FACULTAD DE EDUCACIÓN
INSTITUTO DE INVESTIGACIÓN EN EDUCACIÓN
INIE

INFORME FINAL

Estado de la cuestión sobre modelos de evaluación y medición de programas de Educación Superior en el área de la Salud

No. 724-B7-761

Lucrecia Alfaro Rojas
Landy Chavarría Garita
Natalia Salas Segreda
María Paula Villarreal Galera

Mayo 2018

ÍNDICE GENERAL

Índice de tablas

Índice de anexos

I. INFORMACIÓN GENERAL

- Número de la actividad aprobada: No. 724-B7-761.
- Nombre del proyecto: Estado de la cuestión sobre modelos de evaluación y medición de programas de Educación Superior en el área de la Salud.
- Unidad base de las investigadoras: Sistema de Estudios de Posgrado (SEP)
- Unidad de adscripción: Instituto de Investigaciones en Educación (INIE)
- Programa al que pertenece: Programa de investigación, cambio, desarrollo y gestión de la educación superior.
- Personal investigador:

Nombre de las investigadoras	**Carga asignada**	**Programa que asigna la carga**
Lucrecia Alfaro Rojas	½	SEP
Landy Chavarría Garita	½	SEP
Natalia Salas Segreda	¼	SEP
María Paula Villarreal Galera	1/8	Ad honorem

- Vigencia del proyecto: de 21/04/2017 a 28/2/2018

Resumen:

Mediante esta actividad de investigación se pretendió registrar publicaciones recientes relacionadas con modelos de evaluación y medición de programas de formación universitaria en el área de la Salud. Para ello se utilizó la metodología planteada por Gómez-Luna, Fernando-Navas, Aponte-Mayor y Betancourt-Buitrago (2014), en la cual se efectúa una revisión bibliográfica por medio de cuatro fases: la definición del problema, la búsqueda de información en bases de datos científicas, la organización de la información y el análisis de lo recopilado. Los resultados obtenidos de este estudio fueron: 62 publicaciones que comprendieron desde el año 2006 hasta el 2017. Las técnicas de evaluación encontradas que han sido utilizadas con mayor frecuencia son: la evaluación clínica objetiva estructurada (OSCE), la Mini-entrevista múltiple (MMI), el Mini-CEX, los exámenes escritos y orales. Estos referentes teóricos y metodológicos podrán sustentar un posterior análisis de la situación actual del modelo de admisión al Programa de Posgrado en Especialidades Médicas de la Universidad de Costa Rica.

- Descriptores: evaluación, medición y salud.

II. ANTECEDENTES

II.1. Introducción y antecedentes del proceso investigativo

Entre octubre de 2013 y setiembre de 2014, se llevó a cabo la actividad de investigación No. 723-B3-771: Análisis cualitativo y psicométrico del “Examen de Conocimientos Generales en Medicina” empleado en el concurso nacional para ingresar al Programa de Posgrado en Especialidades Médicas (PPEM), con el propósito de responder a una solicitud emanada del Sistema de Estudios de Posgrado (SEP) en cuanto a proveer insumos para el mejoramiento de los indicadores psicométricos de este instrumento (Cubillo, Rojas y Villarreal, 2014). A partir de los análisis cualitativos y cuantitativos realizados al proceso de construcción y aplicación de la prueba, así como a las características de los ítems, se concluyó que el proceso requería de mejoras para asegurar criterios mínimos de confiabilidad y validez de la prueba.

A nivel cualitativo, se evidenció la falta de protocolos para la construcción de los ítems acordes con una prueba estandarizada de altas consecuencias. Asimismo, se generó una tabla de especificaciones para la construcción de ítems y el ensamblaje de versiones futuras de la prueba.

En cuanto al análisis psicométrico, se encontraron varias deficiencias en las pruebas tanto del 2013 como del 2014, debido a que no se contaba con un constructo meta claramente definido y a que no se realizaban pruebas piloto de los ítems de la prueba final. En general, los ítems no se adecuaban al constructo "conocimientos generales en medicina". Ninguna de las dos pruebas tuvo un nivel aceptable de consistencia interna, con un alfa de Cronbach en el 2013 de 0,74 y en el 2014 de 0,66. El porcentaje de ítems con discriminación regular o deficiente fue de 95% en 2013 y de 99% en 2014, incluyendo discriminación negativa en algunos de ellos.

A partir de los resultados obtenidos mediante los análisis cualitativos y cuantitativos, el equipo investigador generó recomendaciones para establecer criterios y protocolos claros que permitieran desarrollar una prueba estandarizada de altas consecuencias en 2015, así como incrementar la calidad y la confidencialidad en futuros

procesos de construcción de ítems. Ciertamente, esta actividad constituyó un primer paso para examinar el modelo de admisión al PPEM, al enfocarse en las características psicométricas del tipo de preguntas que típicamente han formado parte de esta prueba, así como en la revisión de otros procesos relativos al ensamblaje y la aplicación del examen.

Otro antecedente de esta propuesta es la actividad de investigación No.7234-B5-A79 “Construcción y validación de un banco de ítems para la prueba “Examen de Conocimientos Generales en Medicina” (ECGM) empleada para el ingreso al Programa de Posgrado en Especialidades Médicas”, que tuvo como unidad base el Instituto de investigaciones en Educación (INIE) y que se llevó a cabo entre octubre de 2015 y septiembre de 2016. Esta vez, la actividad de investigación apuntó a la necesidad de empezar a construir un banco de ítems para una prueba estandarizada con evidencias de validez, fundamentadas en las exigencias técnicas y científicas que la medición y la evaluación psicoeducativas demandan. Asimismo, se continuó con la redacción del “Protocolo de construcción de ítems de la prueba” y del “Protocolo de aplicación de la prueba”, cuyos borradores iniciales fueron realizados por el equipo 2013-2014, mientras que una primera versión depurada fue entregada por el equipo actual, junto con el informe final, al Sistema de Estudios de Posgrado.

El trabajo realizado durante esta actividad hizo posible la aplicación del ECGM en el año 2016 a 1.805 profesionales provenientes de las diferentes Escuelas de Medicina del país. La calidad y la organización del proceso hicieron posible una exitosa aplicación del examen, así como la recolección de información en condiciones que mejoraron la estandarización del proceso.

Una vez aplicada la prueba, se realizó el análisis de dificultad del instrumento, el cual señaló que menos de la mitad de los ítems podía considerarse fácil, mientras que la representación de ítems con dificultad alta fue adecuada (mayor que en las oportunidades anteriores). La poca representatividad de ítems con dificultad media sigue siendo un reto para los futuros procesos de construcción de ítems y de

ensamblaje de la prueba. En cuanto al alfa de Cronbach, este alcanzó un valor de 0,68, es decir, subió con respecto al año anterior pero no con respecto al trasanterior. Solo mediante una depuración de la base de datos en la que se excluyó los ítems con la discriminación más baja, el Cronbach llegó a un nivel apenas aceptable de 0,75.

Ambas actividades de investigación relativas al ingreso al PPEM han conducido a la implementación de un proceso amplio y ordenado que, con el seguimiento adecuado, pretende llegar a consolidar un banco de ítems para el ECGM, con las respectivas evidencias de validez y confiabilidad. No obstante, aún no se ha realizado un análisis del proceso de admisión más allá de su primera etapa (que consiste, precisamente, en la aplicación del ECGM).

II.2. Planteamiento del problema

El presente estado de la cuestión surge a partir de la necesidad de mejorar el proceso de selección de aspirantes al Programa de Posgrado de Especialidades Médicas (en adelante PPEM) de la Universidad de Costa Rica, al cual pueden acceder estudiantes de universidades tanto públicas como privadas. Este proceso lleva consigo dos etapas, la primera de ellas que contiene una prueba de opción múltiple para todas las personas aspirantes (sin importar la especialidad a la que desee ingresar), y la segunda etapa, en la cual las personas encargadas de cada una de las especialidades aplican criterios e instrumentos definidos de manera independiente para la selección del estudiantado, que previamente aprobó la primera etapa.

Con el afán de mejorar el procedimiento de la primera y segunda etapa, preliminarmente se requiere contar con una base teórico-metodológica sólida que respalde la valoración que se haga, tanto del modelo actual como de nuevas propuestas que apunten al mejoramiento de la calidad técnica de los procedimientos propios del proceso completo de selección. Así la recolección y el análisis de insumos de tipo científico, será un paso clave en cuanto a garantizar procesos de selección justos y

equitativos, que redundarán en la formación de especialistas más aptos(as) y mejor capacitados(as) para su labor.

II.3. Objetivo general y específico, metas e indicadores

Objetivo general: Sistematizar las experiencias registradas en publicaciones recientes relacionados con modelos de evaluación y medición de programas de formación universitaria en el área de la Salud.

Objetivo específico 1: Analizar publicaciones recientes, relacionadas con modelos de evaluación y medición de programas de formación universitaria en el área de la Salud.

- Meta 1: Recopilar publicaciones recientes de carácter científico, relacionadas con modelos de evaluación y medición de programas de formación universitaria en el área de la Salud.

- Indicador 1: Antología compuesta por literatura científica publicada en los últimos diez años, relacionada con modelos de evaluación y medición en programas de formación universitaria en el área de la Salud, con una ficha resumen para cada una de las publicaciones.

Objetivo específico 2: Identificar tendencias en los modelos recientes de evaluación y medición utilizados en el contexto de programas de formación universitaria en el área de la Salud.

- Meta 1: Análisis descriptivo del conjunto de publicaciones recientes de carácter científico relacionadas con modelos de evaluación y medición en programas de formación universitaria en el área de la Salud.

- Indicador: Capítulo del informe de investigación dedicado al análisis descriptivo del conjunto de publicaciones recientes de carácter científico relacionadas con

modelos de evaluación y medición en programas de formación universitaria en el área de la Salud.

Objetivo específico 3: Comparar los resultados obtenidos por las diferentes investigaciones reseñadas en materia de idoneidad de los métodos de evaluación y medición presentados, con miras al contexto de la selección de estudiantes para el PPEM.

- Meta 1: Documento comparativo con base en los resultados obtenidos por las diferentes investigaciones reseñadas en materia de los alcances de los métodos de evaluación y medición utilizados.

- Indicador: Capítulo del informe de investigación dedicado a la comparación de los resultados obtenidos por las diferentes investigaciones reseñadas en materia de los posibles alcances de los métodos de evaluación y medición utilizados.

III. REFERENTE TEÓRICO

Para el presente estudio se parte de que el estudiante de medicina adquiere en su formación diferentes habilidades y valores dirigidos hacia el quehacer de dicha profesión, esto por medio de la práctica, la indagación y estudio de una considerable cantidad de información biomédica y médica. En este sentido, el médico debe ser educado y evaluado como clínico, científico, humanista y docente, aunado a esto debe prevalecer en el médico-estudiante un interés continuo por la superación académica y el aprendizaje independiente, ya que estos aspectos propician la adaptabilidad para el cambio y habilidad para pensar de manera crítica, para educar, y para comunicarse claramente (Reddy y Vijayakumar, citado en Rodríguez, 2008).

A nivel de posgrado, los procesos de enseñanza-aprendizaje se orientan a mejorar la formación profesional integral, con base en las diversas competencias profesionales y específicas (propias de cada disciplina). Así, la evaluación conserva cierta complejidad, por lo tanto, es necesario utilizar instrumentos que estén contextualizados, asociados a una situación profesional y pertinentes para las competencias profesionales que se desean evaluar, con el fin de tener éxito en el proceso de evaluación y formación, por ello elegir el instrumento adecuado a una dimensión o a las diversas dimensiones a evaluar de una competencia profesional, puede determinar el tipo de profesional que se está eligiendo (Arenis y Pinilla, 2016).

Según Cabrales (citado en Arenis y Pinilla, 2016) la evaluación del aprendizaje en posgrado es un espacio que permite el desarrollo de una evaluación formativa (por la retroalimentación que permite el proceso), ya que el estudiante es un profesional que se inscribe por vocación e interés. En esta línea Morales y Trianes, proponen que:

> … la educación en posgrado debe servir para la formación de profesionales como agentes de cambio social, no sólo en lo concerniente a la creación y gestión de conocimiento, sino también en el ejercicio de una ciudadanía que propenda por una mayor cohesión social, un ejercicio ciudadano responsable, un compromiso con la comunidad y un actuar con valores, guiados desde el

> conocimiento científico académico y la práctica profesional desde y en la universidad (Morales y Trianes, citado en Arenis y Pinilla, 2016, p.52).

Según el estudio realizado por Rodríguez (2008), para evaluar el conocimiento en medicina se han desarrollado diversos procedimientos como el examen con reactivos de opción múltiple; el examen con respuesta estructurada por el estudiante; el examen ante pacientes reales, hospitalizados y externos; examen ante pacientes estandarizados; examen clínico objetivo y estructurado; examen oral, estructurado y no-estructurado; manejo del problema principal de un paciente y portafolio. Todos estos procedimientos tienen sus ventajas y desventajas, por lo tanto, la elección de alguno de ellos debe ser por medio de un análisis meticuloso.

Algunos de los criterios a considerar para su selección se derivan de sus propiedades psicométricas, en particular validez concerniente al contenido, a la construcción y a la predicción (la validez hace referencia a la medida en que el examen mide la competencia que se propone evaluar); y la confiabilidad (la medida en que el puntaje del examen es consistente y puede ser generalizado) (Rodríguez, 2008).

En la selección del procedimiento de evaluación también se considera aspectos como la aceptabilidad, la capacidad discriminatoria y los costos. Con respecto a la aceptabilidad, se refiere a la opinión que alumnos y profesores tienen sobre el tipo de examen y a su disposición para aceptarlo y, en el caso de los profesores, para elaborar los reactivos correspondientes; por otro lado con relación a la capacidad discriminatoria, hace referencia a la capacidad del examen para diferenciar un nivel de conocimientos de otro; y finalmente los costos, considera los recursos económicos y tiempo necesarios para la elaboración, aplicación y calificación de exámenes (Schuwirth y Van der Vleuten, citado en Rodríguez, 2008).

IV. PROCEDIMIENTO METODOLÓGICO

La metodología de investigación fue la revisión bibliográfica, la cual se ajustó a la propuesta de Gómez-Luna, Fernando-Navas, Aponte-Mayor y Betancourt-Buitrago (2014), compuesta de cuatro fases.

La primera fase es la definición del problema, en el caso específico de esta actividad de investigación, el problema se refirió a la necesidad de conocer y valorar experiencias recientes relacionadas con modelos de evaluación y medición de programas de formación universitaria en el área de la Salud. De acuerdo con Gómez-Luna, Fernando-Navas, Aponte-Mayor y Betancourt-Buitrago (2014) esta primera etapa debe ser lo suficientemente clara para poder realizar una búsqueda bibliográfica que responda a las necesidades de la investigación, y que además aporte al estado de la técnica, de manera que conduzca a un escenario bastante amplio y permita la retroalimentación de la investigación.

En segunda instancia Gómez-Luna, Fernando-Navas, Aponte-Mayor y Betancourt-Buitrago (2014), proponen la búsqueda de información, en la cual se considera la recopilación de material informativo, que provengan de fuentes científicas certificadas. Por ello para el presente estado de la cuestión las bases de datos utilizadas fueron: US National Library of Medicine National Institutes of Health, AAA Science On-Line, AAA Science Signaling, Current Contents: Clinical Medicine, Scielo, EBSCOhost: Academic Search Complete, DOYMA, Redalyc, LILACS, Portal Regional de la BVS. Información y Conocimiento para la Salud, ScienceDirect, Investigación en Educación Médica.

Además, como paso previo se eligieron descriptores que orientaron la indagación, entre los cuales están: medicina, salud, evaluación, examen de alto impacto, educación, confiabilidad, validez, pruebas de ingreso a posgrado, evaluación en medicina, ingreso a posgrados en especialidades médicas, exámenes de alto impacto para ingreso a posgrados en medicina, validez de pruebas de altas consecuencias y evaluación diagnóstica en especialidades médicas.

A pesar de que en la búsqueda se encontró más de 200 artículos, se consideró elegir las publicaciones que tuvieran más afinidad con la temática a investigar y los que fueron publicados después del año 2006. Aunado a estos criterios, se optó por artículos tanto de carácter investigativo como teóricos, con diferentes estrategias metodológicas y de diversos países.

Por otro lado, la tercera fase a la cual hace alusión Gómez-Luna, Fernando-Navas, Aponte-Mayor y Betancourt-Buitrago (2014), es la organización de la información, la cual consiste en organizar de manera sistemática la documentación encontrada. Para cumplir con este criterio (de manera básica en un primer nivel), se valoró la pertinencia del documento de acuerdo con el objetivo general de la actividad de investigación y, más adelante, se hizo de forma más detallada con las fuentes seleccionadas, mediante la elaboración de una ficha que consistió, en una síntesis del artículo que contiene la referencia bibliográfica, los objetivos de la investigación, la metodología aplicada, las ideas centrales del texto, y las principales conclusiones.

La organización de la información también se hizo mediante la elaboración de una tabla comparativa, en la que se incluyó: nombre del artículo, objetivo general, año de publicación, país y universidad en la cual se desarrolló la propuesta, metodología, resultados y conclusiones; lo que permitió el manejo de la información de una manera más global y sistematizada.

Finalmente, la cuarta etapa fue el análisis de la información, la cual se llevó a cabo en una integración de la información de cada una de las fuentes, seguida de una argumentación crítica acerca de sus aportes al estudio del tema, así como de sus posibles sesgos y limitaciones (Gómez-Luna, Fernando-Navas, Aponte-Mayor y Betancourt-Buitrago, 2014). Este análisis fue de carácter descriptivo y se hizo a la luz de las necesidades específicas del PPEM en materia de su proceso de admisión, tomando en cuenta sus particularidades como programa único en su tipo en Costa Rica.

V. ANÁLISIS Y DISCUSIÓN DE LOS RESULTADOS

El total de publicaciones analizadas fueron 62 artículos, que abarcaron desde el año 2006 hasta el 2017. De estos, 53 fueron de carácter investigativo y 9 referentes teóricos. Los países representados en la literatura valorada son: Alemania, Arabia Saudita, Argentina, Australia, Bélgica, Canadá, Chile, Colombia, Costa Rica, Dinamarca, España, Estados Unidos, India, Indonesia, Inglaterra, Irán, Israel, Malasia, México, Países Bajos, Pakistán, Perú, Reino Unido, República Checa y Taiwán. A continuación, se presenta un resumen de las instituciones en las cuales se realizaron los artículos estudiados:

Tabla 1. Instituciones en las cuales se realizaron los artículos analizados del 2006 al 2017

País	Institución
Alemania	· Universidad de Luebeck
	· Universidad de Witten/Herdecke
	· Centro Médico Universitario Hamburg-Eppendorf, Hamburgo
Arabia Saudita	· Universidad de King Faisal, Hofuf
Argentina	· Ministerio de Salud de la Nación
	· Universidad Nacional de Comahue
Australia	· Universidad de Sydney
	· Universidades Bond Unive0rsity
	· University of Queensland
	· Institute of Health and Biomedical Innovation & Griffith University
Bélgica	· Universidad de Ghent
Canadá	· Universidad de Ottawa, Ontario
	· Universidad de British Columbia, Vancouver
	· Universidad de Alberta
	· Universidad de McMaster
Chile	· Universidad de La Frontera
	· Universidad San Sebastián

Colombia	· Universidad de Antioquia, Medellín
	· Universidad Nacional de Colombia
Costa Rica	· Universidad de Costa Rica
Dinamarca	· Universidad del Sur de Dinamarca
España	· Universidad de Barcelona
Estados Unidos	· Cleveland Clinic Lerner College of Medicine, Ohio
	· Universidad de California, San Francisco
	· Universidad de Miami
	· Penn State Hershey Medical Center, Penn State College of Medicine & Department of Behavioral Sciences and Education
	· Universidad de Carolina del Norte
	· Universidad de Cincinnati
	· Universidad de Minnesota
	· Universidad de Nueva York
	· Universidad de Chicago
	· University HealthSystem Consortium
	· Universidad de Emory, Atlanta
	· Universidad de Michigan
	· Universidad de Lowa
India	· Dr. Punjabrao Deshmukh Memorial Medical College, Amravati, India.
Indonesia	· Universidad de Indonesia, Universidad de Andalas
	· Universidad de Sebelas Maret
Inglaterra	· Universidad de Kingston
Irán	· Universidad de Ciencias Médicas, Tabriz
Israel	· Instituto Nacional para Pruebas y Evaluaciones, Jerusalen
Malasia	· Universidad de las Ciencias
México	· Universidad Autónoma de México
Países Bajos	· Universidad de Groningen
	· HAN University of Applied Sciences
	· Maastricht University Medical Center
Pakistán	· Universidad Internacional de Riphah, Islamabad
	· Universidad Aga Khan, Karachi
	· Khyber Medical College, Ayoub Medical College, Saidu Medical College & Gomal Medical Colleges

Perú	· Universidad Nacional de Trujillo · Universidad Privada Antenor Orrego · Instituto Regional de Enfermedades Neoplásticas
Reino Unido	· Univerirsity of London · Universidad de Aberdeen · Escuela de Medicina de Hull York · Hospital de Broomfield · Hospital de Mid Essex Trust · Universidad de Cambridge · Universidad de Dundee
Taiwán	· Universidad Nacional Cheng Kung

En la multiplicidad de estudios analizados las variables como la cantidad de participantes, las edades, las especialidades, los años de estudios y el tipo de evaluación varía en función de la estrategia metodológica empleada, cabe destacar que cada artículo fue revisado en forma individual destacando sus particularidades con respecto al resto de información recolectada, lo cual puede ser mostrado por medio de las fichas descriptivas de cada artículo, lo cual se puede consultar en el **anexo 1** del presente documento, producto de la actividad de investigación.

Las evaluaciones más frecuentes empleadas a estudiantes de medicina son: **La Evaluación Clínica Objetiva Estructurada (OSCE)**: Sultana, Akbar khan, Sharif, Mah Khan y Sadia, 2015; Iblher, Zupanic, Karsetn y Brauer, 2015; Lievens, 2013; Pugh et al., 2016; Martinez et al., 2017; Kaliyadan, Sattar, Kuruvilla y Feroze, 2014; Chávez y Barrantes, 2014; Trejo, Martínez, Méndez, Morales, Ruiz y Sánchez, 2014; Gamboa, et al., 2011; López, 2017; Illesca, Cabezas, Romero y Diaz, 2012; Alarcón, 2013; y Adam et al. 2015. **La Mini-entrevista múltiple (MMI)**: Axelson y Kreiter, 2009; Dowell, et al., 2012; Eva y Macala, 2014; Eva, et al., 2009; Knorr y Hissbach, 2014; Mini-interviews help to recruit students for their values, 2013; Rahim y Yusoff, 2016; Rosenfeld, et al., 2008; Shih-Chieh Liao, et al., 2014; Zaidi, et al., 2014; y Patterson, et al., 2016; Gafni,

et al., 2012. Y **el Mini-CEX** Baños, et al., 2015; Brazil, et al., 2012; Fernández, 2011; Fornells-Vallés, 2009; Malhotra, Hatala y Courneya 2008; y Donato, et al., 2008.

Sumado a lo anterior se destacan los diferentes métodos de selección de estudiantes en el área de medicina, los cuales son:

- Pruebas escritas (exámenes estandarizados) (León, Ortiz, Bonilla y Berlanga, 2006; Duré, Dursi, Raffoul y Caffarena, 2014; Findyartin et al., 2015; Westerkamp, Penninga, Kuks y Shotanus, 2013; Delgado y Sánchez, 2012; Salazar, Veléz y Tobón).

- Pruebas de ubicación en Ciencias de la Salud, que incluyen pruebas del lenguaje aplicado a un contexto académico, logros matemáticos, comprensión matemática, y razonamiento científico (Wadee y Cliff, 2016).

- Evaluaciones de conocimiento en ciencias y variables conductuales de interés (Schripsema, van Trigt, Borleffs y Cohen, 2014).

- Examen de ingreso biomédico (Emery, Bell y Vidal, 2011).

- Exámenes prácticos (León, Ortiz, Bonilla y Berlanga, 2006; Duré, Dursi, Raffoul y Caffarena, 2014; Kenny, McInnes y Singh).

- Examen de aptitud (Patterson, et al., 2016).

- Evaluación por medio de centros de selección (Patterson, Zibarras, Kerrin, Lopes y Price, 2014; y Patterson, et al., 2016).

- Pruebas de juicio situacional (Patterson, et al., 2016; Gafni, et al., 2012).

- Ensayo escrito (O'Neill, Vonsild, Wallsted y Dornam, 2013).

- Cuestionario biográfico (Gafni, et al., 2012).

- Entrevistas (Patterson, et al., 2016; Emery, Bell y Vidal, 2011; Duré, Dursi, Raffoul y Caffarena, 2014; Kenny, McInnes, y Singh, 2013).
- Declaraciones personales, evaluación de personalidad e inteligencia emocional (Patterson, et al., 2016).
- Cartas de recomendación y referencias (Patterson, et al., 2016; Hall, Connell y Cook, 2017; Kenny, McInnes, y Singh, 2013).
- Record académico (Patterson, et al., 2016; Parate, Pande y Lokari, 2016; O'Neill, Vonsild, Wallsted y Dornam, 2013; Duré, Dursi, Raffoul y Caffarena, 2014; Adam et al., 2015; Schripsema, Van Trigt, Borleffs y Cohen, 2014).
- Experiencia laboral (Adam et al., 2015; O'Neill, Vonsild, Wallsted y Dornam, 2013).
- Experiencia de intercambio y trabajo voluntario (O'Neill, Vonsild, Wallsted y Dornam, 2013).
- Publicación de artículos (Hall, Connell y Cook, 2017).
- Manejos de idiomas (Duré, Dursi, Raffoul y Caffarena, 2014).

A continuación, se presenta de forma detallada cada uno de los métodos de evaluación en medicina, además de las técnicas utilizadas y analizadas para la selección de estudiantes en las ciencias médicas.

V.1. Evaluación Clínica Objetiva Estructurada

Ante las deficiencias que presentan las evaluaciones tradicionales (como no demostrar si los estudiantes han adquirido competencias prácticas) se propone la OSCE (Objective Structured Clinical Examination, por sus siglas en inglés), la cual consiste en estaciones de simulación donde los aspirantes deben demostrar sus conocimientos y habilidades atendiendo diferentes situaciones clínicas. En este tipo de evaluación según Iblher, Zupanic, Karsetn y Brauer (2015), se debe contar con un médico examinador para cada estación, los cuales deben estar capacitados sobre el proceso de evaluación y adicionalmente deben contar con las instrucciones de los procedimientos en cada lugar de aplicación.

La OSCE es una prueba de evaluación utilizada ampliamente en las ciencias de la salud para determinar las habilidades de comunicación y competencias en procedimientos clínicos en practicantes. Entre los estudios encontrados está el de Sultana, et al. (2015), en el cual se demuestra la influencia de pacientes estandarizados y no estandarizados en la OSCE, ya que los estudiantes que trabajaron con pacientes estandarizados obtuvieron un mejor rendimiento en habilidades de comunicación y técnicas, en comparación con estudiantes que practicaron con pacientes reales, después de ochos semanas de haber iniciado la práctica.

La evaluación mediante el procedimiento OSCE en Sultana et al. (2015) es descrita como un método donde un supervisor evalúa las habilidades deseadas de los practicantes mientras realizan una examinación, en este estudio específico la examinación pélvica en una única estación. Contrario a la investigación realizada por Iblher, Zupanic, Karsetn y Brauer (2015) donde se incorporaron siete estaciones (Soporte cardíaco avanzado, soporte de vida básico, atención de trauma, emergencias en pediatría, síndrome coronario agudo, atención de vías respiratorias y emergencias en obstetricia), con una duración de cinco minutos cada una, además de un minuto para retroalimentación y para desplazarse a la siguiente estación. A lo largo de esta prueba, los participantes tuvieron que resolver diferentes problemas médicos, tanto teóricos

como prácticos, mientras interactuaban con pacientes estandarizados o maniquíes. Todas las tareas fueron leídas paso a paso por los examinadores y fueron evaluadas mediante listas de chequeo, donde se calificaba con máximo de 25 puntos por estación. Es importante indicar que los participantes no fueron informados de su puntuación en curso para no distraerlos de sus tareas.

