Med Teach*. 2022;44(4):401–409. doi:10.1080/0142159X.2021.1998401.
5. Barajas A, Ramos C, Castillo JD, et al. Flaws in the design of the Examen Nacional para Aspirantes a Residencias Médicas produce inequity. *Salud Publica Mex*. 2019;61(2):125–135. doi:10.21149/9790.
6. Crawford C, Black P, Melby V, Fitzpatrick B. An exploration of the predictive validity of selection criteria on progress outcomes for pre-registration nursing programmes – a systematic review. *J Clin Nurs*. 2021;30(17–18):2489–2513. doi:10.1111/jocn.15730.
7. Cunningham C, Patterson F, Cleland J. A literature review of the predictive validity of European dental school selection methods. *Eur J Dent Educ*. 2019;23(2):73–87. doi:10.1111/eje.12405.
8. Hafferty FW, O'Brien BC, Tilburt JC. Beyond high-stakes testing: learner trust, educational commodification, and the loss of medical school professionalism. *Acad Med*. 2020;95(6):833–837. doi:10.1097/ACM.0000000000003193.
9. Panda N, Bahdila D, Abdullah A, Ghosh AJ, Lee SY, Feldman WB. Association between USMLE step 1 scores and in-training examination performance: a meta-analysis. *Acad Med*. 2021;96(12):1742–1754. doi:10.1097/ACM.0000000000004227.
10. Salehi PP, Azizzadeh B, Lee YH. Pass/Fail Scoring of USMLE Step 1 and the Need for Residency Selection Reform. *Otolaryngol Head Neck Surg*. 2021;164(1):9–10. doi:10.1177/0194599820951166.
11. American Educational Research Association, American Psychological Association, National Council on Measurement in Education. *Standards for Educational and Psychological Testing*. 6th ed. Washington, DC: American Educational Research Association, American Psychological Association & National Council on Measurement in Education; 2014.
12. Ferrara S. Our field needs a framework to guide development of validity research agendas and identification of validity research questions and threats to validity. *Interdisciplinary Research and Perspectives*. 2007;5(2–3):156–164. doi:10.1080/15366360701487500.
13. Kreptul D, Thomas RE. Family medicine resident OSCEs: a systematic review. *Educ Prim Care*. 2016;27(6):471–477. doi:10.1080/14739879.2016.1205835.
14. St-Onge C, Young M, Eva KW, Hodges B. Validity: one word with a plurality of meanings. *Adv Health Sci Educ Theory Pract*. 2017;22(4):853–867. doi:10.1007/s10459-016-9716-3.
15. Cronbach LJ, Meehl PE. Construct validity in psychological tests. *Psychol Bull*. 1955;52(4):281–302. doi:10.1037/h0040957.
16. Messick S. Validity. In: Linn RL, ed. *Educational Measurement*. 3rd ed. New York, NY: Macmillan; 1989. p. 13–104. doi:10.1002/j.2330-8516.1987.tb00244.x.
17. Downing SM. Validity: on the meaningful interpretation of assessment data. *Med Educ*. 2003;37(9):830–837. doi:10.1046/j.1365-2923.2003.01594.x.
18. Cook DA, Hatala R. Validation of educational assessments: a primer for simulation and beyond. *Adv Simul*. 2016;1(1):1–12. doi:10.1186/s41077-016-0033-y.
19. Sireci S, Faulkner-Bond M. Validity evidence based on test content. *Psicothema*. 2014;26(1):100–107. doi:10.7334/psicothema2013.256.
20. Auewarakul C, Downing SM, Jaturatamrong U, Praditsuwan R. Sources of validity evidence for an internal medicine student evaluation system: an evaluative study of assessment methods. *Med Educ*. 2005;39(3):276–283. doi:10.1111/j.1365-2929.2005.02090.x.
21. Varkey P, Natt N, Lesnick T, Downing S, Yudkowsky R. Validity evidence for an OSCE to assess competency in systems-based practice and practice-based learning and improvement: a preliminary investigation. *Acad Med*. 2008;83(8):775–780. doi:10.1097/ACM.0b013e31817ec873.
22. Miller GE. The assessment of clinical skills/competence/performance. *Acad Med*. 1990;65(9 Suppl):S63–S67. doi:10.1097/00001888-199009000-00045.
23. Embretson SE. A cognitive design system approach to generating valid tests: application to abstract reasoning. *Psychol Methods*. 1998;3(3):380–396. doi:10.1037/1082-989X.3.3.380.
24. Padilla JL, Benítez I. Validity evidence based on response processes. *Psicothema*. 2014;26(1):136–144. doi:10.7334/psicothema2013.259.
25. Downing SM, Haladyna TM. Validity threats: overcoming interference with proposed interpretations of assessment data. *Med Educ*. 2004;38(3):327–333. doi:10.1046/j.1365-2923.2004.01777.x.
26. Haladyna TM, Downing SM. Construct-irrelevant variance in high-stakes testing. *Educational Measurement*.