Iblher, Zupanic, Karsetn y Brauer (2015) demuestran que los practicantes de medicina lograron puntajes significativamente más altos para las estaciones de atención de trauma, atención de vías respiratorias, emergencias de obstetricia cuando el evaluador era un estudiante de último año, mientras que la puntuación de la estación para emergencias de pediatría fue significativamente mayor para los evaluados por médicos profesionales. Estos hallazgos encontrados en el estudio de Iblher, Zupanic, Karsetn y Brauer (2015) pueden alentar una introducción de estudiantes avanzados como evaluadores, sin embargo, esto puede no ser aplicable en exámenes de alto riesgo y sumativos; en parte, debido a posibles impedimentos legales.

Aunque los objetivos de los currículos de formación en medicina reconocen la importancia de las habilidades interpersonales y las características personales, o las denominadas habilidades blandas, la mayoría los sistemas formales de admisión a las escuelas de medicina tienden a evaluar principalmente los logros académicos en los dominios de la ciencia y las habilidades cognitivas, como el razonamiento verbal o el matemático. Lievens (2013) menciona que la OSCE es un claro indicador que justifica desarrollar y aplicar un enfoque más holístico y sofisticado para la selección, basado en predictores de la atención que sean válidos y relevantes para los pacientes. De esta manera este estudio comparte explícitamente con Iblher, Zupanic, Karsetn y Brauer (2015) la búsqueda por dinamizar la evaluación y formación de profesionales en medicina.

En Bélgica, Lievens (2013) realiza una investigación longitudinal y de cohortes sobre la validez predictiva de pruebas cognitivas y una prueba de habilidades interpersonales basada en video, ambas pruebas aplicadas en el proceso de admisión a las

escuelas de medicina. En este caso se encontró que las pruebas cognitivas son un buen predictor para los cursos de primer año (0.34) con un valor de p<.001, los cursos médicos (0.32) con un valor de p<.001, el promedio de las notas de todo el bachiller universitario (0.37) con un valor de p<.001, el promedio de notas de los cursos de maestría (0.25) con un valor de p<.001, promedio total de todos los cursos llevados en bachillerato y maestría (0.33) con un valor de p<.001 y en una prueba de conocimiento de medicina general aplicada al terminar el internado en esta área (0.15) con un valor de p<.05, esta última realizada 9 años después de haberse aplicado las pruebas cognitivas.

Sobre la prueba de habilidades interpersonales basada en video, esta consistió en observar fragmentos de video y contestar sobre la forma más correcta para abordar la situación, se encontró que predice el rendimiento en los cursos de habilidades interpersonales (0.21) con un valor de p<.001, una entrevista sobre abordaje de un caso (0.19) con un valor de p<.01, evaluación de rendimiento médico (0.15) con un valor de p<.05 y un OSCE sobre habilidades interpersonales (0.12) con un valor de p<.05. Las últimas tres fueron realizadas 9 años después de la aplicación de la prueba de habilidades interpersonales basada en video (Lievens, 2013).

Lamentablemente Lievens (2013) no proporciona mayor detalle de la evaluación OSCE que se realizó, solo indica que fue sobre habilidades interpersonales médicas. Se desconoce sobre la cantidad de estaciones utilizadas y la descripción de estas, sin embargo, el nivel de predicción de las pruebas de habilidades interpersonales basada en video y la OSCE deja entrever que el entrenamiento en habilidades interpersonales durante la educación médica no descalifica el valor de seleccionar estudiantes sobre la base de sus habilidades interpersonales de manera inicial.

Pugh et al. (2016) a diferencia de Lievens (2013) utilizan las evaluaciones OSCE como prueba formativa durante el proceso educativo en Medicina, de esta forma se acerca más a un proceso formativo en competencias médicas. Pugh et al. (2016) explican que con la incorporación de las pruebas se busca promover un aprendizaje

más profundo al cambiar el enfoque de la evaluación del aprendizaje (en el que el objetivo principal es pasar una prueba) a la evaluación para el aprendizaje (en el que el objetivo es la mejora continua, independientemente del resultado en la prueba).

El contenido de la evaluación OSCE de acuerdo con el estudio de Pugh et al. (2016) varió de un año a otro, pero cada evaluación se basó en el plan de estudios completo de residencia de medicina interna. Los casos fueron seleccionados y escritos por internistas con experiencia en facultades. Los mismos residentes desde primero a cuarto año participaron en la evaluación. Los residentes fueron evaluados por médicos examinadores y recibieron retroalimentación verbal (inmediatamente después de cada caso) y retroalimentación numérica proporcionados unas semanas después del examen). Las correlaciones entre los puntajes OSCE (según el año que cursa) y los puntajes de los componentes escritos y de ejecución en el examen objetivo integral de medicina interna del colegio de médicos y cirujanos de Canadá fueron positivos y moderadamente altos, desde 0.3 a 0.54 con un valor de $p<.05$.

Pugh et al. (2016) también indica que a medida que los puntajes en la OSCE aumentaron, la probabilidad de reprobar en cualquiera de los componentes del examen disminuyó. Por ejemplo, por cada aumento de 10 puntos en la OSCE, las probabilidades de que un residente de cuarto año corra el riesgo de fallar el componente escrito de la prueba disminuyeron un 10.5% y las probabilidades de reprobar el componente basado en la ejecución disminuyeron un 14.0%. Estos resultados sorprendieron a las personas investigadoras, ya que no se esperaba tanto impacto de la OSCE en la parte escrita de la prueba.

El estudio de Pugh et al. (2016) brinda la novedad de demostrar que los puntajes de una evaluación OSCE podrían usarse para identificar a los residentes en alto riesgo de reprobar un futuro examen nacional de medicina interna, sin embargo, se debe indicar que esta conclusión solo fue estadísticamente significativa para la cohorte de segundo y cuarto año de residencia.

Martinez et al. (2017) mencionan el uso de la OSCE como herramienta para evaluar el nivel de competencia clínica de los estudiantes al inicio del internado médico de pregrado (10° semestre), aunado a una evaluación teórica computarizada de 232 reactivos con el formato escrito de preguntas de opción múltiple de una sola respuesta, siguiendo las recomendaciones internacionales para un instrumento de esta naturaleza.

Estos autores mencionan que el proceso de elaboración de las estaciones de la OSCE estuvo a cargo de profesores con experiencia en este tipo de pruebas, considerando evaluaciones formativas y sumativas de la Facultad de Medicina. Posteriormente, realizaron la selección y la adecuación de estaciones por expertos en evaluación y elaboraron el material de apoyo para cada estación, como son los resúmenes de escenarios clínicos, las rúbricas con escalas globales y los libretos para los pacientes estandarizados. Es importante mencionar que se realizó una prueba piloto para observar deficiencias y realizar los ajustes pertinentes (Martínez et al., 2017).

Para las estaciones se contó con la participación de personas no enfermas (médicos pasantes en Servicio Social) que se capacitaron para representar un problema de salud en forma coherente y homogénea como pacientes estandarizados. La OSCE quedó conformado por cinco estaciones estandarizadas de doce minutos cada una en dos circuitos simultáneos. Con los estudiantes se realizaron reuniones informativas voluntarias (Martínez et al., 2017).

En relación con los resultados de los atributos de la competencia clínica, Martínez et al. (2017) reportan que los más altos fueron obtenidos en los atributos de habilidades de comunicación y de interrogatorio, en contraste con el atributo con menor valoración, que fue la interpretación de exámenes de laboratorio y gabinete. Entre las estaciones de la OSCE, la de embarazo no deseado tuvo la mayor puntuación (67.5); en contraste, con la menor puntuación (54.8) de la estación de fiebre-cefalea. La puntuación en el alfa de Cronbach (0.59) de la OSCE se encuentra dentro del rango de lo reportado en exámenes de menos de diez estaciones (0.56-0.74). Entre los atributos

de la competencia clínica se documenta que los estudios diagnósticos y de laboratorio, la terapéutica y la exploración física son los que obtuvieron valores menores.

Martinez et al. (2017) afirman que este estudio muestra claramente que se requiere realizar intervenciones educativas en algunas áreas de los programas académicos para hacer los cambios pertinentes, tanto en la enseñanza como en los contenidos; y recalcan la importancia y el significado de las evaluaciones formativas, cuyos resultados pueden realimentar la enseñanza de la medicina y fortalecer el propósito primario de dichos exámenes. Siguiendo con la idea de dinamizar y modificar los procesos educativos en medicina que comparte con Iblher, Zupanic, Karsetn y Brauer (2015) y Kaliyadan, Sattar, Kuruvilla y Feroze (2014).

Kaliyadan, Sattar, Kuruvilla y Feroze (2014) utilizaron la OSCE asistida por computadora, para evaluar a ciento veintinueve estudiantes de medicina en el quinto año de su curso de pregrado. A cada estudiante se le mostraron dieciséis imágenes de casos clínicos comunes y cada imagen fue seguida por cuatro preguntas. Las preguntas más frecuentes se referían a la descripción de las lesiones cutáneas, el diagnóstico y el diagnóstico diferencial, la investigación y el tratamiento. Cinco minutos fueron asignados a cada diapositiva. Los puntajes en la OSCE se correlacionaron estadísticamente con los puntajes obtenidos en la presentación de un caso clínico y el promedio del curso, incluido un examen escrito. Los puntajes de los estudiantes en la OSCE asistido por computadora mostraron una fuerte correlación positiva con los puntajes en la presentación clínica (coeficiente de Pearson 0.923, valor P <0.000) y una buena correlación con el promedio del curso (coeficiente Pearson 0.728, valor P <0.000), lo que indica que este es un método confiable para la evaluación dermatológica.

Estos autores concluyen que uno de los factores más importantes que aumenta la confiabilidad sería aumentar el número de diapositivas, con el fin de garantizar una muestra más amplia que cubra todos los temas "imprescindibles" en la dermatología de pregrado y que además sería esencial para garantizar que la evaluación tenga buenas propiedades psicométricas (Kaliyadan, Sattar, Kuruvilla y Feroze, 2014).

Por otro lado, para el estudio realizado por Chávez y Barrantes (2014), se generaron listas de cotejos, que demuestran la relación de características, aspectos y cualidades de los practicantes, por ello las listas se utilizan para determinar y registrar el grado de competencias alcanzado por los estudiantes durante las estaciones de la OSCE, en este sentido se establece la presencia o ausencia de una conducta durante la evaluación, por medio de un profesor de las especialidades de Cirugía General, Neurocirugía, Oftalmología, Otorrinolaringología y Urología (áreas que evalúa la OSCE en dicha investigación).

Para dicho estudio se contó con diez estaciones, Chávez y Barrantes (2014) cinco fueron activas y contaban con la presencia de un profesor y un paciente simulador (este paciente fue entrenado previamente por el docente de acuerdo con el campo que requería ser evaluado). Y cinco estaciones fueron pasivas y contaron con la presencia de un profesor quien había dispuesto una imagen radiográfica o de tomografía axial computarizada o contaba con un maniquí o una maqueta. Sobre el proceso como tal, los autores indican que los estudiantes ingresaron a los ambientes de examen en número de diez, uno por cada estación de acuerdo con su número de orden. Antes de ingresar a cada cubículo, leyeron las indicaciones del instructivo en el que se expone brevemente el caso o el procedimiento que debe realizar el estudiante para posteriormente ingresar al ambiente en el que se encuentra un profesor además de la imagen a observar.

Chávez y Barrantes (2014) además afirman que en las estaciones activas los practicantes realizaban la anamnesis o el examen físico de acuerdo con la patología planteada, posteriormente finalizaban la rotación respondiendo preguntas escritas en algunas estaciones. El tiempo de rotación por cada estación fue de cinco minutos incluyendo el tiempo para leer el instructivo.

Sobre la evaluación, específicamente, Chávez y Barrantes (2014) mencionan que los profesores solamente observaron el comportamiento y rendimiento de los alumnos, y registraban en su lista de cotejos, de manera dicotómica: si se cumplían o no los

distintos pasos considerados en las respectivas estaciones. Se hace mención de algunos ejemplos, como lo son: si el practicante al ingresar saludaba o no al paciente simulador presentes en la estación, si el alumno se colocaba o no los guantes para realizar un procedimiento y si seguía un orden o era desorganizado. En cuanto a los resultados se reporta que las notas obtenidas de la calificación de cada una de las listas de cotejos de las diez estaciones de la OSCE, varía desde 8,1 hasta 18,4 puntos, siendo el promedio final de 16.2 puntos (en escala de 20).

En el estudio Chávez y Barrantes (2014) indican que se apreció confiabilidad en niveles de 0.33 a 0.90 en seis estaciones y cuatro no tuvieron confiabilidad; las estaciones no confiables, fueron una de cirugía general la que se refiere a anamnesis del dolor abdominal; dos de otorrinolaringología, la prueba de Rinne y la prueba de senos para nasales y una de urología, la referida a la prueba de cólico renal, además la validez varió desde 0.65 hasta 1.0. Estos datos permiten inferir que en la realidad actual la enseñanza y aprendizaje del médico privilegia el conocimiento y es muy poca o casi nula la formación en aspectos educativos y humanos, por lo tanto, recomiendan más experiencia y capacitación de los profesores en procesos de facilitación del aprendizaje.

Trejo, Martínez, Méndez, Morales, Ruiz y Sánchez (2014) concuerdan con Lievens (2013), Sultana et al (2015), Iblher, Zupanic, Karsetn y Brauer (2015), y Martínez et al (2017) sobre la necesidad de formar médicos que sean capaces de proporcionar una atención integral y al mismo tiempo mantener el aspecto humanístico de la medicina. Además, estos autores plantean la OSCE puede evaluar componentes de la competencia clínica como el interrogatorio, la exploración física, la interpretación de los estudios de laboratorio y gabinete, el diagnóstico y plan de manejo y la comunicación en la relación médico-paciente. La OSCE es una prueba flexible en la cual se pueden usar una variedad de métodos para obtener una información amplia que permita evaluar las habilidades clínicas.

Trejo, Martínez, Méndez, Morales, Ruiz y Sánchez (2014) realizan un ejemplo sobre la importancia de explorar varios componentes de la competencia clínica para

mejorar la confiabilidad de la evaluación, “es similar a la formulación de preguntas de opción múltiple para evaluar el conocimiento de los estudiantes, al usarse en combinación con los formatos escritos, puede incrementar la confiabilidad, si se incluye una cantidad de estaciones suficiente para lograr una amplia muestra de situaciones clínicas” (p. 10). Es de suma importancia que los criterios de evaluación sean determinados por las competencias que deben alcanzar los estudiantes de acuerdo con los objetivos del plan de estudios y los programas académicos correspondientes. En este estudio se realizó un pretest antes de realizar el internado en medicina y postest de igual manera.

Se presentaron 315 internos al pretest y 10 meses después se aplicó el postest. La evaluación consistió en 16 estaciones de seis minutos cada una y dos estaciones de descanso, las cuales según Trejo, Martínez, Méndez, Morales, Ruiz y Sánchez (2014) fueron conformadas y validadas por un comité de expertos profesores de la Facultad de Medicina, con al menos 10 años de experiencia clínica, pertenecientes a una especialidad, que han sido acreditados en un curso-taller sobre el OSCE y participado en al menos tres exámenes con esta metodología.

Trejo, Martínez, Méndez, Morales, Ruiz y Sánchez (2014) mencionan que los casos contenían una presentación con el escenario clínico y un formato de instrucciones para que el estudiante se centrase en el problema y siguiera las indicaciones. En las estaciones con pacientes, el examinador, además de calificar en la lista de cotejo, emitía una calificación global sobre las habilidades de comunicación interpersonal, mediante una escala global de 1 a 9, en la que 1 a 3 = insatisfactorio, 4 a 6 = satisfactorio y 7 a 9 = superior. En cuanto a los resultados, los autores mencionan que el alfa de Cronbach, en el pretest fue de 0.62 y en el postest, de 0.64. En la aplicación del examen participaron 108 profesores evaluadores. La media global del OSCE pretest al principio del internado fue de 55.6 ± 6.6 y la media de la medición postest al final del internado, de 63.2 ± 5.7, con una diferencia absoluta del 7.6% ($p < 0.001$). Por lo anterior se puede observar que el área de medicina familiar tuvo la mayor diferencia de medias pre y postest, que fue de 12.1, seguida por el área de cirugía. En contraste, el área de menor diferencia fue la de medicina interna, aunque todas las diferencias fueron significativas.

Gamboa, et al. (2011) reportan el uso de una evaluación OSCE con la finalidad de contar con un instrumento de evaluación de competencias en pediatría que pudiera emplearse con fines de evaluación formativa y acumulativa y, al mismo tiempo, brindar a los profesores el entrenamiento en la elaboración de estaciones para evaluar competencias. Las etapas que la OSCE consideró fueron: primero la planeación en donde se encargaron del diseño de instrumentos, validez de contenido y elaboración del material de apoyo; segundo la organización el día previo al examen donde se dio la preparación del material de apoyo y últimos detalles; tercero el día del examen; y cuarto después del examen la reflexión y evaluación.

La duración total del examen fue de dos horas con veinte minutos, que incluía un minuto para el cambio de estación, para reordenar los materiales y permitir el desplazamiento de los alumnos. No se utilizaron estaciones intermedias de descanso; sin embargo, algunas de las estaciones requirieron menos de cinco minutos para su resolución. Cabe resaltar que, en esta especialidad a diferencia de otras áreas de la Medicina, en Pediatría es difícil contar con pacientes estandarizados, por lo tanto, para el estudio piloto reportado solo en una estación se contó con un niño sano de 9 años; agregan que en otra estación un médico actuó de enfermera y en una más una residente hizo el papel de mamá de un niño enfermo; estos participantes fueron entrenados para aportar los mismos datos a todos los examinados (Gamboa, et al. 2011).

En cuanto a los resultados, estos autores reportan que las estaciones aprobadas de manera global fueron doce y las no aprobadas ocho. El promedio global de todas las estaciones fue de 6.53 (DE 0.62) (en escala de 10). La estación que alcanzó el promedio más alto fue la de Cardiología con una puntuación de 8.90 (DE 1.6) y la que obtuvo el promedio más bajo fue la de Gastroenterología con una puntuación de 3.04 (DE 0.98). En la discusión de esta publicación se hace énfasis a que la OSCE permite evaluar múltiples habilidades clínicas fundamentales de los programas de posgrado que no pueden ser valoradas por métodos tradicionales que coincide con anteriores planteamientos de otros autores analizados en este apartado (Gamboa, et al., 2011).

En Costa Rica se conoce de un antecedente directo del uso de este tipo de evaluación, López (2017), indica que a pesar de que este tipo de evaluación ha sido muy poco utilizada en el estudiantado de obstetricia, se considera una solución efectiva para evaluar habilidades que no son medibles con exámenes tradicionales. Ella reporta el uso de este tipo de evaluación en estudiantes que cursaban el curso de Enfermería Ginecobstetricia y Perinatal I, de la Maestría en Enfermería Ginecológica, Obstétrica y Perinatal. La autora reporta que esta evaluación se llevó cabo en el Centro de Simulación en Salud de la Escuela de Enfermería. En la evaluación se contó con la participación de 24 estudiantes matriculados durante el 2016. Durante la intervención se contó con la participación de dos profesoras, como pacientes estandarizadas, acreditadas en simulación clínica. La modalidad de la OSCE reportada en López (2017) muestra una característica única en comparación con los demás antecedentes revisados en este aparatado, en este caso se reporta el uso de la grabación en cada estación con la finalidad de que los y las estudiantes revisaran su accionar.

Se sistematizó la experiencia de las tres estaciones utilizadas: Consulta de control prenatal; Consulta de planificación familiar; Toma de citología vaginal. De igual manera que en Chávez y Barrantes (2014) y Trejo, Martínez, Méndez, Morales, Ruiz y Sánchez (2014) se diseñó una lista de cotejo dicotómica para establecer la aparición o no de la conducta deseada y agrega la variante de que en caso de que no, se encontraba un espacio para anotaciones o comentarios que justifican el marcado de la casilla. La autora indica que es importante aclarar que las listas de cotejo se construyeron con base en el manual de procedimientos de la Caja Costarricense del Seguro Social (López, 2017).

López (2017) señala como parte de los resultados que al observar los videos, se evaluó el lenguaje verbal y no verbal utilizado por el estudiante, sus habilidades y estrategias de comunicación. La lista de cotejo permitió dar una nota sanativa no solo a la aplicación del conocimiento, sino también en comunicación, al igual que en Sultana, et al. Sadia (2015). En concordancia con los anteriores autores, encuentra que en este método es pertinente para evaluar los conocimientos en acción.

Illesca, Cabezas, Romero y Diaz (2012) comentan sobre el uso de la OSCE en pregrado y postgrado, en ese marco de amplia aceptación de esta metodología de evaluación se propone su uso en la formación de profesionales en enfermería. Por lo tanto, se plantea indagar en profundidad el significado que tuvo para los estudiantes ser evaluados a través de la OSCE al finalizar la práctica clínica. Para esto se estructuraron trece estaciones: ocho para demostración de habilidades clínicas y actitudinales (vía venosa con tapón; curación simple; preparación solución parenteral; administración subcutánea de fármacos; confección de cama en dos tiempos; control de temperatura, control de respiración y pulso; postura de guantes) y cinco orientadas al dominio cognoscitivo (preparación preoperatorio en general; división topográfica del abdomen; referencias anatómicas de la administración de fármaco intramuscular; administración de medicamentos y administración de oxígeno para usuario pediátrico).

La recolección de datos se obtuvo mediante una encuesta semiestructurada aplicada a 53 estudiantes, con preguntas abiertas orientadas a develar por parte de los estudiantes las ventajas, desventajas y sugerencias, la cual fue respondida al final de la evaluación. “Las preguntas orientadoras se referían a: de acuerdo con su experiencia vivida en esta evaluación ¿puede mencionar ventajas?, ¿desventajas?; ¿puede emitir sugerencias para mejorar este examen?” (Illesca, Cabezas, Romero y Diaz, 2012).

Las autoras reportan que encontraron como desventaja, el tiempo de espera para ser evaluados, el tiempo establecido en cada una de las estaciones para demostrar competencias y el momento del semestre para su realización (el cual se planificó una vez que finalizaron todos los estudiantes su práctica clínica). Sin embargo, para el momento y el lugar en que se realizó la OSCE el tiempo de espera no pudo ser menor por la cantidad de estudiantes del curso, lo que se podría revertir si se duplicaran las estaciones, pero por motivo de espacio físico esto resulta casi imposible (Illesca, Cabezas, Romero y Diaz, 2012).

Illesca, Cabezas, Romero y Diaz, (2012) acertaron como ventaja, que la prueba permite el desarrollo de habilidad mental, sistematizar procesos, identificar debilidades

y fortalezas y mejorar errores. Y además destacan positivamente la calidad de la retroalimentación recibida posteriormente por el evaluador. Esta investigación demuestra una nueva vertiente investigativa entorno a la aplicación de la OSCE y muestra el reconocimiento por parte de los estudiantes como una metodología evaluativa acorde a la formación deseada.

Por otro lado, Alarcón (2013) hace referencia una investigación que realizó para analizar la percepción de alumnos de enfermería en la implementación de la OSCE, para relacionarla con otros procedimientos evaluativos utilizados, similar a lo realizado por Kaliyadan, Sattar, Kuruvilla y Feroze (2014). Se contó con una muestra de 22 estudiantes, para los cuales se preparó ocho estaciones donde el estudiante disponía de siete minutos para resolverlas. Dos estaciones consistían en contestar una pregunta, o dos estaciones eran con pacientes simulados, luego dos estaciones con maniquíes y finalmente dos estaciones con diversos procedimientos. En la evaluación se solicitó a los estudiantes sus opiniones sobre la metodología a través de una encuesta, de la cual se obtuvo resultados favorables, estimando que es un método objetivo y beneficioso para su formación. Sin embargo, se debe indicar que los aspectos menos satisfactorios fueron el tiempo destinado a cada estación y el estrés que genera. De las preguntas abiertas se encontró que un 40,91% declaró que el tiempo asignado fue insuficiente en algunas estaciones y un 31,82% lo consideró como una oportunidad de aprendizaje e igual porcentaje lo estimó estresante durante el inicio de la evaluación (Alarcón, 2013).

Por último, sin restar su importancia, dentro de los estudios hallados que utilizan la OSCE, se encuentra el de Díaz-Plasencia, et al. (2016), los cuales efectuaron una investigación con 117 estudiantes para determinar la correlación bivariada entre las puntuaciones del examen teórico, la práctica clínica, el aprendizaje virtual y la OSCE de la serie total con el portafolio. A cada estudiante se le solicitó la elaboración de un portafolio de evidencia de práctica clínica, en función de la competencia por demostrar, en donde se definieron previamente las tareas clínicas y los apartados que se debían realizar, y cada alumno decidió qué padecimiento abordar y qué fuentes bibliográficas consultaría para sustentar la toma de decisiones clínicas realizadas. Este portafolio se

estructuró en tres partes: actividades registrales que acreditaron las habilidades trabajadas por el alumno y el nivel de profundidad con que las trabajó, actividades realizadas para la planificación (autoevaluación) y reflexión.

En los resultados de este estudio se evidenció que el portafolio se correlacionó directamente con la nota teórica, la práctica clínica, el aprendizaje virtual y la OSCE obtenidos al final del curso de Cirugía I, lo cual muestra que esta estrategia metodológica de evaluación tiene validez concurrente con otros formatos de evaluación que consideran aspectos cognitivos y procedimentales. Hubo correlación significativa entre el portafolio con el examen teórico ($r = 0,410$; $p = 0,0001$), la práctica clínica ($r = 0,258$; $p = 0,003$). El examen teórico ($r = 0,38$; $p = 0,0001$) y la OSCE ($r = 0,33$; $p = 0,0001$) se correlacionaron con el portafolio; no fue así con el caso clínico virtual ($r = 0,13$; $p = 0,122$). Hubo correlación significativa entre el aprendizaje autorreflexivo ($r = 0,305$; $p = 0,0001$) y la nota teórica final del curso. La fiabilidad interevaluador del portafolio fue significativa en: caso clínico real ($\alpha = 0,486$; $p = 0,006$), incidente crítico ($\alpha = 0,702$; $p = 0,0001$), aprendizaje autorreflexivo ($\alpha = 0,664$; $p = 0,0001$) y estructura del lenguaje ($\alpha = 0,431$; $p = 0,017$) (Díaz-Plasencia, et al., 2016a).

Estos mismos autores indagaron sobre la validez concurrente de la OSCE con el promedio ponderado, la nota teórica y el portafolio electrónico (que incluía actividades registrales, autoevaluación personal, práctica reflexiva, y estructura y lenguaje escrito) en 123 estudiantes de medicina. Para ello, realizaron cuatro estaciones de cinco minutos cada una, dos para la rotación de especialidades quirúrgicas y otros dos para cirugía general y abdominal. Cada docente aplicó una lista de cotejo o escala de verificación, y al final realizaron la retroalimentación. Además, cada estudiante al inicio de cada circuito introdujo información en un portafolio semiestructurado diseñado en forma electrónica, cuya construcción y entrega de actividades se realizaron al término de cada rotación (Díaz-Plasencia, et al., 2016b).

Los resultados mostraron que hubo correlación bivariada aceptable ($r = 0,65$) entre la nota teórica y la OSCE; correlación moderada ($r = 0,52$) entre el promedio

ponderado y el ECOE; y correlación alta (r = 0,77) entre la nota del portafolio electrónico y el ECOE. Hubo correlación aceptable (r = 0,65) entre la nota teórica y la OSCE estructurada. Hubo correlación moderada (r = 0,52) entre el promedio ponderado previo y el ECOE (Fig. 2). Hubo correlación alta (r = 0,77) entre las notas del portafolio y el ECOE (Díaz-Plasencia, et al., 2016b).

Los dos estudios mencionados muestran que el portafolio se puede utilizar en el pregrado porque es fiable y tiene validez predictiva y concurrente, de manera que permite que el estudiante reflexione sobre sus necesidades de aprendizaje (Díaz-Plasencia, et al., 2016a). Y también evidencia que el portafolio electrónico y la nota teórica utilizados para evaluar a los estudiantes de un curso de cirugía de pregrado tuvieron validez concurrente en el rango de aceptable a alta, y esta información constituye la base para mejorar los estándares de evaluación (Díaz-Plasencia, et al., 2016b).

Aunado a lo anterior se destaca el trabajo de Tetzlaff, Dannefer y Fishleder (2009), los cuales diseñaron e implementaron un portafolio de evaluación de competencias en: investigación; conocimiento médico en lo básico, clínico y ciencias sociales; comunicación; profesionalismo; desarrollo personal; habilidades clínicas; razonamiento clínico; sistemas de salud; y práctica reflexiva. En este documento cada estudiante desarrolló planes de aprendizaje de acuerdo a sus fortalezas y debilidades. También elaboraron ensayos y presentaron la retroalimentación que daban a la evidencia que seleccionaban de su hoja de evaluación (todo esto supervisado por un profesor). Este estudio mostró que los residentes fueron capaces de llegar al mismo nivel de evaluación técnica que sus profesores. La ventaja agregada de este proceso es el aprendizaje adicional del acto de la auto-evaluación y el sentido de responsabilidad de los estudiantes por su propio aprendizaje. Reflexionar acerca de casos desafiantes combinados con mantener datos en una libreta y la retroalimentación de un tercero mejoró las habilidades de auto-evaluación.

Siguiendo la evaluación entre pares, se destaca es estudio de Speyer, Pilz, Van y Wouter-Brunings (2011), los cuales utilizan este método para estimular a estudiantes en la participación de actividades educativas y para mejorar el rendimiento del equipo o determinar el esfuerzo individual. Este tipo de evaluación promueve en los estudiantes una actitud crítica, ya que al juzgar a sus compañeros pueden obtener una idea en su propia actuación. De acuerdo con Gielen (citado en Speyer, Pilz, Van y Wouter-Brunings, 2011) la evaluación de pares tiene cinco objetivos principales: el uso de la evaluación por pares como herramienta de evaluación y herramienta de aprendizaje, la instalación de control en el entorno de aprendizaje, la preparación de estudiantes para el autocontrol y la autorregulación en toda su vida aprendizaje, y la participación activa de los estudiantes en el aula.

V.2. Mini-Entrevista Múltiple

La Mini-Entrevista Múltiple (en adelante MMI, por sus siglas en inglés) es una adaptación de la OSCE y comparte algunas de sus características, como la gran cantidad de personal que se requiere para la evaluación del proceso, lo cual viene a ser el costo más elevado de la prueba, por el tiempo que se requiere de estos evaluadores, el cual tiene un costo elevado (Rosenfeld, et al., 2008).

La mini-entrevista múltiple (en adelante MMI) según Eva y Macala (2014) es un proceso de evaluación que se utiliza para facilitar la selección de residentes. Esta técnica se realiza por medio de entrevistas en diferentes escenarios, que tiene por estrategia reunir y recolectar información a través de una serie de breves observaciones independientes. De acuerdo con Rosenfeld, et al. (2008) hay tres tipos generales de

estaciones: discusión, habilidades interpersonales y cooperación. Con respecto a la estación de discusión, el escenario contiene problemas que cualquiera que aspire a la escuela de medicina debería ser capaz de considerar y discutir, por lo que, con múltiples estaciones, se puede abordar una amplia gama de problemas. Por otro lado, el escenario de habilidades interpersonales expone a la o el solicitante a una situación emocional, la cual en posible medida debe ser resuelta. Y la última estación que evalúa la cooperación, es aquella en la que dos solicitantes deben completar una tarea conjuntamente, lo que requiere trabajo en equipo, en este caso cada candidato(a) tiene una persona evaluadora independiente que le observa y califica.

Las estaciones que contienen las MMI requieren alrededor de ocho minutos de tiempo de contacto (observación del desempeño del solicitante) y dos minutos para completar un formulario de evaluación, variando esto según lo evaluado en cada estación (Rosenfeld, et al., 2008). En esta línea, la duración para completar todas las aplicaciones de las MMI se puede modificar por la cantidad de personas evaluadas, frente a esto la literatura muestra que puede fluctuar entre 1 a 11 días, y de hasta 7 sets por día y 7 circuitos por set (Knorr y Hissbach, 2014).

Tomando como base la investigación realizada por Knorr y Hissbach (2014), la MMI cuenta con una gran variabilidad en los diseños en términos de atributos, tipos de estaciones, detalles del proceso (número de estaciones, conjuntos, circuitos y días, tiempo de preparación, duración de la estación, entre otros) y sistema de puntuación (tipo y uso de escalas y subescalas, rango de escala, número de evaluadores) entre las instituciones y también entre los años subsiguientes en la misma institución, por lo cual se demuestra su factibilidad y posibilidad de ajuste de acuerdo al contexto.

Este procedimiento suele aplicarse a estudiantes de avanzados años de estudio, los cuales aplican para residencias o posgrados. El número habitual de estaciones que deben completar cada estudiante oscila entre 6 y 12, algunas de ellas incluyen tareas como: resolución de problemas, priorización de tareas, tareas creativas, clips de películas, entre otras; y cada una se basa en la selección de atributos o características (los

cuales varían entre 3 y 19), considerando que algunos de ellos son más comúnmente usados, como es el caso de las habilidades de comunicación. La evaluación de cada estación, se hace por medio de una sola escala o la combinación de varias subescalas likert con puntajes de 4 a 10 (Knorr y Hissbach, 2014).

A pesar de las diferencias considerables en los diseños de las MMI´s, la confiabilidad es un punto fuerte de este procedimiento, ya que la mayoría de los valores de consistencia interna, así como la generalización general son satisfactorios (Knorr y Hissbach, 2014; Shih-Chieh Liao, et al., 2014; Zaidi, et al., 2014; Rahim y Yusoff, 2016). El enfoque multiestación, que es el elemento central de todos los formatos MMI, permite lograr un nivel satisfactorio de confiabilidad aumentando el número de estaciones (Knorr y Hissbach, 2014).

La MMI es una prueba que ha demostrado ser capaz de generar datos confiables predictivos del éxito, su confiabilidad de los diferentes tipos de MMI y su eficiencia en el uso del tiempo de entrevista, respalda fuertemente la superioridad del método sobre las técnicas de entrevistas convencionales (Rosenfeld, et al., 2008; Zaidi, et al., 2014; Dowell, et al., 2012; Knorr y Hissbach, 2014). No obstante, con respecto a la validez de la MMI, la tendencia hacia la falta de relación con medidas predominantemente académicas y correlaciones débiles a moderadas con menos medidas académicas sugiere que esta prueba no puede reemplazar las herramientas de admisión tradicionales, sino que miden algo diferente (Knorr y Hissbach, 2014).

Por otro lado, los factores más relevantes con respecto a costos comparados con entrevistas convencionales son los costos de la implementación de la estación y el pago a las personas autoras. No obstante, la ventaja de este formato es especialmente su eficiencia para reducir el tiempo de entrevista. De este modo, permite un mayor número de candidatos a ser entrevistados en un período de tiempo más corto (Knorr y Hissbach, 2014).

Entre los estudios encontrados, se halló que la MMI puede ser diseñada para probar cualidades personales como la empatía, el compromiso y el potencial de

liderazgo (Mini-interviews help to recruit students for their values, 2013); la comunicación (incluyen múltiples perspectivas, reflejo del escenario, articulación, interés en el dilema, comunicación no verbal y habilidades interpersonales) (Zaidi, et al., 2014); el razonamiento crítico, habilidades de comunicación, concientización ética, y conocimiento del sistema de salud, preguntas estándar, desempeño en el idioma, y conducta en general (Rahim y Yusoff, 2016); empatía, respeto por la vida, gestión de crisis, iniciativa, perspicacia, integridad y habilidades de comunicación (Shih-Chieh Liao, et al., 2014).

En línea con lo anterior, Rahim y Yusoff (2016) implementaron la MMI en circuitos de nueve estaciones de siete minutos: cinco de entrevistas y cuatro de descanso. Cada estación evaluó el razonamiento crítico, habilidades de comunicación, concientización ética, y conocimiento del sistema de salud, preguntas estándar, desempeño en el idioma, y conducta en general. Para disipar el aburrimiento de los entrevistadores se les permitió cambiar de estación al haber completado una sesión, lo que hizo difícil cualquier conclusión basada en las diferencias entre sesiones difíciles, por lo tanto, en próximas aplicaciones se les pedirá a los entrevistadores permanecer en la misma estación.

Siguiendo el estudio anterior, es relevante resaltar que participaron como entrevistadores personas no académicas como miembros del cuerpo médico de hospitales y enfermeros, ya que son ellos quienes van a recibir a las y los estudiantes en los hospitales. Al final del proceso estas personas fueron positivas con respecto a su capacidad para evaluar a los candidatos de manera justa. La confiabilidad general del ejercicio arrojó un 0.94, lo cual indica un alto nivel de consistencia interna, siendo así un aceptable nivel para tomar decisiones de alto riesgo. Por otro lado, los candidatos fueron bastante positivos respecto a la calidad e implementación de la prueba, donde la media de las evaluaciones estuvo todas por encima de 5 de un total de 7. Para este estudio hubo 447 participantes y 30 entrevistadores por sesión (Rahim y Yusoff, 2016).

Considerando otros autores que han utilizado la MMI, Zaidi, et al. (2014), presentan en que la validez de constructo de esta técnica con respecto a la interacción ítem-escenario corresponde al 1.4% del total de varianza. Esta baja estimación de varianza atribuible a la faceta de ítem es reforzada por el alfa de Cronbach (0.97) para los siete ítems que se contempló en dicha investigación, lo cual sugiere consistencia interna muy alta entre los atributos medidos por la MMI, estos autores mencionan que la validez de constructo de la MMI ha sido menos estudiada en comparación con la validez predictiva, a pesar de la necesidad de examinar todas las propiedades psicométricas de un proceso de medición que es utilizado para fines de admisión de alto riesgo (Zaidi, et al., 2014). Cabe destacar que el estudio realizado por Axelson y Kreiter (2009), muestra la ventaja de la MMI sobre las entrevistas de admisión tradicionalmente aplicadas en las Escuelas de Medicina.

Apoyando las publicaciones antes presentadas se destaca la investigación efectuado por Gafni, et al. (2012) ya que se ha interesado en proveer estimaciones confiables para la MMI, mediante el estudio de estaciones de simulación y estaciones de entrevista. Para esto contaron con la participación de 2662 estudiantes en MOR (estaciones de simulación) y de 2023 personas en MIRKAM (estaciones de entrevistas).

Para MOR se contó con nueve estaciones individuales (6-9 min c/u) y 2 estaciones grupales (30 min c/u). Tres estaciones de simulación son de interacción entre candidato y un paciente estándar; dos estaciones de información donde se les aplicó una entrevista acerca de su desempeño en la estación de interacción con paciente estándar; una entrevista personal estandarizada sobre lo que piensa el candidato acerca de la profesión médica y los actuales problemas en políticas médicas. En cambio, en las estaciones grupales, se requirió de 6 miembros donde los candidatos se enfrentaron a situaciones interpersonales e intra-grupales. Para la evaluación de los candidatos se utilizaron formularios de evaluación estructurados donde en algunos casos incluían escalas de 1-6 y algunas otras veces eran evaluados por un profesor o por un paciente estándar. Para este proceso participó un total de 348-386 profesores y 33-35 de pacientes estándar cada año (Gafni, et al., 2012).

Para MIRKAM se contó con ocho estaciones de entrevistas (10 minutos cada una(. También hubo tres estaciones de semi-simulación donde se daba un role-play entre entrevistador y candidato. En dos estaciones se le presentaba al candidato un dilema médico ético a discutir. En tres estaciones, se le preguntaba al candidato sobre su historia biográfica (Gafni, et al., 2012).

Para tener los resultados de las pruebas se aplicó un cuestionario de juicio y toma de decisiones que consistía en tres pruebas a forma de ensayo donde los candidatos describían dilemas éticos de la vida real. Los argumentos dados por los candidatos se contaban, y cada argumento recibía de 0-2 puntos basándose en cómo se relacionaba a los valores generales, profesionalismo, y consideraciones morales, de toda esta información se sacó un promedio ponderado de los ensayos. También se aplicó un cuestionario biográfico que consistió en 20 preguntas abiertas enfocadas a la experiencia de vida del candidato y a la concientización emocional de este. Cada pregunta se evaluó en una escala de 1-5. (Gafni, et al., 2012).

Se calcularon coeficientes de generalidad para las estaciones cada año, coeficientes de confiabilidad para todos los centros en general, coeficientes para aquellos que volvían a tomar la prueba, y la correlación de coeficientes entre los centros. Con respecto a los resultados, la generalizabilidad de las estaciones de comportamiento para todas las pruebas fue de 0.73 para las estaciones de comportamiento de MOR. Los coeficientes G son similares al rango reportado por MMI de 0.65 - 0.81. En general las estimaciones tienden a ser más altas para las estaciones de comportamiento MOR que para las estaciones de comportamiento MIRKAM (Gafni, et al., 2012).

Del mismo modo tanto MOR como MIRKAM comprenden estaciones y dos cuestionarios. Tal y como se esperaba, al aumentar el número de estaciones a 14, se encontró que estaba asociado con un incremento en la confiabilidad de un 0.67 (MIRKAM) o 0.76/0.69 (MOR) a un 0.81/0.77 a lo largo de los años. La correlación entre MOR y MIRKAM de los dos cuestionarios en las dos estaciones diferentes son similares (0.28 y 0.23 entre JDQ y las estaciones MOR y MIRKAM respectivamente; 0.51 y 0.47 entre

BQ y las estaciones de comportamiento de MOR y MIRKAM, respectivamente). El coeficiente de la correlación Pearson entre MOR y MIRKAM a lo largo de los 4 años fue de 0.56 (Gafni, et al., 2012).

De acuerdo con los artículos analizados, se destaca a continuación algunos estudios que han incluido en la MMI otro tipo de evaluación, lo cual ha enriquecido el proceso de evalaución:

Shih-Chieh Liao, et al. (2014), muestran que la MMI proporciona un enfoque válido, confiable y defendible para la selección de estudiantes de medicina; no obstante, estas no son suficientes para evaluar las habilidades interpersonales. Por lo tanto, además del circuito de siete estaciones que evaluaron empatía, respeto por la vida, gestión de crisis, iniciativa, perspicacia, integridad y habilidades de comunicación; se incorporó al diseño de la MMI una entrevista grupal, la cual consistió en una sesión de discusión entre compañeros(as). Esta innovación creó una entrevista de selección de estudiantes de medicina más efectiva, sus resultados mostraron que la inclusión de la entrevista grupal con MMI aumentó el alfa de Cronbach de 0.54 a 0.63. Lo que demuestra que la combinación aumentó la consistencia interna de toda la entrevista y la hizo más válida.

Eva y Macala (2014), incluyeron en el proceso una estación de estilo libre, en la cual se permitió a las personas evaluadoras realizar cualquier tipo de pregunta que les ayudara a generar un puntaje con respecto al rol profesional de la persona aspirante. Esto además de las estaciones de juicio situacional y la estación de entrevista de comportamiento. La primera de estas pedía a las y los participantes que imaginaran cómo reaccionarían ante diferentes situaciones planteadas, con el fin de evalaur las habilidades de comunicación, de razonamiento, y profesionalismo del candidato. Y en las estaciones de entrevista de comportamiento, debían pensar en una situación que ellos o ellas hubieran experimentado y que fuera análoga con la situación presentada en la estación, y debían discutirlo con los examinadores.

La correlación entre el promedio de las estaciones dentro de cada tipo del promedio de las 9 estaciones MMI usadas para admisión fue: estación de juicio de situación r = 0.45; estaciones de entrevista de comportamiento r = 0.57, y estación de estilo libre r = 0.42. En general los candidatos consideraron las estaciones de estilo libre fueron más desafiantes, menos claridad en sus respuestas, y más difíciles que el resto de las estaciones y les provocaban más ansiedad que las otras estaciones, situación que las personas evaluadoras no percibieron (Eva y Macala, 2014).

Dowell, et al. (2012) consideraron cuatro estaciones: tres de las cuales eran entrevistas tradicionales uno-a-uno y la cuarta consistía en una evaluación interactiva donde se hacía un role-play. El resultado de este estudio evidenció que de una muestra de 452 de participantes en el 2009 y 477 entrevistados en 2010, el 98% de las y los aspirantes indicaron que el uso de la MMI mejoró su punto de vista de la escuela de medicina de Dundee; el 90% de las personas entrevistadas están totalmente de acuerdo en que la MMI fue justa; el 75% de las personas encuestadas consideraron que las estaciones lograron lo que se propusieron hacer; una gran mayoría (74%) de los que ya fueron entrevistados en otro la escuela de medicina indicó una preferencia por la MMI; el 60% mencionó que recomendarían Dundee a un amigo debido a la MMI; y una minoría considerable de entrevistados (33%) consideró que el MMI fue más estresante que las entrevistas tradicionales.

Eva, et al. (2009) realizó un estudio para comparar puntajes MMI con aquellos obtenidos en pruebas de habilidades clínicas a nivel nacional. En primera instancia se evaluó la estabilidad del desempeño, para lo cual se tomó una muestra de 29 estudiantes residentes, con 9 estaciones MMI. En este proceso los participantes fueron asignados a uno de los 3 circuitos que corrían paralelos. Cada circuito estaba compuesto por estaciones de 10 minutos cada una, en donde había un evaluador en cada una. Posterior a cada desempeño, los participantes eran evaluados en una hoja que contenía 4 ítems con 7 puntos a evaluar. En total participaron 18 evaluadores a lo largo de los 3 circuitos (9 por día). La confiabilidad de cada estación fue de G=0.24, en donde la confiabilidad del puntaje total generada por la ponderación de las 9 estaciones

fue de 0.76. Se sugiere que una evaluación con 12 estaciones podría subir el nivel de confiabilidad a 0.80. La confiabilidad de cualquiera de las estaciones por separado es baja, generalmente <0.25, en donde el promedio para 12 estaciones en total se ha encontrado ser de 0.73.

Posteriormente, se evaluó la validez predictiva del MMI. En esta línea, los estudios previos han demostrado una relación complementaria entre la capacidad de predicción de GPA (promedio ponderado total) y la MMI. La prueba de dos partes MCCQE es una prueba a nivel nacional que todos los estudiantes de medicina deben pasar para licenciarse en el Colegio de Médicos de Canadá. Para evaluar la predicción se correlacionó los puntajes obtenidos en la MMI del estudio de posgrado realizado en la sección anterior con aquellos obtenidos en la parte II del MCCQE. Para esta parte de la investigación participaron 117 aplicantes en el 2002 quienes habían participado en la MMI además de otros 4 procesos de admisión. Estos estudiantes fueron entrevistados ante un panel de 3 jueces y luego participaron en una MMI de 10 estaciones. Los resultados de este apartado muestran que la MMI fue estadísticamente predecible respecto al porcentaje de estaciones que los candidatos pasaron en la segunda parte de la prueba MCCQE ($r = 0.43$, $P < 0.05$) y la cual tendió a ser estadísticamente predictiva del puntaje total ($r = 0.36$, $P < 0.1$). Sin embargo, se encontró que la MMI fue el único predictor estadístico del porcentaje de estaciones pasadas por los candidatos en la parte II de MCCQE (Eva, et al., 2009).

V.3. Mini-CEX

Según Fornells-Vallés (2009), “el Mini-CEX se puede definir como un método de observación directa de la práctica profesional con evaluación estructurada mediante un formulario de ésta y posterior provisión de feedback a la o el residente/estudiante” (Fornells-Vallés, 2009, p.83). El Mini-CEX puede ser usado para evaluar habilidades de entrevista clínica, habilidades de exploración física, profesionalismo, juicio clínico, habilidades comunicativas y organización/eficiencia. Entre las características que tiene

Mini-CEX está: da un feedback inmediato a la persona evaluada, se basa en casos clínicos con pacientes reales, en distintos entornos y con diversos observadores para cada caso y su duración es corta (30 aproximadamente).

De acuerdo con la información recolectada para el presente estado de la cuestión, se encontró que las investigaciones evaluaron el desempeño de estudiantes residentes y pasantes: 27 estudiantes (Baños, et al., 2015), 8 residentes (Fernández, 2011), 12 residentes (Malhotra, Hatala y Courneya, 2008) y 20 pasantes (Brazil, et al., 2012). Por ello, estos resultados no puedan ser extrapolados a otras categorías ni son generalizaciones, por lo tanto, sus derivaciones aplican solo para un análisis descriptivo del presente estado de la cuestión con la muestra en mención.

Los aciertos del Mini-CEX varía según la postura de diversos autores. Por su parte Brazil, et al. (2012) encontraron que la satisfacción en general de internos y asesores fue alta y la mayoría sintió que la prueba era buena, pero como un entrenamiento adjunto, e indicaron que el punto fuerte de la evaluación fue la mejora en la observación directa como medio de transmisión de la evaluación. Lo cual concuerda con Baños, et al., 2015, quienes aluden que la media de la satisfacción de la experiencia de tutores y estudiantes en el estudio que realizaron siempre fue superior a 8.5. Según estos autores el mini-CEX muestra la viabilidad como método de evaluación de competencias clínicas en lo que se refiere a tiempo de dedicación y satisfacción de los participantes. Así mismo Fernández (2011), considera la prueba como confiable, ya que múltiples encuentros, distintos observadores, con diferentes pacientes y situaciones clínicas permiten obtener conclusiones sobre la competencia clínica global.

En la misma línea se identifican otros elementos positivos de la prueba: su validez ya que permite diferenciar entre diferentes niveles de experiencia entre los residentes (Fernández, 2011); su efectividad en identificar los dominios en los cuales los internos tenían deficiencias (Brazil, Ratchiffe, Zhang, y Davin, 2012); y la flexibilidad

para adaptarse a las características específicas de las y los estudiantes que se deseen evaluar (Baños, et al., 2015).

Entre los beneficios del mini-CEX, está su gran impacto educativo, por lo significativo de la devolución constructiva en el aprendizaje de la o el futuro residente (Fernández, 2011); también puede tener otros efectos, ya que entre más veces se repite la experiencia, las y los residentes pueden sentirse más confiados y tranquilos, lo que induce a un mayor provecho a los beneficios educacionales de la prueba (Malhotra, Hatala y Courneya, 2008).

Por otro lado, Donato, et al., (2008) evaluó si un nuevo formulario de evaluación puede mejorar la precisión de miembros de facultad en la detección de desempeños insatisfactorios, generar más observaciones del evaluador y mejorar la calidad de la retroalimentación, para lo que utilizaron Mini CEX para grupo control y la Minicard grupo de intervención. Para esta investigación se registró que el grupo control tenía un ítem más de retroalimentación que el grupo de intervención, sin embargo, este último, tuvo casi el doble del total de observaciones. El grupo de intervención obtuvo resultados más precisos distinguiendo entre desempeños satisfactorios de insatisfactorios, pero menos precisos al identificar desempeños satisfactorios. La Minicard mejoró la precisión general (85% vs 73%), especialmente para desempeños deficientes (96% vs 52%), incrementó el total de observaciones totales (10.8 vs 5.7) e incrementó la confiabilidad a nivel de pasar/quedarse (0.520 vs 0.299) comparado con el formulario ABIM Mini-CEX. Por lo que en este caso la minicard tiene un resultado más eficiente en comparación con el Mini-CEX.

Lo que lleva a plantear algunas de las limitaciones de este método de evaluación: la influencia/afectación por ansiedad de los residentes, al sentirse observados lo que puede cambiar su desempeño académico (Malhotra, Hatala y Courneya, 2008); dificultades prácticas al organizar y conducir las evaluaciones mini-CEX en el sitio de trabajo, así como el formato de evaluación que se efectúe, ya que puede no abarcar todos los dominios de evaluación, y también la inversión monetaria que representa su

aplicación, por el tiempo que los asesores invierten (Brazil, et al., 2012); la amplia variabilidad observada sugiere la necesidad de acotar mejor los objetivos del mini-CEX para evitar una dispersión innecesaria que puede amenazar la viabilidad de la prueba (Baños, et al., 2015); la diferencia del porcentaje obtenido según los diferentes escenarios clínicos donde se desarrolle la observación y las diferencias de criterios de las y los docentes, lo que pueden afectar la confiabilidad del método de evaluación, lo que puede ser mitigado al aumentar el número de docentes que realizan los encuentros, ya que esto disminuye la variabilidad interobservador (Fernández, 2011).

En síntesis, el Mini-CEX tiene un impacto educativo por la devolución constructivista que fomenta, ésta mejora del aprendizaje, favorece la motivación y las habilidades de autorregulación, es decir la capacidad para reconocer debilidades y fortalezas, y así identificar áreas que requieran mejor desempeño. En este proceso el personal docente puede evaluar la capacidad de autocrítica y de reflexión de la o el residente, y guiarle para que la desarrolle. De esta manera se estimula la capacidad de aprender a aprender, que es quizás, la mejor enseñanza que se le puede brindar a la o el residente, ya que esta capacidad le acompañará durante su quehacer profesional.

V.4. Pruebas escritas

El uso de pruebas escritas y de opción múltiple se han utilizado tradicionalmente en la educación superior. Findyartin, Wedhani, Iryani, Rini, Kusumawati, Poncorini y Primaningtas (2015) hacen una primera aproximación al estudio científico de las mismas analizando la Prueba Colaborativa de Progreso (cPT). Findyartin et al. (2015) informan que esta prueba se ha empezado a utilizar para permitir mayor control a los centros de enseñanza sobre el aprendizaje correcto de los conocimientos médicos. El grupo de investigación indica que, en su país, Indonesia, se ha producido un aumentó de 32 a 72 escuelas en los últimos 6 años. Por lo tanto, ahora más que nunca es esencial apelar a la garantía de calidad en todas las escuelas de medicina.

La prueba de progreso permite la evaluación comparativa del enfoque curricular y el rendimiento de los estudiantes en las escuelas de medicina. Findyartin et al. (2015)

mencionan que este también puede ser el objetivo del examen final de acreditación egreso, sin embargo, la prueba de progreso permite una evaluación comparativa más continua de la efectividad curricular y la identificación de problemas en cada escuela desde etapas tempranas.

La cPT es una prueba creada por tres escuelas de Medicina de Indonesia para iniciar con el esfuerzo de garantizar la calidad continua de los planes de estudios médicos. El cPT consiste en 120 ítems de opción múltiple con cinco opciones: una única respuesta correcta. Los autores se proponen como objetivo principal evaluar la validez y la confiabilidad del cPT realizado en las tres escuelas de medicina como parte de una evaluación curricular. Para esto se administró la prueba a estudiantes de primero a quinto año de la carrera en las tres escuelas de medicina, con una participación mínima de 200 estudiantes por escuela (Findyartin et al., 2015).

Findyartin et al (2015) indican que la validez de una evaluación se refiere a qué tan bien la evaluación mide lo que "pretende medir" y dado que el objetivo de la cPT es medir el progreso del conocimiento de los estudiantes de acuerdo con el nivel del año, este estudio intentó proporcionar la evidencia de validez de una manera sistemática. En cuanto a la evidencia de confiabilidad en este estudio se desarrolló con base en la teoría de la prueba clásica.