2004;23(1):17–27. doi:10.1111/j.1745-3992.2004.tb00149.x.
27. Leenen I. Virtudes y limitaciones de la teoría de respuesta al ítem para la evaluación educativa en las ciencias médicas. *Investigación en Educación Médica*. 2014;3(9):40–55. doi:10.1016/S2007-5057(14)72724-3.
28. Rios J, Wells C. Evidencia de validez basada en la estructura interna. *Psicothema*. 2014;26(1):108–116. doi:10.7334/psicothema2013.260.
29. Boer D, Hanke K, He J. On detecting systematic measurement error in cross-cultural research: a review and critical reflection on equivalence and invariance tests. *J Cross Cult Psychol*. 2018;49(5):713–734. doi:10.1177/0022022117749042.
30. Berrío ÁI, Gómez J, Arias EM. Developments and trends in research on methods of detecting differential item functioning. *Educ Res Rev*. 2020;31(March):100340. doi:10.1016/j.edurev.2020.100340.
31. Downing SM. Reliability : on the reproducibility of assessment data. *Med Educ*. 2004;38(9):1006–1012. doi:10.1046/j.1365-2929.2004.01932.x.
32. Dunn TJ, Baguley T, Brunsden V. From alpha to omega: a practical solution to the pervasive problem of internal consistency estimation. *Br J Psychol*. 2014;105(3):399–412. doi:10.1111/bjop.12046.
33. Lyons-Thomas J, Liu Y, Zumbo BD. Validity and validation in social, behavioral, and health sciences: a synthesis of syntheses. In: Zumbo BD, Chan EKH, eds. *Validity and Validation in Social, Behavioral, and Health Sciences*, 1st ed. Cham: Springer International Publishing; 2014:313–319.
34. Lane S. Test-based accountability systems: the importance of paying attention to consequences. *ETS Research Report Series*. 2020;2020(1):1–22. doi:10.1002/ets2.12283.
35. Cook DA, Brydges R, Ginsburg S, Hatala R. A contemporary approach to validity arguments: a practical guide to Kane's framework. *Med Educ*. 2015;49(6):560–575. doi:10.1111/medu.12678.
36. Kane M. Validating score interpretations and uses: Messick Lecture, Language Testing Research Colloquium, Cambridge, April 2010. *Language Testing*. 2011;29(1):3–17. doi:10.1177/0265532211417210.
37. Cook DA, Kuper A, Hatala R, Ginsburg S. When assessment data are words: validity evidence for qualitative educational assessments. *Acad Med*. 2016;91(10):1359–1369. doi:10.1097/ACM.0000000000001175.
38. Kane MT. An argument-based approach to validation. *Psychol Bull*. 1992;112(3):527–535. doi:10.1037/0033-2909.112.3.527.
39. Kane MT. Validating the interpretations and uses of test scores. In: Lissitz RW, ed. *Validity: Revisions, New Directions and Applications*. Charlotte, NC: Information Age Publishing, Inc.; 2009:39–64.
40. Chalhoub M. Validity theory: reform policies, accountability testing, and consequences. *Language Testing*. 2016;33(4):453–472. doi:10.1177/0265532215593312.
41. Brennan R. Commentary on "validating the interpretations and uses of test scores. *J Educational Measurement*. 2013;50(1):74–83. doi:10.1111/jedm.12001.
42. Daniels VJ, Pugh D. Twelve tips for developing an OSCE that measures what you want. *Med Teach*. 2018;40(12):1208–1213. TWELVE doi:10.1080/0142159X.2017.1390214.
43. Hess BJ, Kvern B. Using Kane's framework to build a validity argument supporting (or not) virtual OSCEs. *Med Teach*. 2021;43(9):999–1004. doi:10.1080/0142159X.2021.1910641.
44. Borsboom D, Mellenbergh GJ, van Heerden J. The concept of validity. *Psychol Rev*. 2004;111(4):1061–1071. doi:10.1037/0033-295X.111.4.1061.