Los estudiantes fueron muestreados aleatoriamente de los tres grupos de promedio de calificaciones: (a) <3.25, (b) 3.25-3.50 y (c) >3.50. (en base a 4), basado en el de promedio de calificaciones los estudiantes fueron seleccionados al azar por año. Un total de 223, 219 y 161 estudiantes de 1-5 año de carrera participaron en el cPT. Alrededor de dos tercios de la población eran mujeres. La edad promedio de los estudiantes en las tres instituciones que se analizaron fue de 20 años. La prueba cPT utilizada cubría los principales intereses de la Medicina, según los creadores, las habilidades clínicas básicas (categoría 1), aspectos cognitivos (categoría 2), aspectos de razonamiento (categoría 3), desarrollo y degeneración del crecimiento (categoría 4)

y temas de inmunología e infección (categoría 5), terapias y diagnóstico (categoría 6) y temas individuales de gestión de la salud (categoría 7).

Hubo un aumento del puntaje promedio del año 1 al 5, tanto en los datos combinados (ANOVA de una vía F 174.7 (4), p <.001) y en cada escuela (Faculty of Medicine Universitas Indonesia=ANOVA de una vía F 102.5 (4) p <.001, Faculty of Medicine Universitas Andalas= F 83.0 (4) p <.001, Faculty of MedicineUniversitas Sebelas Maret =F 28.28 (4) p<.001). En cuanto a la correlación de el puntaje de la prueba con el promedio de calificaciones se observó la mayor correlación en los estudiantes que cursaban quinto año de carrera. Entre otros resultados importantes a destacar se menciona que la consistencia interna del cPT fue muy buena en las tres instituciones. La evidencia de la validez de constructo de la cPT fue respaldada por el aumento constante de las puntuaciones medias en el año 1 a 5 con una diferencia estadísticamente significativa.

Otra investigación realizada por Salazar, Veléz y Tobón (2015) nos muestra el efecto de la reducción del número de opciones de respuesta por pregunta sobre los indicadores psicométricos de un examen de ingreso a estudios médicos de posgrado. Se tuvieron en cuenta los exámenes de 2.539 aspirantes a ingresar a 21 programas de posgrado clínico-quirúrgicos de la Facultad de Medicina de la Universidad de Antioquia, Medellín, Colombia, en el año 2014. La prueba consta de 70 preguntas de selección múltiple constituidas por un tallo con la descripción de un caso clínico de cualquiera de las especialidades a las cuales aspiran los evaluados y cuatro alternativas de respuesta, que pueden ser de dos tipos: el primero con una sola respuesta verdadera, y el segundo, con todas las opciones de respuesta verdaderas, pero con una de ellas más adecuada que el resto para la situación clínica específica.

Estos autores justifican que es frecuente que en las pruebas con cuatro o cinco opciones de respuesta se incluyan una o dos alternativas muy obvias o poco razonables por el simple hecho de cumplir con la directriz general del número de opciones. Además, estas opciones traen el resultado de aumentan el tiempo de elaboración de la prueba

por parte del docente y el de lectura para el evaluado. Por lo tanto, las pruebas con tres opciones de respuesta ofrecen la ventaja de ser más fáciles de construir para los docentes, con menor riesgo de incluir alternativas inadecuadas y menor tiempo de lectura, lo cual posibilita el aumento del número de preguntas con lo que se logra una mayor cobertura de contenidos y mayor confiabilidad o reproducibilidad de la prueba (Salazar, Veléz y Tobón, 2015).

Entre los resultados encontrados es importante indicar que solo 52,9% de las preguntas tuvieron tres opciones funcionales de respuesta, además no se encontró diferencia en la dificultad, la discriminación, el error estándar de la medición, el alfa de Cronbach, ni el coeficiente de correlación biserial, tampoco en la medida de dificultad de los ítems, teoría de respuesta al ítem, entre las pruebas con tres y cuatro opciones de respuesta. Estos hallazgos permiten cuestionar la composición de las pruebas que se aplican para ingreso a posgrado, ya que la transición hacia pruebas de tres ítems parece no influir en demasía para la admisión (Salazar, Veléz y Tobón, 2015).

Delgado y Sánchez (2012) realizan una investigación sobre el examen profesional de la Facultad de Medicina de la UNAM, esta es la evaluación sumativa más importante de la carrera de médico cirujano en México. Una fuente de evidencia de validez del examen es el análisis psicométrico de los reactivos, para el que tradicionalmente se ha utilizado la Teoría Clásica de los Test (TCT), la cual tiene algunas desventajas, que la Teoría de Respuesta al Ítem (TRI) pretende resolver. Por lo tanto, se reporta el análisis del examen aplicado en el año 2008 con la TRI. Este examen está compuesto por seis áreas de conocimiento: Medicina interna, Pediatría, Gineco-obstetricia, Urgencias médicas, Cirugía y Medicina familiar, evaluadas con 420 reactivos de opción múltiple.

Para esta investigación se calculó la confiabilidad, la dificultad y la discriminación con la TCT. Se utilizó el modelo de tres parámetros de la TRI. Con las dos aproximaciones se seleccionaron los mejores ítems, y se estimó la longitud de la prueba con la fórmula de Spearman-Brown. Entre los resultados más destacables se debe

mencionar que fue respondido por 882 sustentantes, tuvo un índice de dificultad de 0.55 y una confiabilidad de 0.93. Con el modelo de 3pl-TRI, el examen es informativo en niveles de habilidad cercanos al promedio en la escala theta. Además, se encontró que el parámetro de discriminación fue 0.67, el parámetro de dificultad fue 1.21, y el parámetro de seudoadivinación fue 0.18. Se encontró que es posible reducir el número de reactivos de la prueba y aun así mantener una alta confiabilidad. Este tipo de investigación permite identificar el posible ahorro al eliminar ítems innecesarios y forma parte de las evaluaciones necesarias para pruebas de admisión (Delgado y Sánchez, 2012).

Rivera, Flores, Alpuche y Martínez (2017) se enfocan en la adecuada elaboración de los reactivos de un examen. Los autores mencionan que, a pesar de existir un consenso general sobre las recomendaciones en la elaboración de un buen reactivo, hay diferentes estudios publicados que reportan una alta incidencia de fallas en el apego a las mismas. Por lo tanto, se propone un instrumento para evaluar la calidad en la elaboración de reactivos de opción múltiple y se describe el proceso de obtención de evidencias de validez.

En esta investigación se calculó el índice Kappa (por el modelo propuesto por Fleiss) y la correlación punto-biserial de Pearson para medir la concordancia en los diferentes criterios que evalúa el instrumento. Se realizó un análisis factorial exploratorio para identificar las dimensiones del instrumento y se calculó el alfa de Cronbach como estadístico de consistencia interna. Entre los resultados más importantes se puede mencionar que la concordancia entre múltiples jueces tuvo un valor mayor de 0.8 (acuerdo casi perfecto) para 12 de los 21 criterios, y de 0.19 para el nivel taxonómico. El análisis factorial definió 4 dimensiones con un KMO = 0.666, ($p < 0.01$), una varianza total explicada de 49.979%, y un alfa de Cronbach de 0.627. Los anteriores resultados son positivos y demuestran que el instrumento desarrollado puede ser aplicado para la evaluación de reactivos de opción múltiple, ya que cuenta con evidencia de validez relacionada con el contenido, el proceso de respuesta y estructura

interna y los indicadores psicométricos también son adecuados (Rivera, Flores, Alpuche y Martínez, 2017).

Sharma, Cui, Leighton y White (2012), al igual que Findyartin et al (2015) y Westerkamp, Penninga, Kuks y Shotanus (2013) se interesan por el rendimiento en el entorno clínico. A medida que se amplía la matrícula de la escuela de medicina, también se prevé que mantener una relación de uno a uno puede volverse difícil de mantener para evaluadores y estudiantes.

Este grupo de investigación alude que implementaron un modelo de evaluación basado en el equipo durante la pasantía de tercer año en Cirugía General, Anestesiología y Medicina del Dolor en el año académico 2009-2010. Bajo este modelo, los estudiantes fueron asignados a trabajar con un equipo compuesto por varios médicos, residentes y enfermeras. Como los estudiantes pasaron cortos periodos de tiempo con múltiples miembros del equipo de atención médica se empleó la retroalimentación de múltiples fuentes. A medida que se recopila las opiniones de numerosos observadores se proporciona un medio más exacto para comprender cómo el sujeto está desenvolviéndose en un ambiente de equipo, aspecto que resulta casi imposible de obtener a partir de una evaluación tradicional de un individuo a otro.

Los autores (Sharma, Cui, Leighton y White, 2012) reportan que el instrumento se desarrolló revisando los elementos de evaluación existentes y obteniendo información de los evaluadores y los estudiantes. Durante el proceso los autores reportan el uso de entrevistas y grupos focales para determinar la aceptabilidad de los evaluadores y los estudiantes. En esta investigación se encontró que las estimaciones de consistencia interna para cada formulario de evaluación fueron aceptables (alfa de Cronbach 0,856-0,948). Además, en cuanto a la retroalimentación cada estudiante recibió un promedio de 188 palabras. Las entrevistas revelaron que la mayoría de los estudiantes y evaluadores entrevistados encontraron el método aceptable. Sharma, Cui, Leighton y White (2012) demuestran por medio este estudio que un modelo de evaluación basado en equipos con retroalimentación de múltiples fuentes es una forma

de evaluación viable y aceptable para estudiantes de medicina que están desarrollando su práctica clínica, y tiene algunas ventajas sobre la evaluación tradicional basada en un observador.

Continuando sobre el uso de pruebas, Westerkamp, Penninga, Kuks y Shotanus (2013) mencionan que las pruebas de libro abierto en la evaluación apoyan a los estudiantes en el manejo del cuerpo creciente conocimiento. Durante estas pruebas, los estudiantes pueden consultar referencias cuando sea necesario. Las pruebas de libro abierto permiten reducir la necesidad de memorización de hechos. El equipo de investigación menciona que se desconoce cuánto tiempo es razonable para que un estudiante pueda responder una pregunta de libro abierto. Por lo tanto, se proponen examinar el tiempo que los estudiantes duraron respondiendo preguntas a libro abierto, la cantidad de preguntas por la que los estudiantes consultaron sus referencias (comportamiento de búsqueda) y cómo estas dos variables se relacionan con la nota de la prueba.

A lo largo de todo el programa de bachillerato en Medicina de la Universidad de Groningen, las pruebas de libro abierto se utilizan para evaluar el conocimiento que las y los estudiantes necesitan para comprender y aplicar correctamente. Dos cohortes de estudiantes de medicina participaron en el estudio: 491 estudiantes de segundo año, (la edad promedio fue 20.5) y 325 estudiantes de tercer año (la edad promedio fue 21.3). Para ambos cohortes los exámenes eran de opción múltiple. En segundo año los exámenes contenían ocho preguntas con dos opciones de respuestas 21 preguntas con tres opciones de respuesta y una pregunta con cuatro opciones de respuesta, y en tercer año los exámenes contenían 23 preguntas con dos opciones de respuesta y siete preguntas con tres opciones de respuesta. Se contó con tres horas para finalizar la prueba y para cada pregunta los estudiantes informaron si habían consultado sus referencias para responderlo. Además, Westerkamp, Penninga, Kuks y Shotanus (2013) reportan que midieron el tiempo que estudiantes usaban para responder las preguntas de libros abiertos.

La muestra fue de 142 estudiantes de segundo año (29%) y 61 de tercer año (19%). En cuanto al comportamiento de búsqueda, operacionalizado como el número de preguntas para las cuales los estudiantes consultaron referencias (comportamiento de búsqueda) fue de 26 (87%) para el segundo año y 22 (73%) para los exámenes de tercer año. En promedio, los estudiantes de segundo año tardaron de 3.6 minutos para preguntas con dos opciones de respuesta a 7.2 min para preguntas con cuatro opciones de respuesta. En cuanto a los estudiantes de tercer año Westerkamp, Penninga, Kuks y Shotanus (2013) mencionan que gastaron un promedio de 4.3 minutos por pregunta, que varía de 3,8 min para preguntas con dos opciones de respuesta a 5.8 min para preguntas con tres opciones de respuesta. Es fundamental indicar que no se encontraron relaciones significativas entre el número de preguntas para las cuales los estudiantes consultaron sus referencias y los resultados de sus pruebas (año 2: r0.02, p = 0,80; año 3: r = 0,05; p= 0.73). Tampoco se encontró correlación entre el tiempo que los estudiantes utilizaron para responder todas las preguntas y los resultados de la prueba (año 2: r .0.85; año 3: r 0.07, p = 0.32).

V.5. Procesos de selección

O'Neill, Vonsild, Wallsted y Dornam (2013) argumentan que estudiantes de los estratos sociales más bajos están poco representados en la educación superior y específicamente en las escuelas de Medicina. En Dinamarca, atender esta problemática es de amplio interés para mejorar las estrategias de admisión. En esta investigación se reporta el uso de dos formas de admisión: notas y atributos personales, tales como las habilidades de comunicación verbal y escrita, las habilidades interpersonales y la capacidad de enfrentar adecuadamente el estrés, como criterios de selección complementarios para el ingreso a la carrera de medicina. El objetivo de la investigación de O'Neill, Vonsild, Wallsted y Dornam (2013) fue comparar los efectos de las estrategias de admisión, basadas en notas y basadas en atributos, en la composición social de los estudiantes de medicina admitidos en la Universidad del Sur de Dinamarca durante los años 2002-2007.

Este grupo de autores menciona que en la admisión el componente de habilidades de comunicación se evaluó por medio de un ensayo escrito que valía el 40%, el cual incluía el conocimiento de los cursos y la profesión elegida, reflexiones sobre experiencias pasadas, la elección del estudio y los futuros planes de empleo. También se evaluó, las experiencias laborales anteriores, calificaciones educativas pasadas, experiencias de intercambio y trabajo organizativo o voluntario, todo esto con base en la información proporcionada en el formulario de solicitud que tenía un valor de 60% (O'Neill, Vonsild, Wallsted y Dornam, 2013).

De toda la población se seleccionaron 300 personas, que luego fueron invitados a asistir al proceso de admisión, que consistió en una prueba de conocimiento general y una entrevista de admisión. La primera fue una prueba de opción múltiple de 60 preguntas que se completó en 15 minutos. Finalmente, el interés del sujeto, las expectativas, la edad, las habilidades sociales, la tolerancia al estrés, la empatía y el comportamiento en general se evaluaron en una entrevista semiestructurada de 25 minutos. Posterior a estas evaluaciones se seleccionó a 150 aspirantes con base a un

conglomerado de sus puntajes de información de formularios de solicitud (ponderación: 35%), puntajes de conocimiento general (ponderación: 20%) y puntajes de entrevista (ponderación: 45%). Los coeficientes de generalizabilidad compuestos para la cohorte de ingresos de 2007 se estimaron en 0,45 para el proceso de preselección y 0,82 para el proceso de selección final.

En este estudio se encontró que de los 1074 estudiantes daneses admitidos durante 2002-2007, 454 fueron admitidos a través de la admisión basada en notas y 620 a través de la admisión basada en atributos. Los dos grupos de admisión fueron significativamente diferentes en términos de calificaciones académicas pasadas, y los cuatro puntajes basados en atributos correlacionaron débilmente con calificaciones académicas pasadas ($r < 0.14$). Sin embargo, no hubo diferencias entre los grupos de admisión en ninguna de las variables sociales, que interesaban principalmente (origen étnico, educación de los padres, paternidad, padres que viven juntos, padres con asistencia social), que previamente se han asociado con bajo nivel educativo en Dinamarca. Por lo tanto, O'Neill, Vonsild, Wallsted y Dornam (2013) afirman que el tipo de criterio de admisión utilizado (basado en notas o no) no influyó en la diversidad social de los estudiantes de medicina admitidos.

Parate, Pande y Lokari (2016) también se cuestionan sobre el proceso de admisión a la carrera de Medicina y mencionan que la selección apropiada de estudiantes son un prerrequisito fundamental si las escuelas de medicina deben producir médicos competentes y humanistas. "La pregunta del millón es ¿qué criterios, si se aplican, pueden seleccionar a los mejores candidatos como futuros médicos?" (p.98). Ante esto, los autores procuraron determinar la eficacia de la prueba de ingreso, además de encontrar la correlación entre las notas de las materias de ciencias en la educación secundaria, nota de admisión y en las notas de pruebas finales de primer año de carrera. También se estudió la diferencia de género en el rendimiento del examen de admisión aplicado Asso-CET (*Association of Managements of Unaided Private Medical and Dental Colleges, Maharashtra -AMUPMDC*).

Parate, Pande y Lokari (2016) indican que en los resultados de su investigación no se observa una correlación significativa entre las notas de las materias de ciencias en la educación secundaria y la nota de admisión, así como la nota de admisión y las notas de pruebas finales de primer año de carrera para los años 2009, 2010 y 2011. En la población de ingreso en el año 2012, se observó una correlación significativa entre las notas de las materias de ciencias en la educación secundaria y la nota de admisión, así como la nota de admisión y las notas de pruebas finales, pero esta correlación fue negativa y no corresponde con los hallazgos internacionales. Sin embargo, sí se observó una correlación estadísticamente significativa positiva entre las notas de las materias de ciencias en la educación secundaria y las notas de pruebas finales de primer año de carrera en los cuatro años. Se encontró que las mujeres obtenían mejores notas en secundaria, sin embargo, no se producía cambios estadísticamente significativos en las pruebas universitarias. Los autores al igual que O'Neill, Vonsild, Wallsted y Dornam (2013) hacen un llamado a la revisión de los procesos de admisión a las ciencias médicas.

Al igual que Parate, Pande y Lokari (2016), Adam et al. (2015) se enfocaron en la predicción del rendimiento académico y estudiaron los atributos y cualidades, solas o en combinación, que predicen resultados en notas y desenvolvimiento clínico de la Medicina. Para ello, tomaron un cohorte y dieron un seguimiento longitudinal por cinco años. Los componentes de admisión fueron: el promedio de notas, prácticas profesionales, evaluaciones OSCE y observaciones estructuradas de tutores.

Los resultados presentados en el artículo corresponden a lo que reportaron 140 estudiantes de tercero a quinto año que ingresaron a Hull York Medical School durante el año 2007. Este estudio se encontró tres aspectos interesantes: primero que los estudiantes con mayores puntajes en las pruebas OSCE, pruebas escritas y prácticas clínicas ingresaron a la carrera con menos de 21 años; segundo que las mujeres superaron a los hombres en el rendimiento de las pruebas escritas y las evaluaciones OSCE; y tercero que los ciudadanos de Inglaterra obtuvieron mejores resultados en las pruebas escritas y en las prácticas clínicas que los estudiantes extranjeros.

Sumado a lo anterior, se destaca que el promedio de notas de educación secundaria resultó ser el único valor predictivo del componente de admisión en las pruebas escritas y además existió una correlación negativa con la cantidad de penalizaciones por errores en prácticas clínicas (Adam et al, 2015). Además, estos autores agregan que algunos rasgos no cognitivos, como confianza y empatía predicen el rendimiento en prácticas clínicas para la muestra seleccionada. El rendimiento académico de los primeros dos años de la carrera resultó buen predictor para graduación de honor y deserción. El estudio también reportó resultados que no coinciden con investigaciones anteriores como lo es alto nivel de impulsividad y bajos niveles de auto disciplina predecían rendimiento mayor en prácticas clínicas. Esta investigación confirma lo encontrado por Parate, Pande y Lokari (2016), estableciendo que las notas obtenidas previo a admisión en la universidad son el criterio más válido para predecir rendimiento académico en escuelas de Medicina.

Saliendo de Europa, en Pakistán se ha realizado investigación sobre la validez predictiva de la prueba de admisión en el rendimiento académico de los estudiantes. Ali y Ali (2013) indican que el propósito de este estudio fue examinar la validez predictiva de la prueba de ingreso realizada por la Agencia de Pruebas y Evaluación Educativa para la admisión a todos los colegios médicos de la provincia de Khyber Pakhtunkhwa (KP) de Pakistán. La metodología de este estudio utilizó un seguimiento del rendimiento de 2944 estudiantes (hombres = 1975, mujeres = 968) que asistieron, 4 institutos médicos de KP desde el nivel de ingreso hasta la graduación, inscritos en los cursos lectivos de 2000-2005. Los autores indican que se analizaron los predictores: notas de educación secundaria, examen de admisión y nota de Mérito (la combinación entre el examen de admisión y las notas de secundaria). Además, se realizó análisis regresivo para evaluar la validez de los predictores anteriormente mencionados.

Ali y Ali (2013) demuestran que la nota de mérito y la prueba de admisión es un criterio predictivo válido, excepto para cuarto año. Mientras que las notas de educación secundaria muestran predicción significativa con todas las notas de las pruebas profesionales de los cinco años. Estos autores concluyen que todos los factores de

predicción (notas de educación secundaria, prueba de admisión, y nota de mérito) están relacionadas a las notas obtenidas de los estudiantes durante sus años de estudio. Sin embargo, de los tres, el que mostró una relación más cercana al desempeño académico son las notas de educación secundaria. Este estudio muestra la variabilidad de los resultados según la muestra y prueba de admisión específica, ya que como se puede observar concuerda con Parate, Pande y Lokari (2016) en que las notas de secundaria son el mejor predictor para el rendimiento académico universitario en Ciencias Médicas, sin embargo, contradice esta misma investigación al reconocer también el examen de admisión como criterio predictor válido.

Wadee y Cliff (2016) también realizan un estudio sobre validez predictiva donde ellos expresan que las escuelas de medicina de Sudáfrica han reconocido la necesidad de transformación y consideración de factores académicos y no académicos en el proceso de selección de estudiantes. Estos autores tienen claro que los criterios de selección deben ser confiables y válidos para garantizar el rendimiento académico exitoso a nivel universitario dentro de un programa de escuela de medicina. Las escuelas de medicina de Sudáfrica han adoptado una estrategia adyacente mediante el uso de las pruebas de ubicación de Ciencias de la Salud (HSPT) desarrolladas por el Proyecto de Investigación de Admisiones Alternativas (AARP).

Wadee y Cliff (2016) indican que, aunque los cambios en las políticas de selección comenzaron antes de 1994 y la admisión de estudiantes de medicina en Sudáfrica mostró un progreso con respecto al cambio del perfil demográfico (que demostró una mejor representación de los grupos más desfavorecidos en 1999 en comparación con 1994), la representación equitativa sigue siendo un desafío que debe abordarse.

Los autores mencionan que las HSPT constan de cuatro pruebas, que incluyen pruebas del lenguaje aplicado a un contexto académico, logros matemáticos, comprensión matemática, y razonamiento científico. Ante este contexto Wadee y Cliff (2016) proponen investigar el grado de asociación entre los puntajes en los componentes de

prueba específicas de la HSPT y los puntajes en los exámenes de mitad de año y finales, de las materias de Física, Química, Biología, Fundamentos Médicos and Ciencias Clínicas, Sociología y Psicología, del curso lectivo para los estudiantes en su primer año de estudio hacia un título en Medicina. En este estudio se encontró que los estudiantes de colegios privados obtuvieron mejores resultados en las sub pruebas de HSPT, excepto la de razonamiento científico, Además no se encontraron diferencias entre las escuelas públicas y privadas en los primeros exámenes del año. Sin embargo, en los exámenes finales, los estudiantes de escuelas privadas lograron calificaciones más altas que los estudiantes admitidos en las escuelas públicas en las asignaturas de Fundamentos Médicos and Ciencias Clínicas y Psicología ($p < 0.05$).

Wadee y Cliff (2016) entre otros resultados encuentran que la prueba HSPT en su conjunto mostró ser el predictor más importante de las notas de mitad de año y las pruebas finales de la mayoría de los sujetos. Para los exámenes inciales, la HSPT explicó 32%, 20% y 32% de la varianza en las calificaciones obtenidas por los estudiantes de medicina en Biología, Química y Psicología, respectivamente ($p < 0{,}0001$). Finalmente, Wadee y Cliff (2016) al igual que Parate, Pande y Lokari (2016) y Adam et al (2015) argumentan sobre el uso de pruebas como los HSPT como una herramienta adicional en la selección de estudiantes de medicina estos autores creen que los hallazgos de este estudio enfatizan el valor complementario de la HSPT y pruebas similares en la comprensión de que la preparación académica de los estudiantes de Medicina no es necesariamente visible sobre la base de resultados de rendimiento académico preuniversitario.

La mayoría de los programas de Doctorado en Medicina reciben más solicitudes de admisión de las que pueden aceptar cada año, lo que exige un proceso de admisión selectivo (Hall, Connell y Cook, 2017). Estas autoras y autor indican que los criterios de selección típicos incluyen puntajes de prueba estandarizadas, promedio de calificaciones de pregrado, cartas de recomendación, un currículum y / o declaración personal que destaque publicaciones o experiencia profesional relevante entrevistas con el equipo encargado del programa de posgrado. Las decisiones de admisión a

menudo se basan en la suposición de que estos componentes de la aplicación se correlacionan con el éxito en producción académica para el posgrado, pero estos supuestos no se han probado rigurosamente.

Por lo tanto, este grupo de investigadores, se enfocaron en determinar si los componentes de la admisión son criterios predictores de la productividad académica medida por las publicaciones de los estudiantes como primer autor y el tiempo hasta la finalización de estudios. Los datos estudiados comprendieron para los estudiantes de posgrado que ingresaron al programa de doctorado biomédico de primer año en la Universidad de Carolina del Norte en Chapel Hill desde 2008 a 2010. Los resultados más sobresalientes de esta investigación indican que no hay correlación, entre las calificaciones en las pruebas, la cantidad de experiencia de investigación previa o las calificaciones de las entrevistas de admisión, con alta o baja productividad entre los solicitantes que fueron admitidos. Hall, Connell y Cook, (2017) indican que las calificaciones obtenidas en las cartas de recomendación fueron significativamente más fuertes para los estudiantes que publicaron varios trabajos como primer autor que para aquellos que no publicaron artículos en el mismo período de tiempo. Los autores coinciden en que la prueba estandarizada más comúnmente utilizada, Graduate Record Examination, es una herramienta predictiva particularmente ineficaz y las evaluaciones cualitativas de quienes llenan las cartas de recomendación anteriores son más propensas a identificar a los estudiantes que tendrán éxito en la investigación de postgrado biomédica.

Debido a lo anterior, Hall, Connell y Cook, (2017) concluyen que los comités de admisión deben evitar la dependencia excesiva en cualquier componente individual de la aplicación y restar importancia a las métricas que son mínimamente predictivas de la productividad de los estudiantes. Además, recomiendan el seguimiento continuo de las prácticas de admisión ya que, para ellos, el objetivo principal de la formación de doctorado biomédico es desarrollar habilidades científicas de alto nivel a través del proceso de generación de un cuerpo significativo de investigaciones originales publicadas. Sin embargo, reconocen que el simple conteo de publicaciones es una medida imperfecta

de la productividad. Las frecuencias de publicación varían según las disciplinas y los laboratorios, y sin duda un artículo de gran influencia puede representar un volumen sustancial de trabajo, mientras que tres o más trabajos pequeños pueden considerarse solo productividad promedio en algunos casos. En el artículo de Hall, Connell y Cook, (2017) se muestra que diferencia de las suposiciones ampliamente aceptadas por los profesores y administradores de admisiones sobre el poder predictivo de las notas y los puntajes generales de los exámenes de admisión, no se encontró ninguna correlación entre estas métricas y las publicaciones estudiantiles o la finalización del doctorado.

El rendimiento académico en centros de educación superior también interesa a Schripsema, Van Trigt, Borleffs y Cohen (2014) quienes se preguntan: ¿Es el rendimiento de la escuela de medicina diferente entre los estudiantes que fueron admitidos a través de diferentes procesos de admisión? ¿Los estudiantes que fueron aceptados en el proceso de selección multifacético superan a los estudiantes que fueron rechazados en este proceso? ¿Los estudiantes admitidos por medio de la selección multifacética superan a los estudiantes que no participaron en esta modalidad?

La investigación se realizó en la Universidad de Groningen en los Países Bajos. El plan de estudios consiste en un bachillerato preclínico de 3 años y un programa de maestría clínica de 3 años. Schripsema, Van Trigt, Borleffs y Cohen (2014) indican que en los Países Bajos existe una política nacional de admisión a la escuela de medicina según la cual los solicitantes pueden ser admitidos a través de una de tres modalidades. En la primera modalidad, los estudiantes con notas preuniversitarias ≥ 8 (en una escala que varía de 1 = pobre a 10 = excelente) obtienen admisión en la escuela de medicina de su elección sin evaluación adicional. Este promedio de calificaciones se calcula al promediar la nota promedio de cada solicitante en los exámenes preuniversitarios y la calificación promedio en los exámenes finales nacionales. Aproximadamente solo 4% de los estudiantes obtienen esta nota o superior.

La segunda modalidad de ingreso consiste en una evaluación multifacética a cargo de cada escuela de Medicina que se basan generalmente en evaluaciones de

conocimiento en ciencias y variables conductuales de interés. La tercera modalidad es una lotería de admisión en donde pueden concursar quienes no fueron admitidos en la evaluación multifacética y quienes no concursaron en esa modalidad. Para esto se establecen cuatro categorías de notas preuniversitarias: 7.5–7.9; 7.0–7.4; 6.5–6.9, y < 6.5. La proporción de solicitantes admitidos por categoría es 9: 6: 4: 3 (Schripsema, van Trigt, Borleffs y Cohen, 2014).

Schripsema, Van Trigt, Borleffs y Cohen (2014) estudian datos de 1055 estudiantes admitidos al programa de bachillerato holandes de Medicina en la Universidad de Groningen en 2009, 2010 y 2011 (69% mujeres, edad media al inicio del primer año: 18,6 años, promedio de notas preuniversitarias: 7,3). El grupo investigador definió cuatro grupos de estudiantes: estudiantes que ingresaron en base a notas preuniversitarias de ≥ 8 de 10 (n = 143; 71% mujeres; edad promedio al inicio del primer año 18.0 años; promedio de notas preuniversitarias 8.2); estudiantes que fueron aceptados en el proceso de selección multifacético (n = 295; 74% mujeres; edad promedio al inicio del primer año 18.5 años; promedio de notas preuniversitarias 7.1); estudiantes admitidos en lotería que habían sido rechazados anteriormente en el proceso de selección multifacética (n = 315; 69% mujeres; edad promedio al inicio del primer año 18.5 años; promedio de notas preuniversitarias 7.1), y estudiantes admitidos en lotería que no habían participado en el proceso de selección (n = 302, 63% mujeres, edad promedio al inicio del primer año 19.1 años; promedio de notas preuniversitarias 7.0).

A partir de las anteriores modalidades de admisión se buscó predecir las notas en pruebas escritas, progreso académico (creditaje aprobado), notas en profesionalismo y deserción del programa. Schripsema, Van Trigt, Borleffs y Cohen (2014) reportan que las notas en las pruebas escritas variaron según los grupos $F(3,1025) = 63.20$; $p < 0.001$. Las pruebas de comparación múltiple post hoc de Bonferroni mostraron que el grupo de notas preuniversitaria superior tenía puntaje mayor en las pruebas que to-

dos los demás grupos. El grupo que fue aceptado en el proceso de selección multifacética logró puntajes más altos que el grupo admitido en lotería que no había participado en este proceso (DM: 0.30, SE: 0.08; p <0.01).

Los números de créditos de curso obtenidos en el primer año difirieron entre los grupos F (3,1025) = 17.50; p <0.001. El grupo de notas preuniversitaria superior obtuvo más créditos en el primer año que todos los demás grupos (DM: 3.9-.5, SE: 1.20-1.22; p <0.01). El grupo admitido por lotería que no participó en el proceso de selección multifacética obtuvo menos créditos de curso en el primer año que el grupo que había sido aceptado en este proceso y el grupo admitido en lotería que había sido rechazado en el proceso de selección multifacética (Schripsema, Van Trigt, Borleffs y Cohen, 2014).

El grupo de notas preuniversitaria superiores recibió la puntuación más alta posible con mayor frecuencia, en el profesionalismo, seguido por el grupo que había sido aceptado en el proceso de selección multifacético. Las diferencias entre este último grupo y ambos grupos admitidos en lotería fueron significativas, como lo fue la diferencia entre el grupo de notas preuniversitaria superiores y el grupo admitido por lotería que no había participado en el proceso de selección multifacética. Schripsema, Van Trigt, Borleffs y Cohen (2014) además indican que no se encontraron diferencias estadísticamente significativas en la tasa de deserción según la modalidad de admisión.

Este grupo investigador encontró que los diferentes procesos de admisión se relacionan con las diferencias en el rendimiento del estudio. Los estudiantes que fueron admitidos en base a notas preuniversitarias superiores obtuvieron mejores resultados en todas las medidas. Estudiantes que fueron aceptados en el proceso de selección multifacética y sus pares rechazados difirieron significativamente en puntajes de profesionalismo, pero no en las otras medidas de resultado. Estudiantes admitidos en lotería que no participaron en el proceso de selección multifacética mostró el nivel más bajo de rendimiento. El hallazgo de que los estudiantes admitidos en base a notas preuniversitarias superiores se desempeñaron mejor es congruente con resultados

demostrados anteriormente en este escrito (Schripsema, Van Trigt, Borleffs y Cohen, 2014).

En la Universidad de Cambridge Emery, Bell y Vidal (2011) analizan la predicción de la prueba BMAT en el rendimiento académico de los estudiantes en primer año de la carrera de Medicina, analizando además los efectos de predicción según tipo de escuela y género. Es importante resaltar que el examen de ingreso biomédico (BMAT) forma parte del proceso de admisión médica de pregrado en la Universidad de Cambridge. Los autores hacen énfasis en la importancia de investigar imparcialidad de las pruebas de admisión, pues en muchas pruebas estandarizadas se pueden estar afectando poblaciones específicas. Este grupo de investigación se planteó analizar las relaciones entre las variables de antecedentes de los solicitantes y los puntajes de BMAT, si se les ofreció un lugar o si se rechazaron y, para los admitidos, el rendimiento en los exámenes del primer año. Para lograr lo anterior se emplearon modelos de regresión multinivel con datos de tres cohortes de solicitantes combinados. Las tasas de admisión para diferentes grupos se investigaron con y sin controlar el rendimiento de BMAT. La equidad del BMAT se investigó determinando, para los admitidos, si las puntuaciones predijeron el rendimiento del examen de forma equitativa.

A pesar de algunas diferencias en el rendimiento de BMAT de los solicitantes por tipo de escuela y sexo, las puntuaciones de BMAT predijeron que el examen medio marca de manera equitativa todas las variables de fondo consideradas. Sin embargo, la probabilidad de lograr resultados superiores en los primeros exámenes fue ligeramente inferior a la prevista para aquellos admitidos de escuelas que ingresan relativamente pocos solicitantes, esto no es de extrañar producto de la poca cantidad de datos que se tienen disminuyendo la heterogeneidad de la muestra (Emery, Bell y Vidal, 2011).

Emery, Bell y Vidal (2011) hacen énfasis en que la prueba constituye solo una parte de un sistema de admisión compuesto con otros factores, como el rendimiento de la entrevista y logros importantes, por lo tanto, estos autores concluyen que los

resultados respaldan la equidad del BMAT. En esta investigación se muestra el interés por parte del equipo de investigación al igual que O'Neill, Vonsild, Wallsted y Dornam (2013) y Schripsema, van Trigt, Borleffs y Cohen (2014) en brindar acceso a educación superior a toda la diversidad de clases sociales.

En Latinoamérica también se ha demostrado interés por estudiar el rendimiento académico de los estudiantes de Medicina. En la Universidad Autónoma de México Ponce de León, Ortiz, Bonilla y Berlanga (2006) describen el procedimiento para evaluar de manera integral la competencia clínica mediante el Examen Profesional de la Carrera de Médico Cirujano y presentan los resultados obtenidos por los sustentantes. Para esto se realizó un estudio descriptivo y transversal donde se analizaron los resultados de los exámenes profesionales en su fase teórica, y en la fase práctica por áreas de la competencia clínica para los años 2001 y 2002.

La fase práctica se realizó a través de un examen escrito, con base en casos clínicos, que abarcan 84 problemas comunes en el ejercicio profesional del médico general, correspondientes a las cuatro áreas troncales de la Medicina; Medicina interna, Pediatría, Ginecoobstetricia y Urgencias medicoquirúrgicas. La fase práctica se realizó para certificar los conocimientos, actitudes, habilidades y destrezas clínicas utilizados por el alumno, en el manejo de un paciente con un problema clínico real. Esta última fase se desarrolló en 38 hospitales y se conforman jurados con dos o tres profesores del área clínica y uno del área básica para evaluar mediante una guía los elementos básicos de la competencia clínica. Ponce de León, Ortiz, Bonilla y Berlanga (2006) mencionan que estos elementos básicos son relación médico paciente, habilidades de comunicación para la obtención de información en el interrogatorio y en la exploración física, capacidad del alumno para integrar la información obtenida, habilidades de razonamiento clínico, diagnósticos, terapias y pronóstico.

Este grupo investigativo reportan que el número de alumnos que se presentaron a la Fase teórica fue de 1 185 y 1 159 y a la Fase práctica de 927 y 610 respectivamente. Se tuvo una confiabilidad de 0.987 y 0.94 para el examen teórico, sin embargo, no se identificó para la fase práctica. El rendimiento promedio en conjunto para ambos exámenes fue de 231 aciertos (56%), resultado que se ubica dentro del rango esperado para reactivos con dificultad media (.40 a .60). Ponce de León, Ortiz, Bonilla y Berlanga (2006) reportan que la desviación estándar de 37 aciertos (8.8%), muestra una gran dispersión en el rendimiento académico probablemente debida a lo heterogéneo de la población. El rendimiento de los alumnos regulares (quienes no se atrasaron en el plan de estudios), en las dos fases, fue mayor que el de los que tardaron más años en cubrir el plan de estudios. Además, se encontró que pediatría fue el área con mayor rendimiento en promedio, en contraposición con las áreas de diabetes e hipertensión.

De los hallazgos encontrados por Ponce de León, Ortiz, Bonilla y Berlanga (2006) se puede indicar que este tipo de evaluación demuestra que son un elemento más que permite retroalimentar a diversas instancias académicas para incorporar acciones correctivas. En cuanto al análisis realizado al Examen Profesional Teórico-Práctico, se evidencia que los instrumentos utilizados permiten evaluar satisfactoriamente lo que un egresado de la carrera de medicina debe saber, saber hacer y hacer, es decir, su competencia de egreso.

Kenny, McInnes, y Singh (2013) muestran los resultados obtenidos en una investigación con respecto a cuál de la información disponible para la selección de residentes está asociada con el desempeño del médico-residente. Estos resultados ponen de relieve una clara diferencia entre las estrategias de selección basadas principalmente en el examen y las estrategias de selección más subjetivas para las que no existe un resultado estandarizado o numérico (por ejemplo, cartas de referencia, cartas de los decanos, entrevistas). En general, las estrategias de selección basadas en los exámenes tienen una relación más fuerte con resultados no basados en el examen que las estrategias de selección subjetiva. Asimismo, se encontró que las notas

de las escuelas de medicina estaban asociadas con el desempeño clínico en las residencias, mientras que los registros de desempeño en la escuela médica tenían menor correlación con la Evaluación del Residente en Entrenamiento. También las cartas de referencia y la experiencia de investigación tenían las asociaciones más débiles con el desempeño de los residentes. El estudio demuestra que las calificaciones de las escuelas de medicina y el puntaje en las pruebas estandarizadas tienen las asociaciones positivas más fuertes disponibles actualmente para la selección de residentes.

Por otro lado, Duré, Dursi, Raffoul y Caffarena (2014) menciona que los concursos de ingreso se han organizado en general contemplando una prueba escrita, el promedio como uno de sus ponderadores, entrevista personal por parte de los jefes de servicio, antecedentes académicos, idioma, los cuales se combinan en distintas proporciones. Este estudio muestra que el pertenecer a regiones lejos del centro de estudio dificulta el traslado para los estudiantes, por esto se resalta inequidad en el acceso a estas plazas de formación. De la misma forma, en muchos de los centros la generación del examen y la entrevista las realizó un solo actor, dando lugar a una pérdida de transparencia en el proceso.

Para concluir con este apartado se presenta los resultados de la revisión bibliográfica con respecto a los métodos de selección de estudiantes en medicina realizada por Patterson, et al. (2016), estos autores tomaron como referencia para el estudio la eficacia (fiabilidad y validez), el procedimiento, la aceptabilidad y la rentabilidad de cada evaluación.

En primera instancia se menciona el **examen de aptitud**, el cual presenta un conflicto en su efectividad, ya que algunas pruebas sugieren que los estudiantes seleccionados por medio de este método suelen ser más capaces y estar más motivados para el estudio de medicina, no obstante, la utilidad de este examen depende del tipo de prueba aplicada y algunos estudios sugieren que algunas de estas pruebas

están relacionadas con los antecedentes del candidato, o que favorecen a ciertos grupos, pero nada está definido (Patterson, et al. 2016).

Otro elemento encontrado en la revisión de Patterson, et al. (2016) es el **record académico**, ya que se evidencia la validez predictiva de los expedientes académicos en la selección del estudiante de medicina, pues los estudiantes con un mejor record tienen más probabilidad de tener éxito en la escuela de medicina. Según esta exploración el logro educativo previo forma la columna vertebral académica de la selección y la progresión a través de la escuela de medicina y más allá. En este sentido, las pruebas internacionales sugieren que los candidatos admitidos sobre la base de sus registros académicos tenían niveles de abandono más bajos que los que no lo eran. Sin embargo, existe la preocupación de que el poder discriminatorio de los records puede estar disminuyendo a medida que un número creciente de solicitantes tienen buenas calificaciones.

Las **declaraciones personales** también han sido importantes en la selección de estudiantes, a pesar de que su efectividad es variada, se sugiere que podrían tener valor para hacer a los solicitantes conscientes de las características del título médico que están solicitando, lo que puede ayudarles a tomar una decisión más informada. Los candidatos suelen usar declaraciones personales para presentarse de una manera que consideran atractiva para los comités de admisión, pero que no necesariamente son exactas. El contenido de las declaraciones personales también puede parcializar el juicio de los que toman las decisiones de selección (Patterson, et al. 2016).

Patterson, et al. (2016) también encontró que las **referencias** no tienen valor predictivo en el desempeño de los médicos, ya que la información suele ser cuestionable y no aporta información que hagan sobresalir a los solicitantes, además puede parcializar a los comités de admisión. Lo cual discrepa con los resultados de la encuesta elaborada por Hillebrand, Leinum, Desai, Pettit y Fuller (2015), donde encontraron que, para la selección de candidatos los factores más importantes han sido la revisión del currículo y las cartas de recomendación.

Las **pruebas de juicio situacional** (SJT), se ha contemplado como un método valido y confiable para la selección de estudiantes; aunque la forma de aplicación puede afectar la validez (las de video son más efectivas que las escritas). Los SJT son complejas de desarrollar y hay una amplia gama de opciones disponibles en relación con formatos de artículos, instrucciones y puntuación, pero cuando estas opciones se calibran apropiadamente, se evidencia la fuerza en la selección de estudiantes de medicina para evaluar los atributos no académicos (Patterson, et al. 2016).

Por otro lado, se halló que **la evaluación de personalidad e inteligencia emocional** mantiene asociación entre los rasgos de personalidad y el rendimiento en la educación médica y la formación. También se ha demostrado que proporciona una validez incremental sobre los métodos cognitivos en un proceso de selección de facultades de medicina. Algunos rasgos dificultan la eficiencia de los médicos, por ejemplo, los rasgos de personalidad "disfuncionales" en estudiantes de medicina (incluyendo paranoicos, evasivos, pasivos agresivos, antisociales, rasgos narcisistas y no cooperativos), estos han sido reportados como asociados con calificaciones académicas más bajas. Sin embargo, no se ha encontrado correlación positiva entre la personalidad y las habilidades médicas. Con este aspecto sobresale la preocupación de que se puedan “falsificar” los test de personalidad (Patterson, et al. 2016).

Otro componente esencial propuesto por Patterson, et al. (2016) son las **entrevistas y mini entrevistas múltiples** (MMI), ya que la entrevista tradicional no suele ser un método fuerte y carece de valor predictivo; mientras que la entrevista múltiple suele ser más consistente, sin embargo, su validez sigue siendo exploratoria y en gran parte inconclusa. Se resalta que las entrevistas cara a cara estandarizadas pueden no ser comparables con las estaciones planteadas por las MMI basadas en escenarios que usan actores de rol estandarizados y distintas estaciones, las MMI según los estudios son más apropiadas para evaluar una gama de competencias y son un proceso justo, preferible a la entrevista tradicional.

Patterson, et al. (2016) alude como último elemento los **centros de selección** (SC), los cuales han sido son altamente aceptados, aunque puede ser un método caro y logísticamente complejo. De acuerdo con Patterson, Zibarras, Kerrin, Lopes y Price (2014) los SC suelen usar pruebas de muestras de trabajo de alta fidelidad ("juegos de rol") y se ha demostrado que estos exhiben alta validez de criterio y son populares en la educación y evaluación médica. Las simulaciones pueden utilizarse para evaluar la competencia de los médicos, a la vez que proporcionan un contexto real para comprender las complejas necesidades de atención del paciente. La estandarización del desempeño de los simuladores asegura la coherencia entre las experiencias, necesario para hacer comparaciones justas y confiables entre los candidatos. Sin embargo, es relevante mencionar que en los SC, los evaluadores están obligados a observar, registrar y evaluar el desempeño de los candidatos usando escalas de calificación estandarizadas, por tanto, los estudios demuestran que las habilidades de estas personas son vitales para el éxito de cualquier proceso de SC.

En el estudio de Patterson, et al. (2016) se evidencia que los expedientes académicos, las MMIs, las pruebas de aptitud, las pruebas de juicio situacional y los centros de selección son los métodos más eficaces y que son generalmente más justos que las entrevistas tradicionales, las referencias y las declaraciones personales.

VI. CONCLUSIONES

En América Latina, nueve países cuentan con un examen de ingreso centralizado a nivel nacional para incorporarse a posgrados en las especialidades médicas (Bolivia, Honduras, Costa Rica, México, República Dominicana, Chile, Paraguay, Perú y Uruguay), de igual modo que sucede en otros países del mundo (Duré, Dursi, Raffoul y Caffarena, 2014). Adicionalmente este tipo de pruebas puede eventualmente ser acompañadas por otros instrumentos como portafolios, la evaluación clínica objetiva estructurada, las rúbricas, el ejercicio de examen clínico reducido (Arenis y Pinilla, 2016), exámenes orales, resolución de casos clínicos, uso de pacientes virtuales, entre otros (Baradaran, Salek, Nikasa, 2015).

La combinación de pruebas según Baradaran, Salek, Nikasa (2015) es una herramienta confiable para evaluar las habilidades de razonamiento clínico en estudiantes de medicina, aunque estos resultados pueden no ser generalizables para toda la población de estudiantes. Para estos autores el razonamiento clínico es una acción en la que la información relativa a un problema clínico se mezcla con el conocimiento y la experiencia previa de los médicos para solucionar un determinado problema. Este trabajo propicia que se piense en situaciones clínicas complejas, las cuales conducen a la toma de decisiones.

En concordancia con lo descrito, Sánchez-Mendiola y Delgado-Maldonado (2017) proponen que uno de los principios más importantes en evaluación educativa, ya que la evaluación con exámenes de alto impacto tiene factores positivos y negativos, por lo que los argumentos académicos ceden a los aspectos afectivos y de intereses gremiales, alimentados por una ubicua falta de conocimiento de la metodología moderna de elaboración de exámenes y de los conceptos actuales de validez en evaluación.

Entre los efectos negativos, está: el enseñar para los exámenes, los cursos de preparación para exámenes, los efectos en currículo formal y oculto y las inferencias inapropiadas de los resultados de los exámenes. Mientras que los potenciales efectos positivos son: la motivación para estudiar, la estandarización de la evaluación, la mejora de la calidad educativa, la unificación de criterios y las consecuencias positivas no intencionales que según Gregory Cizek (citado en Sánchez-Mendiola, y Delgado-Maldonado, 2017) son: el desarrollo profesional, la acomodación de conocimiento sobre evaluación, la colección de la información, el uso de la información, las opciones educativas, el sistema de rendimiento de cuentas, la familiaridad de los docentes con sus disciplinas y la calidad de los exámenes.

La mayoría de los autores estudiados por Rodríguez (2008) concluyen que se requieren varios formatos de evaluación para establecer con certeza el grado de aprendizaje de los estudiantes y la competencia clínica global de los egresados. En esta línea, el dominio de los conocimientos de ciencias básicas puede ser mejor evaluado con exámenes de opción múltiple, exámenes orales y ensayos, pero se requieren procedimientos más sofisticados para evaluar las diferentes facetas de la competencia clínica; entre ellos los formatos que utilizan pacientes estandarizados, reales, y el denominado examen Clínico Objetivo y Estructurado. Lo que concuerda en gran medida con los hallazgos de la literatura estudiada para el presente estado de la cuestión.

De las publicaciones analizadas, las evaluaciones que han sido empleadas con mayor frecuencia a estudiantes de medicina son la Evaluación Clínica Objetiva Estructurada, la Mini-entrevista múltiple (MMI) y el Mini-CEX. Aunado a esto se sobresalen diferentes métodos de selección de estudiantes en el área de medicina, como las pruebas de ubicación, evaluaciones de conocimiento en ciencias, exámenes escritos de ingreso estandarizado, exámenes prácticos, ensayos referentes a la materia, entrevistas personales, evaluación de personalidad e inteligencia emocional, cartas de recomendación, el record académico, la experiencia laboral y de voluntariado, la publicación de artículo y el manejo de idiomas.

Enfatizando en la prueba OSCE la cuál ha sido más estudiada en los artículos encontrados, se destaca que esta es una estrategia evaluativa que se ha desarrollado y expandido en el área de las ciencias médicas. Las investigaciones demuestran que se ha convertido en un instrumento viable para formar profesionales con orientación humanista y completa, enfatizando en las relaciones humanas y trato a los pacientes y no únicamente en el conocimiento técnico de los profesionales. Es importante resaltar la necesidad de desarrollar este tipo de evaluación que corresponda según los objetivos y conocimientos planteados en los programas de estudio.

Los estudios en el presente estado de la cuestión muestran la necesidad de evaluar las pruebas utilizadas para la admisión de estudiantes a posgrados en ciencias médicas y evaluar de igual manera las pruebas implementadas durante el curso lectivo. En esta revisión se encuentran algunas sugerencias de cambios que pueden beneficiar los estudios de posgrado como lo son la realización de pruebas con únicamente tres reactivos por ítem y el uso de retroalimentación grupal escrita y oral hacia los practicantes.

Sin lugar a duda, el conocimiento y habilidades de los estudiantes son de amplio interés para el quehacer académico de las universidades, específicamente en los procesos de admisión. Los estudios de validez predictiva ayudan a comprender las variables que influyen en el rendimiento académico y profesional de sus estudiantes y con ello tomar mejores decisiones e ingreso y rechazo de estudiantes. Sin embargo, son pocos los estudios lo que investigan sobre los efectos de los procesos en grupos minoritarios y esto resulta ser una práctica necesaria e importante para la educación moderna, con el afán de incorporar personas de poblaciones vulnerabilizadas socialmente.

VII. REFERENCIAS

__(2013). Mini-interviews help to recruit students for their values (2013). Nursing Standar. 28 (2), 6.

Adams, J., et al. (2015). Predictors of professional behavior and academic outcomes in a UK medical school: A longitudinal cohort study. *Medical Teacher,* 37, 868-880.

Alarcon. A. (2013). Incorporación del Examen Clínico Objetivo Estructurado (ECOE) en la Carrera de Enfermería. *Rev Educ Cienc Salud. 10* (1): 18-22

Ali, A. & Ali, Z. (2013). Admission policy of medical colleges: Evaluating validity of admission test in Khyber Pakhtunkhwa, Pakistan. *Journal of Research and Reflections in Education, 7*(1), 77-88. Recuperado de http://www.ue.edu.pk/jrre

Allerup, et al. (2007). Use of 360-degree assessment of residents in internal medicine in a Danish setting: a feasibility study. *Medical Teacher,* 29, 166-170. Doi: 10.1080/01421590701299256

Arenis, Y. y Pinilla, A. (2016). Evaluación de estudiantes de posgrado en ciencias de la salud. *Acta Médica Colombiana.* 41(1), 49-57.

Axelson, R. & Kreiter, C. (2009). Rater and occasion impacts on the reliability of pre-admission assessments. *Medical Education*, 43, 1198–1202.

Baños, J., Gomar-Sancho, C., Grau-Junyent, J., Palés-Argullós, J. & Sentí, M. (2015). El mini-CEX como instrumento de evaluación de la competencia clínica. Estudio piloto en estudiantes de medicina. *FEM; 18* (2): 155-160.

Baradaran, M., Salek, F., Nikasa, P. (2015). Different Methods to Assess Clinical Reasoning in Undergraduate Medical Students. *Res Dev Med Educ, 4*(2), 113-114.

Brazil, V., Ratchiffe, L., Zhang, J., & Davin, L. (2012). Mini-CEX as a workplace-based assessment tool for interns in an emergency department - Does cost outweigh value?. *Medical Teacher,* 34, 1017-1023. Doi: 10.3109/0142159X.2012.719653

Chávez, M. & Barrantes, G. (2014). Confiabilidad y validez de las listas de cotejos del Examen Clínico Objetivo Estructurado para el aprendizaje por competencias de Cirugía. *Ciencia y Tecnología, 10*(3) 115-128.

Delgado, L. & Sánchez, M. (2012). Análisis del examen profesional de la Facultad de Medicina de la UNAM: Una experiencia de evaluación objetiva del aprendizaje con la teoría de respuesta al ítem. *Investigación en Educación Médica. 1*(3), 130-139.

Díaz-Plasencia, J., Moreno-Castillo, P., Calmet-Ipince, J., Yan-Quiroz, E., Díaz-Villazón, M., Iglesias-Obando, A., Zegarra-Castillo, K. & Urquiaga-Ríos, K. (2016). Validez concurrente del examen clínico objetivo estructurado con el portafolio electrónico, examen teórico y promedio ponderado en estudiantes de cirugía de la Universidad Privada Antenor Orrego. *FEM, 19*(5), 237-245.

Díaz-Plasencia, J., Sánchez, E., Guzmán-Gavidia, C., Valencia-Mariñas, H., García-Cabrera, J., Yan-Quiroz, E. & Díaz-Villazón, M. (2016). Fiabilidad y validez de un portafolio reflexivo en la evaluación de la práctica clínica de los estudiantes del capítulo de Cirugía Oncológica del curso de Cirugía. *FEM. 19*(4), 175-185.

Dirección Nacional de Capital Humano y Salud Ocupacional Subsecretaría de Políticas, Regulación y Fiscalización Secretaría de Políticas, Regulación e Institutos (2014). Examen Único de Ingreso a Residencias Médicas. Análisis estadístico de Examen Único 2013. Ministerio de la Salud. Presidencia de la Nación.

Donato, A., Pangaro, L., Smith, C., Rencic, J., Diaz, Y., Mensinger, J. & Holmboe, E. (2008). Evaluation of a novel assessment form for observing medical residents:

a randomised controlled trial. *Medical Education.* 42, 1234-1242. Doi:10.1111/j.1365-2923.2008.03230.x

Dowell, J., Lynch, B., Till, H., Kumwenda, B. & Husbands, A. (2012). The multiple mini-interview in the UK context: 3 years of experience at Dundee. *Medical Education*, 34, 297-304. Doi: 10.3109/0142159X.2012.652706

Emery, J., Bell, J., & Vidal-Rodeiro, C. (2011). The biomedical admissions test for medical student selection: issues of fairness and bias. *Medical Teacher*, 33, 62-71. Doi: 10.3109/0142159X.2010.528811

Eva, K. & Macala, C. (2014). Multiple mini-interview test characteristics: 'tis better to ask candidates to recall than to imagine. *Medical Eduacation*, 48, 604-613. doi:10.1111/medu.12402

Eva, K., Reiter, H., Trinh, K., Wasi, P. Wasi, P., Rosenfeld, J. & Norman, G. (2009). Predictive validity of the multiple mini-interview for selecting medical trainees. *Medical Education, 43*, 767–775.