45. Cizek GJ. Defining and distinguishing validity: interpretations of score meaning and justifications of test use. *Psychol Methods*. 2012;17(1):31–43. doi:10.1037/a0026975.
46. Sireci SG. On the validity of useless tests. *Assess Educ*. 2016;23(2):226–235. doi:10.1080/0969594X.2015.1072084.
47. Russell M. Clarifying the terminology of validity and the investigative stages of validation. *Educational Measurement*. 2022;41(2):25–35. doi:10.1111/emip.12453.
48. Camargo SL, Herrera AN, Traynor A. Looking for a consensus in the discussion about the concept of validity: a Delphi study. *Methodology*. 2018;14(4):146–155. doi:10.1027/1614-2241/a000157.
49. Sireci SG. Comments on valid (and invalid?) Commentaries. *Assess Educ*. 2016;23(2):319–321. doi:10.1080/0969594X.2016.1158694.
50. Young M, St-Onge C, Xiao J, Vachon Lachiver E, Torabi N. Characterizing the literature on validity and assessment in medical education: a bibliometric study. *Perspect Med Educ*. 2018;7(3):182–191. doi:10.1007/s40037-018-0433-x.
51. Cizek GJ. *Validity: An Integrated Approach to Test Score Meaning and Use*. New York, NY: Routledge; 2020.
52. Kinnear B, Kelleher M, May B, et al. Constructing a validity map for a workplace-based assessment system: cross-walking Messick and Kane. *Acad Med*. 2021;96(7S):S64–S69. doi:10.1097/ACM.0000000000004112.
53. Kane MT. Validating the interpretations and uses of test scores. *J Educ Measure*. 2013;50(1):1–73. doi:10.1111/jedm.12000.
54. Kane MT. Validation as a pragmatic, scientific activity. *J Educ Meas*. 2013;50(1):115–122. doi:10.1111/jedm.12007.
55. Crooks TJ, Kane MT, Cohen AS. Threats to the valid use of assessments. *Assess Educ*. 1996;3(3):265–286. doi:10.1080/0969594960030302.
56. Gasmalla HEE, Tahir ME. The validity argument: addressing the misconceptions. *Med Teach*. 2020;43(12):1453–1455. doi:10.1080/0142159x.2020.1856802.
57. Moss PA. The role of consequences in validity theory. *Educ Meas*. 1998;17(2):6–12. doi:10.1111/j.1745-3992.1998.tb00826.x.
58. Nichols PD, Williams N. Consequences of test score use as validity evidence: roles and responsibilities. *Educ Meas*. 2009;28(1):3–9. doi:10.1111/j.1745-3992.2009.01132.x.
59. Haertel E. Getting the help we need. *J Edu Meas*. 2013;50(1):84–90. doi:10.1111/jedm.12002.
60. Kinnear B, Schumacher DJ, Driessen EW, Varpio L. How argumentation theory can inform assessment validity : a critical review. *Med Educ*. 2022;56(11):1064–1075. doi:10.1111/medu.14882.

FACULTAD DE MEDICINA

Revista Investigación en
Educación Médica

Investigación
en Educación Médica

Facultad de Medicina, UNAM
av. Universidad 3000
Ciudad Universitaria,
Cd. Mx. 04510. México

(55) 5622 66 66
ext. 82318

riem@unam.mx,
revistainvestedu@gmail.com
www.riem.facmed.unam.mx

Estimada Blanca Ariadna Carrillo Avalos,

En relación con el manuscrito titulado "Evidencias de validez del proceso de admisión a una escuela de medicina en México", postulado por usted y sus coautores, me complace comunicarles nuestra decisión de aceptarlo para su publicación. Como resultado de las modificaciones realizadas tras el proceso de revisión por pares, su artículo se integrará en el número 50 correspondiente al período abril-junio de 2024.

Agradezco su interés por difundir el producto de su trabajo de investigación en el campo de las ciencias de la salud en las páginas de la revista, y le envío un cordial saludo.

A t e n t a m e n t e
"POR MI RAZA HABLARÁ EL ESPÍRITU"
Ciudad Universitaria, Cd. Mx., a 16 de noviembre de 2023.

Dra. Teresa I. Fortoul van der Goes
Editora Asociada