Fernández, G. (2011). Evaluación de las competencias clínicas en una residencia de pediatría con el Mini-CEX (Mini-Clinical Evaluation Exercise). *Archivos argentinos de pediatría, 109*(4), 314-320.

Findyartini, A., et al. (2015). Collaborative progress test (cPT) in three medical schools in Indonesia: The validity, reliability and its use as a curriculum evaluation tool. *Medical Teacher*, 37, 366–373.

Fornells-Vallés, J. (2009). El ABC del Mini-CEX. *Educación Médica, 12*(2), 83-89.

Gafni, N., Moshinsky, A., Eisenberg, O., Zeigler, D. & Ziv, A. (2012). Reliability estimates: behavioural stations and questionnaires in medical school admissions. *Medical Education,* 46, 277-288. Doi:10.1111/j.1365-2923.2011.04155.x

Gamboa-Salcedo, T., García-Durán, R. Martínez-Viniegra, M., Sánchez-Medina, J., Peña-Alonso, Y. & Pacheco-Ríos, A . (2011). Examen Clínico Objetivo Estructurado como instrumento para evaluar la competencia clínica en Pediatría. Estudio piloto*. Bol Med Hosp Infant Mex. 68*(3), 184-192.

Hall, J., O'Connell, A. & Cook, J. (2017). Predictors of student productivity in biomedical graduate school applications. *PLos ONE, 12*(1), 1-14. Doi:10.1371/journal.pone.0169121

Hillebrand, K., Leinum, C., Desai, S., Pettit, N., & Fuller P. (2015). Residency application screening tools: A survey of academic medical centers. *American Society of Health-System Pharmacists, 72*(1), 16-19. Doi: 10.2146/ajhp150093

Iblher, P., Zupanic, M., Karsten, J. & Brauer, K. (2015). May student examiners be reasonable substitute examiners for faculty in an undergraduate OSCE on medical emergencies? *Medical Teacher*, *37*, 374-378.

Illesca. M, Cabezas. M, Romo. M & Díaz, P. (2012). Opinión de estudiantes de enfermería sobre El Examen Clínico Objetivo Estructurado. *Ciencia y enfermería XVIII* (1), 99-109.

Kaliyadan, F., Khan, A., Kuruvilla, J., & Feroze, K. (2014). Validation of a computer based objective structured clinical examination in the assessment of undergraduate dermatology courses. *Indian Journal of Dermatology, Venereology, and Leprology, 80* (2), 134-136. Doi: 10.4103/0378-6323.129386.

Kenny, S., McInnes, M. & Singh, V. (2013). Associations between residency selection strategies and doctor performance: a meta-analysis. *Medical Education, 47*, 790–800.

Knorr, M. & Hissbach, J. (2014). Multiple mini-interviews: same concept, different approaches. Medical Education, *48*, 1157-1175. doi: 10.1111/medu.12535.

Lievens, F. (2013). Adjusting medical school admission: assesing interpersonal skills using situation judgement tests. *Medical Education, 47*, 182-189. Doi:10.1111/medu.12089.

López, L. (2017). Evaluación clínica objetiva y estructurada (ECOE) en la maestría de Enfermería Ginecobstétrica y Perinatal: una sistematización de la experiencia. *Enfermería Actual de Costa Rica. Universidad de Costa Rica.* (33), 1-17.

Malhotra, S., Hatala, R., & Courneya, C. (2008). Internal medicine residents' perceptions of the Mini-Clinical evaluation exercise. Medical Teacher, 30, 414-419.

Martínez-González, A., Lifshitz-Guinzberg, A., Trejo-Mejía, J., Torruco-García, U., Fortoul-van der Goes, T., Flores-Hernández, F., Peña-Balderas, J., & Sánchez-Mendiola, M. (2017). Evaluación diagnóstica y formativa de competencias en estudiantes de medicina a su ingreso al internado médico de pregrado. *Gaceta Médica de México, 153*, 6-15.

Monroe, K. (2016). The relationship between assessment methods and self-directed learning readiness in medical education. *International Journal of Medical Education, 7*, 75-80. Doi:10.5116/ijme.56bd.b282.

O'Neill, L., Vonsild, M., Wallstedt, B., & Dornan, T. (2013). Admission criteria and diversity in medical school. In Medical Education, 47, 557–561.

Obsnorne, A., Hawkins, S., Pournaras, D., Chandratilake, M. & Welbourn, R. (2014). An evaluation of operative self-assessment by UK postgraduate trainees. Medical Teacher, 36, 32-37. Doi:10.3109/014259x.2013836268.

Patterson, F., Knight, A., Dowell, J., Nicholson, S., Cousans, F., & Cleland, J. (2016). How effective are selection methods in medical education? A systematic review. *Medical Education, 50*, 36-60.

Patterson, F., Zibarras, l., Kerrin, Lopes & Price. (2014). Development of competency models for assessors and simulators in high-stakes selection processes. *Medical Teacher. 36,* 1082–1085.

Ponce de León-Castañeda, M., Ortiz-Montalvo, A., Bonilla-González, I. & Berlanga-Balderas, F. (2006). Evaluación profesional de la competencia clínica a través del Examen Profesional. *Archivos en Medicina Familiar, 8*(2). 119-129.

Pugh, D., Bhanji, F., Cole, G., Dupre, J., Hatala, R., Humphrey-Murto, S., Touchie, C. & Wood, T. (2016). Do OSCE progress test scores predict performance in a national high-stakes examination?. *Medical Education, 20*, 351-358. Doi: 10.1111/medu.12942.

Rahim, A., & Yusoff, M. (2016). Validity evidence of a multiple mini interview for selection of medical students: Universiti Sains Malaysia experience. *Education in Medicine Journal, 8*(2), 49-63. Doi: 10.5959/eimj.v8i2.437

Rivera, J., Flores, F., Alpuche, A., & Martínez, A. (2016). Evaluación de reactivos de opción múltiple en medicina. Evidencia de validez de un instrumento. *Investigación en Educación Médica, 6*(21), 8-15. Recuperado de https:///dx.doi.org/10.1016/j.riem.2016.04.005.

Rodríguez, R. (2008). La evaluación del conocimiento en medicina. *Revista de la Educación Superior, XXXVII*(147), 31-42.

Rosenfeld, J., Reiter, H., Trinh, K. & Eva, K. (2008). A Cost Efficiency Comparison Between The Multiple Mini-Interview and Traditional Admissions Interviews. *Advances in Health Sciences Education*, *13*, 43-58.

Salazar, O., Vélez, C. & Zuleta. J. (2015). Evaluación de conocimientos con exámenes de selección múltiple: ¿tres o cuatro opciones de respuesta? Experiencia con el examen de admisión a posgrados médico-quirúrgicos en la Universidad de Antioquia. *IATREIA. 28*(3), 300-311.

Sánchez-Mendiola, M. & Delgado-Maldonado, L. (2017). Exámenes de alto impacto: implicaciones educativas. *Investigación en Educación Médica, 6*(21), 52-62.

Schripsema, N., Van Trigt, A., Borleffs, J., & Cohen-Schotanus, J. (2014). Selection and study performance: comparing three admission processes within one medical school. *Medical Education, 48*, 1201-1210. Doi: 10.1111/medu.12537

Schweitzer, R., Khawaja, N., Strodl, E., Lodge, J., Coyne, J., & King, R. (2014). Towards a model for student selection in clinical psychology. *Clinical Psychologist, 18*, 125-132. Doi:10.1111/cp.12025.

Sharma, N., Cui, Y., Leighton J., & White, J. (2012). Team-based assessment of medical students in a clinical clerkship is feasible and acceptable. *Medical Teacher, 34*, 555-561. Doi: 10.3109/0142159X.2012.669083

Shih-Chieh Liao, Tzuen-Ren Hsiue, Chyi-Her Lin & A-Min Huang (2014). Multiple mini-interviews combined with group interviews in medical student selection John. *Medical Education, 48*, 1104.

Speyer, R., Pilz, W., Van Der Kruis, J., & Wouter-Brunings, J. (2011). Reliability and validity of student peer assessment in medical education: A systematic review. *Medical Teacher, 33*, 572-585. Doi: 103109/0142159x.2011.610835

Sultana, Nadim, Akbar khan, Sharif, Khan, & Sadia. (2015). Learning the pelvic examination by clerkship medical students: evaluating skills by standardized patient model. Pak Armed Forces Med J: 65(4): 548-52.

Tetzlaff, J., Dannefer, E. & Fishleder, A. (2009). Competency-based assessment in a medical school: A natural transition to graduate medical education. *Journal of Education Research, 2*(4), 241-255.

Tiller, D., O'Mara, D., Rothnie, I., Dunn, S., Lee, L., & Roberts, C. (2013). Internet-based multiple mini-interviews for candidate selection for graduate entry programmes. *Medical Education, 47,* 801-810. Doi: 10.1111/medu.12224

Trejo, J., Martínez, A., Méndez, I., Morales, S., Ruiz, L. & Sánchez, M. (2014). Evaluación de la competencia clínica con el examen clínico objetivo estructurado en el internado médico de la Universidad Nacional Autónoma de México. *Gaceta Médica de México*. 2, 150.

Vrushali R Parate, Sushma S. Pande & Pushpa O. Lokare. (2016). Admission To Medical Colleges- Predictive Validity Of Selection Criteria. *In National Journal of Integrated Research in Medicine 7*(4), 98-105.

Wadee, A., & Cliff, A. (2016). Pre-admission tests of learning potential as predictors of academic success of first-year medical students*. South African Journal of Higher Education, 30*, 264-278. Recuperado de: http://dx.doi.org/10.20853/30-2-619.

Westerkamp, C., Heijne-Penninga, M., Kuks, J. & Cohen-Schotanus. (2013). Open-book tests: Search behaviour, time used and test scores. *Medical Teacher, 35*, 330-332.

Zaidi, N., Swoboda, C., Wang, L., & Manuel, S. (2014). Variance in attributes assessed by the multiple mini-interview. *Medical Teacher, 36*, 794-798. Doi: 10.3109/0142159X.2014.909587.

VIII. ANEXOS

Anexo 1: Literatura científica, relacionada con modelos de evaluación y medición en programas de formación universitaria en el área de la Salud, con una ficha resumen para cada una de las publicaciones.

1. O'Neill, L., Vonsild, M., Wallstedt, B., & Dornan, T. (2013). Admission criteria and diversity in medical school. In Medical Education, 47, 557-561.

Se realizó un estudio observacional y transversal, en el cual se recolectaron datos por 5 años para medir la diversidad social y económica de los estudiantes de medicina. Se seleccionaron 1047 estudiantes, de los cuales 454 fueron admitidos por sus notas y 620 por otros atributos. Para explorar la condición social de los candidatos en ambos grupos se obtuvo la información de los índices sociales asociados con los logros educacionales. Para ello se tomaron en cuenta los siguientes datos: un padre doctor, origen étnico, educación de los padres, paternidad del candidato al momento de elección, convivencia con los padres y padres beneficiarios de programas sociales.

Ambos grupos fueron comparados con las variables de los datos obtenidos. Las diferencias entre ambos fueron analizadas usando las dos muestras del t-test para variables continuas (edad y puGPA). Las diferencias en proporciones entre los dos grupos en las variables restantes fueron analizadas usando el chi-squared test para variables categóricas. Cálculos estadísticos fueran realizados usando STATA/ IC Versión 12 (StataCorp LP, College Station, TX, USA).

En los resultados se encontraron que los dos grupos de admisión fueron significativamente diferentes en términos de calificaciones académicas, y que las notas de las pruebas basadas en atributos solo se relacionan débilmente con pu-GPA. Sin embargo, no hubo diferencia en la admisión en ninguna de las variables sociales que habían sido asociadas con logros educacionales bajos. Por tanto, el tipo de criterio de admisión utilizado (grado o no) no diferenciaba la diversidad social de los estudiantes

de medicina admitidos. El resultado no logró satisfacer esperanzas y sugerencias previas en la educación médica, pero al menos no contradice la abrumadora evidencia de la investigación en ciencias sociales que indica que las experiencias culturales anteriores (niveles socio-económicos) son los principales predictores de los logros educativos.

2. **Vrushali R Parate, Sushma S. Pande & Pushpa O. Lokare. (2016). Admission To Medical Colleges- Predictive Validity Of Selection Criteria. In National Journal of Integrated Research in Medicine, 7(4), 98-105.**

El propósito del estudio fue encontrar la eficacia del examen de admisión, siendo este el único criterio de selección de estudiantes de cursos de MBBS (Bachillerato en Medicina y Bachillerato en Cirugía). También se trató de encontrar la correlación entre la puntuación en HSC (Higher School Certificate), la puntuación de entrada y la puntuación en el MBBS. En la institución analizada los estudiantes son admitidos por medio de un examen, el cual es completamente objetivo y consiste en 200 preguntas de selección única (MCQ) sobre física, química y biología. El criterio de elegibilidad está basado en el desempeño de los estudiantes en HSC en las materias de ciencias o la prueba equivalente en la que el estudiante debe obtener un mínimo del 50% de la nota en las materias de física, química y biología en conjunto.

Se estudiaron cuatro lotes consecutivos de estudiantes que lograron la admisión en los años 2009-2012. Para ese entonces se admitían 100 estudiantes por año, pero el número de estudiantes de la muestra varió cada año porque los estudiantes que no cumplían con los criterios de inclusión en el estudio fueron excluidos. Las principales variables estudiadas fueron el porcentaje combinado de las notas en las materias de física, química y biología obtenidas por los estudiantes en los exámenes de HSC, el porcentaje de las notas obtenidas en el examen de entrada; el porcentaje de notas obtenidas por los estudiantes en el primer año de MBBS.

Los datos obtenidos fueron pasados a una hoja Excel, se usó la versión SPSS - 16 de parámetros estadísticos. Spearman Rank fue usado para encontrar la relación entre el puntaje de HSC, el puntaje de entrada y el puntaje de I MBBS. T-test fueron utilizados para determinar el nivel de significancia ($p<0.05$). También se usó el análisis de regresión lineal y el de logística para encontrar la asociación independiente de varios factores.

El estudio evidencia que la puntuación en HSC se relaciona fuertemente con el desempeño de los estudiantes en todos los lotes de I MBBS, revelando así un valor predictivo del HSC en la selección de candidatos para el MBBS. No obstante, el examen de entrada no muestra ninguna relación con HSC ni con MBBS. La observación muestra que una puntuación alta en HSC no asegura el éxito en la admisión a una universidad médica por medio del sistema actual de entrada. Asimismo, los que lograron una nota alta en el examen de entrada no necesariamente lo hicieron bien en el examen MBBS. Usando la regresión lineal MBBS no mostró relación con el desempeño en el examen de entrada. Solo el HSC muestra una relación significativa positiva con MBBS.

En las conclusiones se hace referencia a otros estudios con los cuales compara resultados. Se llega a la conclusión de que un examen con solo un criterio de selección no es apropiado para elegir estudiantes, sugiere que el examen de entrada necesita una revisión. Ya sea que el examen sea modificado o que se utilicen y se combinen varias herramientas de selección. Es necesario evaluar la eficacia del actual criterio de selección en predicción del desempeño de los estudiantes no solo en el año preclínico, sino en los cursos médicos. Depender solamente de un dominio cognitivo no es suficiente, ya que esto no predice el desempeño de estudiantes de medicina. Exámenes de admisión reflejan solo la habilidad para memorizar hechos aislados y no habilidades esenciales para ser un buen médico. Es necesario identificar e incorporar en los criterios de selección herramientas que evalúen las características, habilidades y personalidades.

3. **Tekian. A., Hodges, B., Roberts, T., Schuwirth, L & Norcini, J. (2015). Assessing competencies using milestones along the way. Medical Teacher, 37: 399-402.**

El artículo consiste en un resumen de un simposio llevado a cabo en Praga en el 2012. Parte del contexto de discusión surgió de que existen dos modelos o discursos con respecto a la naturaleza de la educación. El primero es el modelo basado en tiempo, el cual consiste en asumir que la educación sucede en un periodo fijo determinado. Se espera que los estudiantes adquieran todo el conocimiento médico durante cierto periodo de tiempo. El segundo modelo es el basado en resultados. Este modelo consiste en enfocarse en competencias específicas en las cuales el estudiante trabaja para alcanzarlas. Por lo tanto, el tiempo (como una medida) tiene importancia mínima.

En relación con las perspectivas canadienses respecto a competencia. En Canadá hay un movimiento que plantea 7 roles o competencias necesarias en un médico: experto en medicina, comunicador, colaborador, administrador, defensor de la salud, académico y profesional. Se espera que el estudiante sea holísticamente competente al enfatizar en “roles intrínsecos”. Este movimiento ha ejercido influencia en varias partes del mundo.

Por otro lado, se argumenta que los comentarios y evaluaciones hacia los estudiantes deben ser redefinidas y reevaluadas, y que se debe utilizar un enfoque de narrativa integral basado no sólo en porcentajes o números sino también holístico que tome los elementos de forma colectiva y los utilice para describir al estudiante. Los métodos convencionales reúnen puntuaciones como un promedio, sin embargo, la medición en salud no puede hacerse así.

En cuanto al contexto Europeo y estadounidense:

- Europa: cada país tiene su propio método, pero hay cuestiones en común que discutir sobre las competencias y los objetivos: desempeño, transparencia y resultados; medidas simplificadas, momento en que son útiles, qué papel juega el tiempo.

- USA: se adoptaron primero 6 competencias: cuidado del paciente, conocimiento médico, aprendizaje basado en la práctica, profesionalismo, habilidades interpersonales y comunicación. Estas se obtienen desarrollando otras sub-competencias. Asimismo, han surgido otros conceptos y elementos prácticos: reduccionismo, incompatibilidad con el currículum basado en el tiempo, falta de un sistema de educación continua, número de objetivos/metas, formas de evaluación, comité de competencias clínicas, desarrollo de facultad.

4. **Kenny, S., McInnes, M. & Singh, V. (2013). Associations between residency selection strategies and doctor performance: a meta-analysis. Medical Education, 47. 790-800.**

El objetivo del estudio fue usar un meta-análisis para establecer cuál de la información disponible para la selección de residentes está asociada con el desempeño del médico o del residente.

El ranking de candidatos se basa en la suposición de que las estrategias de selección, incluidas la puntuación de las evaluaciones de escuelas médicas, cartas de referencia y el desempeño en la entrevista, pueden predecir el desempeño futuro. El análisis fue realizado considerando diferentes estudios publicados hasta setiembre del 2012, y usando el meta-análisis de estudio observacional en las guías para informes epidemiológicos.

Se planteó lo siguiente: considerando a los candidatos que solicitan una residencia médica, ¿cuáles estrategias de selección disponibles para el comité de selección del residente están asociadas con el futuro desempeño médico? La hipótesis nula fue que las estrategias de selección no están asociadas con el resultado del futuro desempeño médico. El rendimiento del médico se refiere a los resultados tanto durante la formación en residencia como durante el resto de la carrera.

Los resultados demostraron una amplia gama de tamaños de efecto generalmente positivo en relación con las asociaciones entre las estrategias de selección y varios resultados encontrados en la literatura actual. Estos resultados ponen de relieve una clara diferencia entre las estrategias de selección basadas principalmente en el examen y las estrategias de selección más subjetivas para las que no existe un resultado estandarizado o numérico (por ejemplo, cartas de referencia, cartas de los decanos, entrevistas). En general, las estrategias de selección basadas en los exámenes tienen una relación más fuerte con resultados no basados en el examen que las estrategias de selección subjetiva. Se encontró que las notas de las escuelas de medicina estaban asociadas con el desempeño clínico en las residencias, mientras que los registros de desempeño en la escuela médica (MSPR) tenían menor correlación con los ITER (Evaluación del Residente en Entrenamiento) de residencia. También, que las cartas de referencia y la experiencia de investigación tenían las asociaciones más débiles con el desempeño de los residentes.

El estudio demuestra que las calificaciones de las escuelas de medicina y el puntaje en las pruebas estandarizadas tienen las asociaciones positivas más fuertes disponibles actualmente para la selección de residentes.

5. **Findyartini, A., et al. (2015). Collaborative progress test (cPT) in three medical schools in Indonesia: The validity, reliability and its use as a curriculum evaluation tool. Medical Teacher, 37, 366–373.**

El objetivo del estudio fue evaluar la validez y credibilidad del cPT (Prueba de Progreso colaborativa/ Collaborative Progress Test) llevado a cabo en 3 escuelas de medicina como parte del currículo de evaluación.

El estudio se realizó con estudiantes de 1-5 año de tres escuelas de medicina en Indonesia (223 estudiantes de FM UI, 219 de FM UNAND, y 165 de FM UNS) participaron en la evaluación del proceso de aprendizaje del currículo actual. Para esto se les aplicó un examen de progreso colaborativo (cPT) de 120 preguntas de respuesta múltiple. Un examen de progreso es una prueba escrita administrada regularmente a todos los estudiantes al mismo tiempo en un mismo programa médico. La prueba se administra a todos los participantes independientemente del año del curso. Esta prueba se considera importante como una herramienta de evaluación formativa y sumativa.

El examen se aplicó a un 20% de la población de estudiantes de cada año en cada una de las escuelas. El examen evaluó siete áreas: habilidades clínicas básicas, aspectos cognitivos, razonamiento, desarrollo del crecimiento y degeneración, infección-inmunología, detección y diagnóstico y temas de gestión individual de la salud.

Se realizó un muestreo aleatorio basado en el promedio ponderado de los estudiantes. La validez del constructo se estableció evaluando el aumento correspondiente de la puntuación media del cPT al nivel del año estudiantil. Por último, la fiabilidad de la cPT se calculó utilizando el coeficiente Cronbach Alpha.

Se demostró que hubo diferencias significativas y el aumento de la puntuación media de cPT en los diferentes años. Aunque para el 3 año no hubo mucho crecimiento con respecto al segundo año, se cree que se debe al agotamiento de los estudiantes en ese año. El contenido y la validez del constructo de la cPT fueron evidentes. Hubo un aumento de la puntuación media del 1 año al 5.

También se evidenció que la cPT era una prueba válida y fiable para medir el aumento de conocimiento de los estudiantes de medicina y también fue útil para proporcionar retroalimentación para la evaluación del currículo en las tres escuelas de medicina. Los hallazgos del estudio apoyan el uso de cPT en la evaluación de la implementación del currículo. El proceso de desarrollo de la cPT y el análisis del ítem revelaron puntos de evaluación útiles para cada escuela de medicina. La evidencia de validez de contenido fue apoyada por la incorporación de artículos con los niveles

aceptados de índice de dificultad y discriminación y revisando la comprensión de los ítems de la prueba.

6. **Patterson, F., Zibarras, l., Kerrin, Lopes & Price. (2014). Development of competency models for assessors and simulators in high-stakes selection processes. Medical Teacher. 36, 1082–1085.**

En el estudio se centraron en lograr la estandarización y garantizar la calidad de la participación de los evaluadores y simuladores en el proceso de selección de Práctica General, para lo cual desarrollaron dos modelos de competencias que describen los conocimientos, habilidades y atributos asociados con cada rol utilizando una metodología de análisis de trabajo previamente validado.

Estudiaron un proceso de selección validado a gran escala al que se presentan alrededor de 6000 candidatos para 3000 puestos. La etapa final del proceso, un centro de selección (SC), que implica una prueba escrita y tres consultas simuladas, para las cuales se requieren evaluadores y simuladores. En SC, los evaluadores están obligados a observar, registrar y evaluar el desempeño de los candidatos usando escalas de calificación estandarizadas. Por tanto, los estudios demuestran que las habilidades de los evaluadores son vitales para el éxito de cualquier proceso de SC.

Los SC suelen usar pruebas de muestras de trabajo de alta fidelidad ("juegos de rol"). Se ha demostrado que estos exhiben alta validez de criterio y son populares en la educación y evaluación médica. Las simulaciones pueden utilizarse para evaluar la competencia de los médicos, a la vez que proporcionan un contexto real para comprender las complejas necesidades de atención del paciente. La estandarización del desempeño de los simuladores asegura la coherencia entre las experiencias, necesario para hacer comparaciones justas y confiables entre los candidatos.

El análisis cualitativo final resultó en dos modelos de competencias, cada uno de los cuales abarca ocho dominios de competencias. Tanto el modelo del asesor como el de simulador resultaron ser relevantes y válidos para la estandarización. Existe una superposición sustancial entre los dominios de competencia para los evaluadores y los simuladores; sin embargo, para cada modelo los indicadores de comportamiento varían, reflejando los conocimientos específicos, habilidades, comportamientos y actitudes requeridos para cada rol. Los modelos de competencias de evaluador y simulador se desarrollaron en respuesta a la necesidad de una mayor estandarización y calibración de estos roles en la selección nacional de GP. Los resultados de validación inicial indican que los modelos podrían mejorar la estandarización de la entrega de la metodología de selección y la calidad del proceso de selección en general.

Ambos modelos se utilizan actualmente en la práctica para garantizar la calidad y los propósitos de formación. Se concluye que los modelos de competencia se pueden utilizar de tres maneras: (1) reclutamiento de evaluadores/simuladores; (2) medición del desempeño de los evaluadores/simuladores y áreas de destaque para el desarrollo potencial; (3) pueden ser utilizados para la capacitación de evaluadores/simuladores.

7. **Baradaran Binazir, M., Salek Ranjbarzadeh, F., Nikasa, P. (2015). Different Methods to Assess Clinical Reasoning in Undergraduate Medical Students. Res Dev Med Educ, 4(2), 113-114.**

El razonamiento clínico son acciones en las que la información relativa a un problema clínico se mezcla con el conocimiento y la experiencia previa de los médicos y se utiliza para realizar un determinado problema. Se utiliza para que los estudiantes estén listos para situaciones clínicas complejas y para que aprendan racionalmente, los estudiantes aprenden a pensar sobre las señales y cómo estas conducen a la toma de decisiones clínicas.

Una de las formas más recomendadas para evaluar el razonamiento clínico es la Prueba de Concordancia de Guiones (SCT), pues es fácil de formar y puntuar. La prueba consiste en un conjunto de escenarios clínicos cortos. Es un sistema de puntuación complejo, que utiliza las respuestas de un grupo de expertos de la facultad como referencia y en el que varias opciones pueden ser aceptadas como respuestas. SCT distingue el razonamiento clínico del conocimiento médico en la evaluación.

Otras pruebas que evalúan el razonamiento clínico son las características clave (KF), la concordancia de guiones (SCT), los problemas de razonamiento clínico (CRP) y los rompecabezas integrativos integrales (CIP). La combinación de pruebas es una herramienta confiable para evaluar las habilidades de razonamiento clínico en estudiantes de medicina, aunque estos resultados pueden no ser generalizables para toda la población de estudiantes. Los CIP y las pruebas KF muestran el mayor potencial para evaluar el razonamiento clínico.

El uso de pacientes virtuales no es muy confiable pues tienen poca representatividad en comparación con el tema, consume mucho tiempo y su diseño es costoso. También se han usado exámenes orales para la evaluación. En este caso se diseñan casos clínicos realistas.

El uso de distintos tipos de preguntas ya existentes es mejor que buscar o diseñar un nuevo método. La combinación de diferentes tipos de preguntas aumenta la validez de la prueba considerando la evaluación de las habilidades de razonamiento clínico de los estudiantes. Además, el uso de arreglos de preguntas estáticas permitirá a los profesores diseñar buenas preguntas y aumentar la validez de construcción de estos exámenes.

8. Patterson, F., Knight, A., Dowell, J., Nicholson, S., Cousans, F., & Cleland, J. (2016). How effective are selection methods in medical education? A systematic review. Medical Education: 50: 36–60

Luego de una búsqueda formal de literatura en 194 artículos se encontraron 8 principales métodos de selección. Cada método se revisó en función de cuatro criterios: eficacia (fiabilidad y validez), procedimiento, aceptabilidad y rentabilidad.

Examen de aptitud: se presenta un conflicto en su efectividad. Algunas pruebas sugieren que los estudiantes seleccionados por medio de este método suelen ser más capaces y estar más motivados para el estudio de medicina. La utilidad de este examen depende del tipo de prueba aplicada. Estudios sugieren que algunas de estas pruebas están relacionadas con los antecedentes del candidato, o que favorecen a ciertos grupos, pero nada está definido.

Record académico: se evidencia la validez predictiva de los expedientes académicos en la selección del estudiante de medicina, pues los estudiantes con un mejor record tienen más probabilidad de tener éxito en la escuela de medicina. El logro educativo previo forma la columna vertebral académica de la selección, la progresión a través de la escuela de medicina y más allá. Las pruebas internacionales sugieren que los candidatos admitidos sobre la base de sus registros académicos tenían niveles de abandono más bajos que los que no lo eran. Sin embargo, existe la preocupación de que el poder discriminatorio de los records puede estar disminuyendo a medida que un número creciente de solicitantes tienen buenas calificaciones.

Declaraciones personales: la efectividad es variada. Se sugiere que podrían tener valor para hacer a los solicitantes conscientes de las características del título médico que están solicitando, lo que puede ayudarles a tomar una decisión más informada. Los candidatos suelen usar declaraciones personales para presentarse de una manera que consideran atractiva para los comités de admisión, pero que no necesariamente son exactas. El contenido de las declaraciones personales también puede parcializar el juicio de los que toman las decisiones de selección.

Referencias: se encontró que no tienen valor predictivo en el desempeño de los médicos. La información en las referencias suele ser cuestionable y no aporta información que hagan sobresalir a los solicitantes. Las referencias suelen parcializar a los comités de admisión.

Pruebas de juicio situacional (SJT): es un método valido y confiable; aunque la forma de aplicación puede afectar la validez (las de video son más efectivas que las escritas). Los SJT son complejos de desarrollar y hay una amplia gama de opciones disponibles en relación con formatos de artículos, instrucciones y puntuación, cuando estas opciones se calibran apropiadamente, se evidencia la fuerza en la selección de estudiantes de medicina para evaluar los atributos no académicos.

Evaluación de personalidad e inteligencia emocional: la asociación entre los rasgos de personalidad y el rendimiento en la educación médica y la formación es compleja. La evaluación de la personalidad de manera más amplia, también se ha demostrado que proporciona una validez incremental sobre los métodos cognitivos en un proceso de selección de facultades de medicina. Algunos rasgos dificultan la eficiencia de los médicos. Rasgos de personalidad "disfuncionales" en estudiantes de medicina (incluyendo paranoicos, evasivos, pasivos agresivos, antisociales, rasgos narcisistas y no cooperativos) han sido reportados como asociados con calificaciones académicas más bajas. Sin embargo, no se ha encontrado correlación positiva entre la personalidad y las habilidades médicas. También hay preocupación de que se puedan “falsificar” los test de personalidad.

Entrevistas y mini entrevistas múltiples (MMI): la entrevista tradicional no suele ser un método fuerte y carece de valor predictivo; mientras que MMI suelen ser más consistentes, sin embargo, su validez sigue siendo exploratoria y en gran parte inconclusa. Las entrevistas cara a cara estandarizadas pueden no ser comparables con las estaciones MMI basadas en escenarios que usan actores de rol estandarizados y distintas estaciones (MMIs puede medir más de una construcción por estación / pregunta de la entrevista). Se percibe que las MMI son más apropiada para evaluar una

gama de competencias y que son un proceso justo, preferible a la entrevista tradicional. Los centros de selección (SC) son altamente aceptados, aunque puede ser un método caro y logísticamente complejo.

Se evidencia que los expedientes académicos, las MMIs, las pruebas de aptitud, los SJTs y los SC son los métodos más eficaces y que son generalmente más justos que las entrevistas tradicionales, las referencias y las declaraciones personales.

9. **Sultana, Nadim, Akbar khan, Sharif, Khan, & Sadia. (2015). Learning the pelvic examination by clerkship medical students: evaluating skills by standardized patient model. Pak Armed Forces Med J: 65(4): 548-52**

El objetivo del estudio era comparar la efectividad en el entrenamiento del examen pélvico (PE) de los estudiantes de medicina en los pacientes estandarizados (SP) con el entrenamiento en pacientes regulares (RP) durante las rotaciones clínicas, para deducir cuán vital y eficiente será el rol de SP si se incorpora al currículo de pregrado de esa institución. Para tal efecto, se realizó un estudio a 67 estudiantes de 5 año de Obs/Gine (32 entrenados en RP y 35 en SP), estos fueron incluidos mediante un muestreo aleatorio simple. Luego del entrenamiento, las habilidades de PE de ambos grupos fueron evaluadas a través de OSCE. Al momento del examen, ninguno de los evaluadores sabía en cual método había sido entrenado el estudiante. Se calcularon las puntuaciones medias de evaluación, las calificaciones técnicas y las calificaciones de habilidades de comunicación de ambos grupos y se obtuvo la desviación estándar. Todos los resultados obtenidos fueron analizados estadísticamente. Después de aplicar la prueba t, se calculó el valor de p. Se consideró que el valor de p menor que 0,05 era estadísticamente significativo.

El promedio de puntuaciones de la evaluación del OSCE del examen modular combinado del Grupo RP y del Grupo SP fueron de 6,0 y 7,7, respectivamente. Del mismo modo, el puntaje promedio de evaluación para el grupo RP y SP para las

habilidades técnicas en el examen modular combinado fue de 6,0 y 7,75, y para las habilidades de comunicación fue de 6,25 y 8,0, respectivamente. Se calculó el valor de p estadísticamente significativo de <0,001. Se encontró significativamente que los estudiantes formados en SP eran más competentes en técnica y en habilidades de comunicación de la exploración pélvica en comparación con los estudiantes entrenados en RP.

Para concluir, los SP son una alternativa mucho más útil y eficiente a los RP para la formación clínica de los estudiantes de medicina de PE. Los estudiantes entrenados en SP parecen ser más hábiles. Los estudiantes entrenados en SP indicaron que el uso de pacientes profesionales ayudó a mejorar no sólo las habilidades técnicas y clínicas, sino que también mejoró su poder de decisión, autoevaluación y confianza en la aplicación del conocimiento. Además, de que tanto para estudiantes como para profesores resultó una estrategia educativa amena.

10. **Iblher, P., Zupanic, M., Karsten, J. & Brauer, K. (2015). May student examiners be reasonable substitute examiners for faculty in an undergraduate OSCE on medical emergencies? Medical Teacher. 37: 374–378**

Los investigadores postulaban que los estudiantes de medicina senior son capaces de cumplir con la responsabilidad del examinador tan apropiadamente como lo haría un médico en esta modalidad estandarizada. Para el estudio se planteó el siguiente objetivo: Comparar el efecto de los estudiantes examinadores (SE) con el de los profesores examinadores (FE) sobre el desempeño de los examinandos en un OSCE, así como sobre la evaluación posterior a la prueba en el área de manejo de medicina de emergencia. Para esto, se plantearon 2 preguntas de investigación: ¿Qué diferencias podrían evaluarse entre las puntuaciones de la OSCE para los grupos SE y FE? y ¿Qué diferencias podrían evaluarse entre la evaluación y la post-evaluación de cada estación de la OSCE en los grupos SE y FE?

Un OSCE de siete estaciones (Soporte de Vida Cardíaca Avanzado (ACLS), Soporte de Vida Básico (BLS), Manejo de Trauma (TM), Emergencias Pediátricas (PE), Síndrome Coronario Agudo (ACS), Manejo de Vías respiratorias (AM), y Emergencias Obstétricas (OE)) se aplicó a 207 estudiantes de tercer año de medicina después de completar su entrenamiento de emergencias de pregrado. Dos pistas idénticas fueron ejecutadas simultáneamente con los SE o FE. El grupo SE consistió en estudiantes de último año de medicina, quienes completaron su rotación en anestesiología y participaron voluntariamente.

Los participantes fueron asignados aleatoriamente a una de las dos pistas de ejecución simultánea: ya sea con SE (n ¼110) o con FE (n ¼98). Todos los participantes tenían 20 minutos antes del inicio de la evaluación para las instrucciones. Cada estación duró cinco minutos. A lo largo de la OSCE, los participantes tuvieron que resolver diferentes problemas médicos teóricos y prácticos mientras interactuaban con actores o maniquíes estandarizados. Luego del examen se pidió a los estudiantes que evaluaran cada estación de la OSCE y que proporcionaran su percepción general de la OSCE mediante un cuestionario estandarizado. Se utilizó la prueba t de muestras independientes y se calcularon los tamaños de efecto (Cohens d).

En general, se encontraron puntuaciones significativamente más altas en las estaciones de TM, AM y OE, así como en la puntuación global del OSCE. La estación de PE obtuvo puntuaciones significativamente más altas en la pista FE. No se encontraron diferencias entre las estaciones BLS, ACLS y ACS. Para diferencias significativas, se calcularon tamaños de efectos pequeños; excepto para la estación TM, que produjo tamaños de efecto medios. En la parte de evaluación posterior a la evaluación del estudio, los de la pista FE dieron calificaciones significativamente más altas para la estación ACS y para la "evaluación general de la OSCE “.

El estudio demostró que SE obtuvo una mejor puntuación significativa en la mayoría de las estaciones de la OSCE, pero con sólo tamaños de efecto menor. La evaluación posterior a la evaluación no mostró diferencias entre los grupos de estudio

a excepción de dos mejoras significativas para FE, aunque también con tamaños de efecto menor. Es bastante admisible y justificado alentar a los estudiantes de medicina a oficiar como examinadores en una bien definida y estructurada prueba de pregrado en medicina de emergencia de la OSCE, y así mejorar la eficiencia de la gestión de los recursos clínicos. Sin embargo, esto puede no ser aplicable en exámenes de alto interés o sumativos; en parte, debido a posibles implicaciones legales.

11. **__(2013). Mini-interviews help to recruit students for their values. Nursing Standar. 28 (2), 6.**

La Universidad de Kingston y la Escuela de Medicina de Saint George en Londres han estado usando mini-entrevistas múltiples, en las que se pide a los futuros estudiantes de enfermería completar un circuito de seis entrevistas de cinco minutos cada una. Cada entrevista tiene un valor de 5 puntos para una puntuación total de 30. Cada una es realizada por un evaluador diferente, lo que permite obtener la opinión de 6 personas. En las entrevistas los candidatos se enfrentan a escenarios diseñados para probar cualidades personales, incluyendo la empatía, el compromiso y el potencial de liderazgo.

El análisis realizado por la Universidad de Kingston mostró que los candidatos de enfermería que tuvieron dificultades o fallaron un curso habían obtenido 18 puntos o menos en la prueba de mini-entrevista múltiple. Esto permitió a los académicos vincular los puntajes a la probabilidad de tener estudiantes adecuados. Las universidades utilizan ahora una puntuación mínima de 18 para decidir a quienes aceptar.

Los escenarios utilizados en las mini-entrevistas no están necesariamente relacionados con la atención médica. Por ejemplo, uno requiere que el candidato diga cómo actuaría si perdiera el conejo de un amigo después de aceptar cuidarlo.

Además de las mini-entrevistas las universidades siguen utilizando otros procesos de selección.

12. **Shih-Chieh Liao, Tzuen-Ren Hsiue, Chyi-Her Lin & A-Min Huang (2014). Multiple mini-interviews combined with group interviews in medical student selection John. Medical Education, 48, 1104.**

Las mini-entrevistas múltiples (MMI) proporcionan un enfoque válido, confiable y defendible para la selección de estudiantes de medicina; sin embargo, estas no son suficientes para evaluar las habilidades interpersonales. Por tanto, el objetivo de la investigación fue agregar una entrevista grupal a una MMI de 7 estaciones en las entrevistas de admisión de estudiantes de medicina en la Universidad Nacional Cheng Kung (NCKU), Taiwán, con el fin de fortalecer la evaluación de las habilidades interpersonales de los aspirantes y construir una entrevista de admisión más holística y válida.

Las siete estaciones evalúan empatía; respeto por la vida; gestión de crisis; iniciativa; perspicacia, integridad y habilidades de comunicación. En la entrevista de grupo, se les muestra un video de 5 minutos, luego se realizan cuatro preguntas estandarizadas para estimular la discusión en grupo. Cada grupo consiste en seis o siete entrevistados y tres entrevistadores que monitorean y evalúan la discusión entre los entrevistados.

En base a análisis estadísticos, las correlaciones entre cada una de las siete estaciones de MMI y la entrevista grupal fueron todas positivas, variando de 0,15 a 0,42. Los coeficientes de correlación fueron todos significativos (todos $p < 0,05$), la única excepción fue para la estación 5 (perspicacia) ($p = 0,10$). El alfa de Cronbach para las siete estaciones MMI fue de 0,54. Cuando se agregaron los datos de la entrevista del grupo al análisis, el alfa de Cronbach para el MMI combinado con la entrevista grupal aumentó a 0,63.

La relación entre entrevistado-entrevistador difiere en cada tipo de entrevistas, por lo que la correlación entre ambas es buena. Además, se presenta una indicación de validez convergente, lo que significa que ambas midieron competencias similares de acuerdo con la misión y filosofía del Departamento de Medicina de NCKU.

La inclusión de la entrevista grupal con MMI aumentó el alfa de Cronbach de 0,54 a 0,63. Esto demuestra que la combinación aumentó la consistencia interna de toda la entrevista y la hizo más válida. Este estudio agregó una entrevista grupal en forma de una sesión de discusión entre compañeros al diseño de MMI y demostró que esta innovación creó una entrevista de selección de estudiantes de medicina más válida.

13. **Westerkamp, C., Heijne-Penninga, M., Kuks, J. & Cohen-Schotanus. (2013). Open-book tests: Search behaviour, time used and test scores. Medical Teacher. 35: 330–332**

La aplicación de exámenes a libro abierto se ha vuelto una necesidad para poder manejar el siempre creciente corpus de conocimiento en medicina. Este tipo de prueba reduce la necesidad de memorización y se corresponden mejor con la realidad profesional del médico. Sin embargo, los estudiantes tienden a prepararse menos cuando se trata de un examen de este tipo. Durante los exámenes a libro abierto, los estudiantes suelen consultar sus materiales de referencia muy seguido y depender de ellos totalmente. Como consecuencia utilizan más tiempo en completar un examen a libro abierto que uno sin la ayuda de libros.

El objetivo del estudio fue examinar el tiempo que los estudiantes tardan contestando preguntas a libro abierto, el número de preguntas en las que se consultó la referencia y cómo estas dos variables se relacionan con la puntuación obtenida.

En este estudio participaron 491 estudiantes de medicina de segundo año y 325 de tercero. Los exámenes del segundo año contenían ocho preguntas con dos opciones

de respuesta, 21 preguntas con tres opciones de respuesta y una pregunta con cuatro opciones de respuesta; los exámenes de tercer año contenían 23 preguntas con dos opciones de respuesta y siete preguntas con tres opciones de respuesta. El tiempo de la prueba fue de tres horas. Para cada pregunta de libro abierto, los estudiantes informaron si habían consultado sus referencias para contestarla. Para determinar las relaciones, se calcularon las correlaciones de Spearman y Pearson.

Los estudiantes de segundo y tercer año consultaron sus referencias para el 87% y el 73% de las preguntas. En promedio, los estudiantes de segundo año pasaron de 5,0 minutos en responder a una pregunta de libro abierto, variando de 3,6 minutos para preguntas con dos opciones de respuesta a 7,2 minutos para preguntas con cuatro opciones de respuesta. Los estudiantes de tercer año pasaron en promedio 4,3 minutos por pregunta de libro abierto, variando de 3,8 minutos para preguntas con dos respuestas a 5,8 minutos para preguntas con tres opciones de respuesta

El uso de cinco minutos para contestar una pregunta de libro abierto es mucho más que el tiempo necesario para responder a una pregunta de libro cerrado, que varía de 50 a 75 segundos. Los resultados también mostraron que los estudiantes usaron sus referencias para responder casi todas las preguntas, y que el comportamiento de búsqueda no estaba relacionado con el resultado de la prueba. Una posible razón para no encontrar tal relación es que tanto los buenos estudiantes como los de menor desempeño no pasan suficiente tiempo en la preparación de la prueba de libros abiertos.

Restringir el tiempo de la prueba de libro abierto podría estimular a los estudiantes a prepararse de una manera más profunda.

14. Adams, J., et al. (2015). Predictors of professional behavior and academic outcomes in a UK medical school: A longitudinal cohort study. Medical Teacher, 37, 868-880.

El objetivo del estudio es determinar si variables particulares en el momento de selección podrían distinguir entre aquellas con mayor probabilidad de convertirse en buenos médicos profesionales, para esto se siguió un grupo de estudiantes de la escuela de medicina y se analizaron todos los datos utilizados para la selección de los estudiantes, su desempeño en una variedad de pruebas de selección de potencial, evaluaciones académicas y clínicas a lo largo de sus estudios y registros de conducta profesional durante los cinco años de la carrera.

En una base de datos longitudinal se recopiló la siguiente información de los 146 estudiantes admitidos en el 2007 en la Escuela de Medicina de Hull York en UK: datos demográficos, desempeño en cada componente del proceso de selección, desempeño en pruebas psicométricas cognitivas y no cognitivas, comportamiento profesional en tutorías y en otros contextos clínicos, desempeño académico, habilidades clínicas y de comunicación en evaluaciones sumativas. Los errores profesionales se monitorean rutinariamente como parte de los procedimientos de aptitud prácticas. Se buscaron correlaciones entre las variables predictoras y las variables de criterio elegidas para demostrar la gama completa de resultados (no completar la carrera hasta la graduación con honores) y para revelar fortalezas y debilidades clínicas y profesionales.

La demografía resultó ser un importante predictor de resultados (mujeres, estudiantes más jóvenes y los ciudadanos británicos obtuvieron un mejor rendimiento general). La variable basada en el desempeño académico anterior fue un predictor significativo en los resultados de los exámenes clínicos de los estudiantes de 4 y 5 año. Algunas pruebas cognitivas también mostraron relevancia en los exámenes escritos para los de 5 año y las pruebas no cognitivas fueron predictores para el desempeño clínico en 4 y 5 año y en el comportamiento profesional. La calificación tutorial fue un predictor significativo en todos los resultados, tanto los deseables como los indeseables. Los resultados en los exámenes escritos de 1er y 2 años fueron predictores del desempeño clínico y escrito en 5 año.

Las medidas de una serie de atributos de selección y cualidades personales pertinentes pueden predecir los logros intermedios y finales de la carrera en los ámbitos académico, clínico y profesional, también ayudaron a este fin las evaluaciones tutoriales tempranas.

15. **Sánchez-Mendiola, M. y Delgado-Maldonado, L. (2017). Exámenes de alto impacto: implicaciones educativas. Investigación en Educación Médica. 6(21), 52-62.**

El objetivo del estudio fue revisar algunas de las implicaciones educativas más importantes de los exámenes de alto impacto, la literatura científica que las sustenta o no, y realizar algunas reflexiones sobre la relevancia de estos conceptos en escuelas y facultades de ciencias de la salud. Para ello realizaron búsquedas en julio-agosto de 2016, en las bases de datos: Medline, Google Scholar, Psycinfo, ERIC, Latindex, SciELO, Redalyc.

De acuerdo con la información recolectada, plantean que la evaluación con exámenes de alto impacto tienen factores tanto positivos como negativos, por lo que los argumentos académicos ceden a los aspectos afectivos y de intereses gremiales, alimentados por una ubicua falta de conocimiento de la metodología moderna de elaboración de exámenes y de los conceptos actuales de validez en evaluación. Entre los potenciales efectos positivos se encuentran: la motivación para estudiar, la estandarización de la evaluación, la mejora de la calidad educativa, la unificación de criterios y las consecuencias positivas no intencionales que según Gregory Cizek (uno de los investigadores en evaluación educativa más reconocido a nivel internacional) son: el desarrollo profesional, la acomodación de conocimiento sobre evaluación, la colección de la información, el uso de la información, las opciones educativas, el sistema de rendimiento de cuentas, la familiaridad de los docentes con sus disciplinas y la calidad de los exámenes. Por otro lado, con respecto a los potenciales efectos negativos, está: el enseñar para los exámenes, los cursos de preparación para

exámenes, los efectos en currículo formal y oculto y las inferencias inapropiadas de los resultados de los exámenes.

Un ejemplo que exponen de exámenes de alto impacto en educación en ciencias de la salud en México, es el examen de selección para realizar un curso de especialización médica. Siendo uno de los retos educativos y de recursos humanos en salud más importantes de México, dado al creciente desbalance entre los aspirantes a realizar una residencia médica y los espacios disponibles para ello en los hospitales e instituciones de educación superior. Los datos recientes del Examen Nacional de Aspirantes a Residencias Médicas de México hablan por sí mismos: en el año 2016 para 7,948 plazas concursaron 35,884 aspirantes.

El estudio propone que uno de los principios más importantes en evaluación educativa, es no tomar decisiones de muy alto impacto basados en el uso de un solo instrumento. Por ejemplo un Consejo de certificación de una especialidad médica, debe determinar la aptitud de los graduados de todas las universidades e instituciones de salud del país que egresen especialistas de esa rama del saber, y debe tomar la decisión de acreditar o no a cada uno de los sustentantes, basándose solamente en su desempeño en el examen de certificación. Esto tiene muchas implicaciones al no tomar en cuenta su origen geográfico, el contexto de las limitaciones del hospital en que se entrenó, del tipo de pacientes a los que se enfrentó durante sus rotaciones, las diferencias en los programas educativos de las universidades mexicanas y las grandes diferencias interinstitucionales en la manera como se abordan diferentes enfermedades y tipos de pacientes. Los exámenes de certificación de estos profesionales no deben desaparecer, sino realizarse con mayor profesionalismo educativo, con el fin de seleccionar a médicos generales y especialistas más competentes.

Por otra parte, el estudio demuestra la relativa escasez de conocimiento original empírico sobre los exámenes de alto impacto y sus efectos en el currículo, métodos de enseñanza de los docentes y de estudio de los estudiantes, dado que hay una ausencia casi total de investigación original en educación superior sobre esta temática en países

como México, lo que dificulta aún más el establecimiento de una discusión que arroje resultados contundentes o de consenso. La investigación encontró que más del 90% son artículos de opinión y un porcentaje bajo de investigación empírica. Es decir, la mayoría de las publicaciones sobre exámenes de alto impacto son opiniones, anécdotas o datos obtenidos sin metodología apropiada, no solo a nivel global sino local. De los trabajos de investigación la gran mayoría están publicados en la literatura anglosajona.

16. **Delgado, L. y Sánchez, M. (2012). Análisis del examen profesional de la Facultad de Medicina de la UNAM: Una experiencia de evaluación objetiva del aprendizaje con la teoría de respuesta al ítem. Investigación en Educación Médica. 1(3), 130-139.**

El artículo tuvo como propósito explorar los beneficios del uso de la Teoría de Respuesta al Ítem (TRI), para documentar evidencia de validez en un examen de altas consecuencias en educación médica. Para esto efectuaron un análisis psicométrico del Examen Profesional Teórico de la Facultad de Medicina de la UNAM, aplicado en el 2008 a estudiantes que finalizaron el quinto año del Plan Único de Estudios, de la carrera de Médico Cirujano, en la Facultad de Medicina de la UNAM.

La prueba consistió en un examen de opción múltiple, acerca de seis áreas de conocimiento: Medicina interna, Pediatría, Gineco-obstetricia, Urgencias médicas, Cirugía y Medicina familiar, evaluadas con 420 reactivos de opción múltiple, con cinco opciones de respuesta. La prueba fue aplicada en condiciones estandarizadas para todos los y las sustentantes, con papel y lápiz. Los resultados del examen se colectaron en hojas de lector óptico, que fueron capturadas para generar los datos utilizados en el análisis psicométrico.

Calcularon la confiabilidad, la dificultad y la discriminación con la Teoría Clásica de los Test (TCT). Y utilizaron el modelo de tres parámetros de la TRI. Con las dos

aproximaciones se seleccionaron los mejores ítems, y se estimó la longitud de la prueba con la fórmula de Spearman-Brown. El examen fue respondido por 882 sustentantes y los resultados obtenidos fueron un índice de dificultad de 0.55 y una confiabilidad de 0.93. Con el modelo de 3pl-TRI, el examen es informativo en niveles de habilidad cercanos al promedio en la escala theta. El parámetro de discriminación promedio (a) fue 0.67, el parámetro de dificultad (b) fue 1.21, y el parámetro de seudoadivinación (c) fue 0.18.

El estudio muestra que el Examen Profesional Teórico de la Facultad de Medicina de la UNAM era susceptible de reducirse en longitud, obteniéndose o incluso mejorando la precisión en la estimación de los niveles de habilidad de los sujetos. Manteniendo una alta confiabilidad. La mayoría de los ítems en la prueba original (84%) tuvieron un buen ajuste al modelo 3pl-TRI, y en la versión acortada la gran mayoría (97%) tuvieron un ajuste similar.

En línea con lo anterior, una prueba de menor longitud puede traer ventajas como: mejorar la eficiencia del instrumento: disminución de cansancio y desgaste por parte de los sustentantes al enfrentarse a un examen más corto, ahorro de recursos (de tiempo y económicos) en el diseño y aplicación de una prueba con menor número de ítems, ingreso a la prueba de reactivos nuevos con fines de conocer su calidad métrica, con el objetivo de crear y nutrir un banco de reactivos calibrados y con un amplio repertorio para medir distintos niveles de habilidad, particularmente en el rasgo de interés. Por lo anterior se sugiere trabajar un banco de reactivos de manera permanente, con ítems calibrados y que cubran el constructo a evaluar, para estar en condiciones de aplicar instrumentos de evaluación que identifiquen apropiadamente las habilidades necesarias en las y los sustentantes.

Una de las conclusiones importantes de este trabajo es que los modelos de TCT y TRI, si bien tienen diferencias substanciales, en la práctica se pueden utilizar de manera complementaria para lograr una práctica de evaluación educativa más profesional y eficaz, ya que cada uno tiene virtudes y limitaciones que se pueden

ponderar de acuerdo a la situación de evaluación específica. De manera particular, el modelo de TRI permite analizar de una manera más integral los ítems que componen un test, permitiendo seleccionar aquellos que muestren mejores parámetros en cuanto a los valores de dificultad, discriminación y seudoadivinación y, con un menor número de ítems, determinar la habilidad de los examinados. Además, permite identificar los reactivos que proporcionen mayor información de los niveles de rasgo en los que se tenga particular interés. Con esto, se logran seleccionar a priori los reactivos cuyo error de medición sea menor en los niveles de rasgo que se pretenden medir y así conformar la prueba más precisa a esos valores de dominio.

17. **Ponce de León-Castañeda, M., Ortiz-Montalvo, A., Bonilla-González, I. y BerlangaBalderas, F. (2006). Evaluación profesional de la competencia clínica a través del Examen Profesional. Archivos en Medicina Familiar. Vol.8 (2) 119-129**

El objetivo del estudio fue describir el procedimiento para evaluar de manera integral la competencia clínica mediante el Examen Profesional de la Carrera de Médico Cirujano en la Facultad de Medicina de la Universidad Nacional Autónoma de México y presentar los resultados obtenidos por las y los sustentantes.

En la investigación analizaron los resultados de dos exámenes profesionales (teórico-práctico), aplicados en los años 2001 y 2002. Con respecto a la fase teórica utilizaron las bases de datos con los resultados de los exámenes aplicados a 1 185 alumnos en 2001 y a 1 159 en 2002. El programa utilizado para la calificación y el análisis estadístico del examen fue el Kalt versión 5.0, el cual emite número de respuestas correctas totales y por área, así como índices de grado de dificultad, poder de discriminación y confiabilidad. Analizaron las frecuencias simples de los aciertos totales y por cada una de las cuatro áreas troncales.

En la fase práctica utilizaron las bases de datos con los resultados de 927 (2001) y 610 (2002) sustentantes (en este último año 339 alumnos optaron por ser evaluados

mediante un examen tipo ECOE). Contaron además con una guía de evaluación previamente verificada en cuanto a validez y confiabilidad mediante análisis por jueces, factorial y discriminante. La guía quedó estructurada en cuatro áreas: relación médico-paciente, interrogatorio, exploración física y réplica oral. Utilizaron también una escala tipo Likert para su evaluación (no realizado 1, excelente 5). Las anotaciones de los profesores se transformaron a valores numéricos 1 a 5, y la confiabilidad fue analizada mediante alfa de Cronbach. Los datos de las guías se capturaron y analizaron en una base de datos apoyados en el programa FoxPro versión 4.0. El presidente del jurado, indicó en la hoja de criterios para la selección del caso clínico el diagnóstico del paciente. A partir de ello realizaron un listado de los padecimientos seleccionados y se analizó su frecuencia.

En cuanto a los resultados, en la fase teórica el número de alumnos que se presentaron fue de 1 185 y 1 159 y en la Fase práctica fue de 927 y 610 respectivamente. Se tuvo una confiabilidad de 0.98 y 0.94 en el examen teórico, ésta no se determinó en la parte práctica. El rendimiento de los alumnos regulares, en las dos fases, fue mayor que el de los que tardaron más años en cubrir el plan de estudios. Pediatría es el área con mejor rendimiento. Áreas como diabetes e hipertensión presentaron bajo rendimiento. Se encontraron puntajes bajos en tres de las cuatro dimensiones exploradas en la Fase práctica.

La media de aciertos en ambos exámenes tanto en la fase teórica como en la práctica fue mayor en los alumnos regulares, en tanto que a medida que la generación de procedencia se aleja de la estudiada (alumnos no regulares), el rendimiento va disminuyendo progresivamente. Se encontró además que son los alumnos regulares, los que tienen promedio más alto en la carrera. Las patologías exploradas en los casos clínicos, son problemas de manejo frecuente del médico general, sin embargo el examen teórico permitió conocer cuáles son aquellas áreas en donde el porcentaje de aciertos es muy bajo y por lo tanto es necesario enfatizar el contacto de los alumnos con estos problemas y su manejo terapéutico como en el caso de epilepsia, diabetes

mellitus e hipertensión arterial, todos ellos padecimientos presentes en el perfil epidemiológico y en los motivos de mayor demanda en el primer nivel de atención.

Según los resultados de este estudio, se requiere reforzar el entrenamiento clínico de las y los alumnos en las acciones que explora la fase práctica del Examen Profesional; principalmente en la réplica oral (integración y razonamiento clínico, toma de decisiones), ya que en esta área es donde los alumnos en general presentan menor rendimiento. El análisis realizado al Examen Profesional Teórico-Práctico, muestra que los instrumentos utilizados permiten evaluar satisfactoriamente lo que un egresado de la carrera de medicina debe saber, saber hacer y hacer, es decir, su competencia de egreso. Establece en los alumnos una relación positiva, entre su promedio en la carrera y el resultado obtenido en el Examen Profesional. Esta relación nos permite verificar que esta forma de evaluación es congruente con las calificaciones otorgadas por los profesores que participaron en la formación del alumno, por lo tanto nos permite determinar si un alumno posee las competencias básicas terminales de la Carrera de Médico Cirujano.

18. **Salazar, O., Vélez, C. y Zuleta. J. (2015). Evaluación de conocimientos con exámenes de selección múltiple: ¿tres o cuatro opciones de respuesta? Experiencia con el examen de admisión a posgrados médico-quirúrgicos en la Universidad de Antioquia. IATREIA. 28(3), 300-311.**

La investigación tuvo por objetivo evaluar el efecto de la reducción del número de opciones de respuesta por pregunta sobre los indicadores sicométricos de un examen de ingreso a estudios médicos de posgrado.

El estudio fue de carácter descriptivo para el cual se utilizaron dos enfoques teóricos de la medición: la teoría clásica y la teoría de respuesta al ítem. Para ello tuvieron en cuenta los exámenes de 2.539 aspirantes a ingresar a 21 programas de posgrado clínico-quirúrgicos de la Facultad de Medicina de la Universidad de Antioquia,

Medellín, Colombia, en el año 2014. Dicha prueba evaluada consta de 70 preguntas de selección múltiple constituidas por un tallo con la descripción de un caso clínico de cualquiera de las especialidades a las cuales aspiran los evaluados(as) y cuatro alternativas de respuesta, que pueden ser de dos tipos: el primero con una sola respuesta verdadera, y el segundo, con todas las opciones de respuesta verdaderas, pero con una de ellas más adecuada que el resto para la situación clínica específica.

Las preguntas fueron elaboradas por profesores(as) de las diferentes especialidades, considerando el perfil epidemiológico de Colombia. Una comisión de cuatro profesores(as) con experiencia en elaboración de pruebas, evaluaron una a una las preguntas con el fin de garantizar su validez. En este proceso, mantuvieron el tallo de la pregunta, es decir, los elementos estrictamente necesarios para entender la situación problemática que se presenta; en caso necesario se mejoró la redacción, privilegiaron las preguntas positivas y evitaron que hubiesen trucos o aspectos diferenciadores para la respuesta correcta que no sean los verdaderamente importantes desde el punto de vista clínico. Con respecto a las opciones de respuesta, buscaron alternativas incorrectas, pero que parecieran admisibles, es decir, que no sean descartadas de manera obvia, sino que tuvieran la posibilidad de atraer a las y los aspirantes que tienen menos, pero no a las y los que tienen más conocimiento del tema; que fueran de igual extensión y con forma y estilo gramatical similares, concordantes con la pregunta, que no dieran claves de respuesta para esa o para otra pregunta; ordenadas de manera aleatoria o en un orden lógico cuando la pregunta lo amerite. Para la investigación, eliminaron la opción de respuesta elegida con menor frecuencia y se la reemplazó por azar de entre las tres restantes.

Los resultados arrojaron que en 33 preguntas (47,1%) las tres opciones incorrectas fueron funcionales, 29 (41,4%) tuvieron dos opciones incorrectas funcionales, 7 (10%) tuvieron solo una opción incorrecta funcional y una pregunta (1,4%) no tuvo opciones incorrectas funcionales, es decir, solo 52,9% de las preguntas tuvieron tres opciones funcionales de respuesta. No se encontró diferencia en la dificultad, la discriminación, el error estándar de la medición, el alfa de Cronbach ni el coeficiente de

correlación biserial (teoría clásica de la medición); tampoco en la medida de dificultad de los ítems o de habilidad de las personas (teoría de respuesta al ítem) entre las pruebas con tres y cuatro opciones de respuesta. La prueba con tres opciones conservó un buen ajuste. Por lo cual concluyen que una prueba con tres opciones de respuesta se comportó tan bien como su contraparte de cuatro opciones. La diferencia más importante es el aumento de preguntas con opciones de respuesta funcionales con la prueba de tres opciones, mientras que los diferentes índices y demás parámetros de evaluación son bastante similares entre ellas. Los índices más importantes para evaluar las pruebas desde esta teoría son la dificultad y la discriminación. El estudio muestra que si el examen hubiera tenido preguntas con tres opciones de respuesta, en vez de cuatro, la coincidencia en la decisión hubiera sido del 95,4% y solo en 5 casos de los 2.539 (0,2%) se habría tomado una decisión diferente.

19. **Arenis, Y. y Pinilla, A. (2016). Evaluación de estudiantes de posgrado en ciencias de la salud. Acta Médica Colombiana. 41(1), 49-57.**

Este artículo tuvo como propósito sensibilizar e invitar a docentes y estudiantes de posgrado a reflexionar y renovar procesos educativos y en especial la evaluación, desde su acepción paradigmática, metodológica y axiológica. Esto con el fin de optimizar la formación profesional integral que abarca las diversas competencias profesionales (CP) como las específicas (propias de cada profesión o disciplina) y las transversales o comunes éticas, administrativas, pedagógicas y educativas, comunicativas e investigativas, de tal forma que se comprenda por qué es prioritario redimensionar la evaluación como un proceso dinámico, flexible, participativo, concertado; además, diagnóstico (de entrada), formativo y sumativo o terminal por medio del cual un docente brinda apoyo a un estudiante, quien es un profesional en formación posgraduada.

A nivel de posgrado, la literatura evidencia que los procesos de enseñanza-aprendizaje e investigación se deben orientar al desarrollo de competencias

profesionales, a su vez, el proceso de evaluación se articula en éstos según los nuevos enfoques pedagógicos, así, la evaluación es soporte para el aprendizaje y la formación integral de un estudiante de posgrado. En los programas de posgrado en ciencias de la salud, las competencias profesionales incluyen las genéricas o transversales y las específicas, las cuales tienen dimensiones que se desarrollan de forma progresiva. Según la pirámide de Miller, un estudiante para desarrollar una competencia avanza a través de sus diferentes dimensiones: saber-conocer, saber-hacer o saber-cómo, saber-emprender y saber-ser. Entonces, es indispensable contemplar la pirámide de cada competencia profesional, actualizar el concepto de evaluación como proceso eje para el aprendizaje y la enseñanza, todos estos enmarcados en un modelo pedagógico que oriente el enfoque del currículo el cual defina lo qué es evaluación, qué se debe evaluar, cómo evaluar y para qué evaluar. Profundizando más en la complejidad de la evaluación, es necesario entender cómo un instrumento debe estar adecuado a una dimensión o a las diversas dimensiones a evaluar de una competencia profesional; en programas de posgrado de ciencias de la salud, se pueden utilizar nuevas técnicas e instrumentos de evaluación, tales como: la carpeta de aprendizaje (CA) o portafolio, la evaluación clínica objetiva estructurada (ECOE), las rúbricas, el ejercicio de examen clínico reducido (Mini CEX), entre otras. Por lo anterior, es necesario repensar el paradigma educativo y utilizar los instrumentos pertinentes para las competencias profesionales que se desean evaluar, los cuales deben estar contextualizados y asociados a una situación profesional para tener éxito no sólo en el proceso de evaluación sino en la formación de los futuros egresados. Por tanto, se aportan elementos para que los docentes reflexionen sobre su actuación como evaluadores y, para que la evaluación enriquezca a docentes y discentes en lugar de ocasionarles conflicto en diversos niveles y escenarios.

20. Duré, I., Dursi, C., Raffoul, M. y Caffarena, W (2014). Examen Único de Ingreso a Residencias Médicas. Análisis estadístico de Examen Único 2013. Ministerio de la Salud. Presidencia de la Nación. Recuperado de http://www.msal.gob.ar/observatorio/images/stories/desctacados/examenunico/Analisis-estadistico-EU-2013.pdf

En América Latina, 9 países cuentan con un examen de ingreso centralizado a nivel nacional (Bolivia, Honduras, Costa Rica, México, República Dominicana, Chile, Paraguay, Perú y Uruguay), de igual modo que sucede en otros países del mundo. La realización y la toma del examen en muchos de ellos están a cargo del Ministerio de Salud y en otros del Ministerio de Salud y de Educación en forma conjunta.

En Argentina realizan El Examen Único de Ingreso a Residencias Médicas (EU), el cual es parte de un proceso de construcción federal en torno a la formación de profesionales de salud, particularmente dirigido a mejorar la gestión y la calidad de las residencias.

Los concursos de ingreso se han organizado en general contemplando una prueba escrita y tomando el promedio como uno de sus ponderadores, pero incorporando una diversidad de otros elementos como la entrevista personal por parte de los jefes de servicio, antecedentes académicos, idioma, pertenencia a la jurisdicción, los cuales se combinan en distintas proporciones. En años anteriores, esta situación, sumada a las dificultades de algunos jóvenes para trasladarse, configuraba un marco de inequidad en el acceso a estas plazas de formación. En muchas jurisdicciones la generación del examen y la entrevista las realizaba un solo actor, dando lugar a una pérdida de transparencia en el proceso. La falta de criterios comunes para realizar los exámenes potenciaba la distribución inequitativa de aspirantes.

De este modo, el EU se configura como un proyecto de centralización del ingreso a las residencias que se instala sobre tradiciones e identidades culturales diversas. Constituye la posibilidad de trabajar en favor de la equidad de oportunidades en el acceso de formación de especialistas, ya que promueve una mejor distribución de los

jóvenes profesionales entre las provincias y simplifica el proceso de concurso para los aspirantes, en tanto estos pueden concursar en distintas jurisdicciones con el puntaje que obtienen en un mismo examen.

Las características más salientes del proceso en la actualidad son:

- La unificación de la instancia de preinscripción, el cronograma de los concursos y el instrumento de examen.
- La posibilidad de readjudicar cargos vacantes en otras especialidades o provincias participantes con el resultado del examen, si no se hubiera accedido a la vacante deseada.
- La existencia de un Comité Técnico de Examen Único conformado por un referente de cada provincia participante. Todas las decisiones se toman por consenso en el Comité y quedan asentadas en actas.
- La elaboración conjunta de la tabla de especificaciones, el temario del examen y la lista de bibliografía. Las provincias remiten preguntas basadas en este temario y el examen se consolida en el Ministerio de Salud de la Nación.
- La prueba de selección múltiple única, que se aplica el mismo día y a la misma hora en todas las provincias, y utiliza igual escala de calificación.
- La posibilidad de que los profesionales puedan rendir examen en cualquiera de las sedes provinciales, independientemente del concurso al que apliquen.
- El proceso de seguimiento online de los postulantes y asignación de cargos a través del Sistema Integrado de Información Sanitaria Argentino (SIISA), que permite la integración de las distintas etapas y transparenta las acciones ante todas las jurisdicciones participantes.

- La corrección centralizada, a partir de un proceso de escaneo local de los exámenes y lectura óptica a cargo de la Universidad Tecnológica Nacional

21. Lievens, F. (2013). Adjusting medical school admission: assesing interpersonal skills using situation judgement tests. Medical Education, 47, 182-189. doi:10.1111/medu.12089

El objetivo de la investigación fue examinar la validez de la prueba de juicio de situación y su predicción a medidas de resultado por medio de un estudio del diseño de un estudio multitudinal de características compartidas.

El estudio toma lugar en Bélgica con población de habla holandesa y se enfocó en recolección de información de la prueba de admisión de la Escuela de Medicina Holandesa en Bélgica entre 1999 y 2002. Los participantes fueron 5444 candidatos que hicieron la prueba de admisión a la universidad. De esta población 2161 (39.7%) pasaron el examen, de la personas que aprobaron 1788 comenzaron el primer año en la escuela de medicina y 373 eligieron no para estudiar medicina.

Con el fin de eliminar criterios de juicio, el estudio combinó múltiples grupos de estudiantes de medicina para que la muestra fuera más grande. Se estandarizaron las medidas de predicción y las de resultado para que los puntajes tuvieran las mismas implicaciones. Y a la hora de seleccionar la población candidata, la variabilidad de los resultados se fue reduciendo ya que sólo quienes son admitidos alcanzan estudios en medicina. Algunos dejaron la universidad.

Se evaluaron los cursos en interpersonales y no interpersonales. Se incluyó una escala de calificación desempeñando la función de médico. Después de 7 años, 261 estudiantes (28.2%) ingresaron a esta evaluación práctica por un periodo de 2 años.

- Medidas de resultado: GPA's de: primer año, cursos de comunicación interpersonal, cursos no interpersonales, bachillerato, maestría, último año (después de 7 años).

- Medidas de control: GPA de educación secundaria. 3049 estudiantes lo reportaron (56.0%).

En este estudio se utilizaron los datos de rendimiento para los 1432 estudiantes de medicina de los tres universidades más grandes de la parte holandesa de Bélgica, con promedio de edad entre 18 años y 10 meses, siendo 36.7% hombres; 63.3% mujeres.

Para ello se utilizaron diversas técnicas para la recolección de datos, entre ellas:

- Pruebas cognitivas: las pruebas cognitivas incluyeron cuatro pruebas de conocimiento científico (biología, química, matemáticas, física), con 10 ítems compuestos por 4 opciones de respuesta (respondidas en 180 minutos). Una prueba general de habilidad mental constaba de 50 elementos (verbales, numéricos o figurales), cada uno con cinco posibles respuestas (respondidas en 50 minutos).

- Pruebas de juicio situacional: evalúa las habilidades interpersonales, por medio de la simulación de una situación real, la cual era filmada y presentada a las y los candidatos (por ejemplo dar una mala noticia a un paciente, indicándole a un paciente que no tome medicina tradicional). El SJT consistió en 30 preguntas de opción múltiple con cuatro posibles respuestas. Los candidatos tenían 25 segundos para responder a la pregunta ('¿Cuál es la respuesta eficaz? ') relacionado con la escena.

- Cuestionario anónimo: para recolectar información sobre las percepciones sobre la prueba de admisión, evaluaban los componentes de la evaluación (relevancia con la profesión) y el nivel de dificultad en una escala Likert.

Los resultados respecto a la validez de predictibilidad de SJT's interpersonales, muestran que SJT's y las pruebas cognitivas se complementan la una a la otra. La prueba SJT predijo el puntaje en una OSCE en comunicación y habilidades interpersonales, desempeño en un caso basado en una entrevista de panel, y el desempeño como médico después de 9 años. La validez de las pruebas cognitivas fueron más altas que las de SJT.

Para establecer la validez de las pruebas para el grupo de candidatos total, los coeficientes de la prueba cognitiva y la SJT's basada en video fueron correlacionados para el total de la población (n = 5444). La correlación para la restricción de rango tuvo un mayor impacto en la validez de las pruebas cognitivas (incremento de alrededor de 0.10.) que en las SJT.

El resultado medio de validez de SJT para predecir el resultado de un sólo curso interpersonal fue de 0.31 y el resultado medio de las pruebas cognitivas para predecir el resultado de un curso de medicina fue de 0.44.

El valor agregado de SJT interpersonal mostró valor agregado significativo en la predicción de 4 resultados: interpersonal GPA, desempeño OSCE, desempeño médico, y desempeño en la entrevista de panel, con proporciones de variación de 4.4%, 1.4%, 2.2% y 3.4% respectivamente.

Con relación a la SJT y diferencias de género, los hombres (n = 1996) tuvieron puntajes ligeramente más altos que las mujeres (n= 3448) en las pruebas cognitivas ($t(5442) = 8.97$, $p < 0.01$, $d = 0.25$). En la SJT sucedió el resultado contrario ($t(5442) = -9.13$, $p < 0.01$, $d = -0.26$). Por lo tanto, utilizando ambas pruebas podría llevar a proporciones equitativas de candidatos que pasen la prueba de admisión.

En esta línea, la SJT y las percepciones de los candidatos, mostraron que la SJT interpersonal tenía más validez que las pruebas cognitivas (t (1470) = 20.50, p < 0.01, d = 0.55). Esto sugiere que la SJT fue percibida como más relacionada a la profesión. Se percibió la SJT parcialmente más fácil que las pruebas cognitivas (d= -0.98)

22. **Ali, A. & Ali, Z. (2013). Admission policy of medical colleges: Evaluating validity of admission test in Khyber Pakhtunkhwa, Pakistan. Journal of Research and Reflections in Education, 7(1), 77-88. Recuperado de http://www.ue.edu.pk/jrre**

El objetivo de la investigación fue examinar la validez de la prueba de ingreso realizada por la Agencia de Pruebas Educativas y de Evaluación (ETEA) para la admisión a todas las universidades de medicina de Khyber Pakhtunkhwa (KP), provincia de Pakistán.

Se escogieron todas las universidades de medicina y odontología de KP que están bajo el control administrativo del gobierno provincial (Khyber Medical College, Ayoub College, Saidu Medical College, y Gomal Medical College). El total de estudiantes fue de 2944 estudiantes, siendo 1975 hombres y 968 mujeres. Desde inicio hasta su graduación que estuvieron matriculados del 2000 al 2005.

El criterio de medición que se utilizó para este estudio fueron las notas de los estudiantes. Dichas notas fueron recolectadas por la oficina de control de evaluaciones de la University of Peshawar, Hazara University Mansehra y University of Malakand. Los datos de estudiantes cuya información estaba incompleta no fueron incluidos en el análisis. La información recolectada fue organizada, tabulada e ingresada en SPSS-16 para el análisis de los siguientes procedimientos estadísticos: Se analizaron las notas, las correlaciones de la prueba de admisión entre predictores incluyendo notas de F.Sc, notas de ingreso, nota de Mérito (nota de mérito es la combinación entre el examen de admisión y F.Sc). Además se realizó análisis regresivo para evaluar la validez de los

predictores anteriormente mencionados para predecir criterios como notas y notas de exámenes.

Los resultados obtenidos fueron que la asociación entre la nota de mérito y la de admisión también fue significativa excepto para cuarto año. Y que las notas F.Sc (con una media de 2.0) muestran correlación significativa con todas las notas de las pruebas profesionales de los cinco años. Además el análisis muestra que hay asociaciones significativas para F.Sc, mérito y admisión para la mayoría de los casos.

Se concluye que todos los factores de predicción (F.Sc, prueba de admisión, y notas en general) están relacionadas a las notas generales de los estudiantes durante sus años de estudio. Sin embargo, de los tres, el que mostró una relación más cercana al desempeño académico es el F.Sc.

23. **Obsnorne, A., Hawkins, S., Pournaras, D., Chandratilake, M. & Welbourn, R. (2014). An evaluation of operative self-assessment by UK postgraduate trainees. Medical Teacher, 36, 32-37. doi:10.3109/014259x.2013836268**

El objetivo del estudio fue investigar la validez, la auto-evaluación basada en procedimientos en el quirófano y evaluar las necesidades de aprendizaje y cambio en la práctica identificada. La población de estudio fue de 25 pasantes de cirugía en el Reino Unido que solicitaron evaluación externa de PBA en apendicectomía, para la elección de participantes no hubo ningún criterio de exclusión. El total de estudiantes fueron: 14 hombres y 11 mujeres, con una media de edad de 29 (rango de 25-33). Su nivel medio de experiencia era de pasante de especialidad de tercer año, habiendo realizado una media de 5 apendicectomías laparoscópicas previamente.

Se realizó un análisis cuantitativo para la comparación de resultados globales y para la satisfacción del pasante se utilizó la “prueba de rangos con signo de Wilcoxon para datos apareados”. Se analizó la correlación Spearman y también la distribución no

Gaussiana. El análisis se realizó utilizando el software GraphPad (2012): GraphPad Prism versión 5.00 para Windows.

En el análisis cualitativo se recolectaron todos los comentarios de los pasantes y los supervisores. Los cuales fueron comparados a las necesidades de aprendizaje identificadas en la PBA externa. Finalmente, el estudiante y el supervisor llenan una encuesta de satisfacción en una escala Likert.

Los estudiantes realizaron una autoevaluación del PBA y los supervisores también llenaron una evaluación sin saber los resultados de la autoevaluación PBA. Y se compararon ambos resultados para ver si los estudiantes se sobre estimaron o se subestimaron.

Se realizó la evaluación basada en el procedimiento (PBA), método de evaluación externa en donde los supervisores intervienen sólo si los estudiantes solicitan ayuda o si se observa que el estudiante no es competente para continuar y la seguridad del paciente está en riesgo. Este método utiliza checklists en 6 secciones tituladas: consentimiento, plan preoperatorio, preparación pre operatoria, exposición y cierre, técnica intraoperatoria, y administración post operatoria. El entrenador lleva un registro de ello y se les evalúa en una nota resumida entre 1 (incapaz de realizar el procedimiento bajo supervisión) a 4 (competente de realizar el procedimiento sin supervisión).

En cuanto a los resultados no se encontró diferencia significativa en los resultados de satisfacción de los pasantes y los resultados de las escalas Likert de las evaluaciones externas PBA y la autoevaluación PBA. Hubo correlaciones positivas entre los resultados globales de las autoevaluaciones PBA y los niveles de entrenamiento (r1/4 0.88, 0.73-0.95, correlación Spearman, $p<0.01$) así como el número de apendicectomías previas realizadas por los pasantes (r1/4 0.79, 0.57-0.91, correlación Spearman, $p<0.01$). No se mostró diferencia significativa entre la evaluación externa y la autoevaluación PBA.

Los temas principales de las habilidades técnicas fueron evidentes en ambos grupos (evaluación PBA externa y autoevaluación PBA). Sin embargo, habilidades no técnicas mostraron diferencias significativas: evaluación externa identificó comunicación y chequeos de seguridad de la OMS (con el equipo de anestesia y las enfermeras), mientras que la evaluación identificó concientización de la situación, toma de decisiones y liderazgo.

Información que considere importante: La confiabilidad tiene diferentes retos debido a la rapidez de cambio de las condiciones en el quirófano, incluyendo pero no limitado al paciente, el equipo y el grupo de trabajo. Retroalimentación individualizada de calidad por parte del supervisor sigue siendo la forma preferida de recibir retroalimentación.

24. López, L. (2017). Evaluación clínica objetiva y estructurada (ECOE) en la maestría de Enfermería Ginecobstétrica y Perinatal: una sistematización de la experiencia. Enfermería Actual de Costa Rica. Universidad de Costa Rica. (33). 1-17

Como parte de las premisas que da lugar a este artículo se afirma, que una de las grandes desventajas de los ECOE es el estrés que puede provocar en el estudiante, lo cual afectaría su desempeño, no obstante, a pesar de esto, han sido bien recibidos. De igual manera apunta que una limitante es la posibilidad de ser exhaustivo en la evaluación.

Ante esto nace el objetivo de definir la pertinencia del ECOE en cuanto a cumplir los objetivos de aprendizaje del curso, como metodología en la evaluación la participación de 24 estudiantes matriculados en el curso PF-0512 Enfermería Ginecológica, Obstétrica y Perinatal I, el cual se impartió en el primer semestre del año 2016. Todos los participantes rotaron por los tres escenarios simulados (estaciones); además se utilizaron guías de simulación clínica para poder dar un valor sumativo.

Los estudiantes conocían los procedimientos y los habían realizado en sus prácticas clínicas, ante esto se evalúa la capacidad para poner en práctica elementos de escucha y comunicación activa, la capacidad para organizar la consulta de tal manera que la información recibida por la persona atendida (paciente estandarizado) sea comprendida y la capacidad para aplicar los conocimientos y brindar la recomendación correcta según la situación de la persona.

La elaboración de un escenario y la creación de la matriz de evaluación necesarias para realizar la prueba pueden omitir aspectos importantes de la formación que pueden ser evaluados también, por lo que se sistematiza el proceso de las tres estaciones que se emplearon en la prueba (consulta de control prenatal, consulta de planificación familiar, toma de citología vaginal).

Para llegar a este punto es importante mencionar que se realizaron estudios de varios usos del ECOE en otros países, por ejemplo, sobre un estudio de tipo observacional realizado en La Escuela de Enfermería y Partería de la Universidad de Babol de Ciencias Médicas, este concluyó que la combinación de la práctica observada con un ECOE, es confiable para medir competencias en los estudiantes.

También se revisó un estudio de España sobre un grupo de matronas de la Asociación Andaluza de Matronas quienes participaron en el diseño de una evaluación clínica objetiva estructurada para residentes, en este caso las conclusiones indican que la participación fue amplia y la valoración global fue positiva y además que la información obtenida a nivel individual permite priorizar las áreas de conocimiento y adaptar los métodos de aprendizaje.

Es importante a la hora de realizar una evaluación como el ECOE considerar que cada persona tiene un estilo diferente de consulta lo cual puede estar influenciado por la personalidad del estudiante, siendo un elemento importante a la hora de evaluar.

Este artículo apunta que como aspecto a tener cuenta en la implementación de la metodología, es la cantidad de recursos utilizados y el planeamiento cuidadoso y

anticipado que requiere la simulación clínica, sin dejar de lado la disponibilidad de pacientes estandarizados. “El ser riguroso en la aplicación de la metodología, y el uso adecuado de los recursos, permiten una adecuada implementación de las estaciones que, definitivamente, repercutirá en el éxito de la prueba”. (López, 2017, p. 15)

Sobre los ECOE, se afirma que:

> Tienen mayor objetividad que cualquier otra evaluación práctica, amplio rango de diferentes examinadores y por lo tanto menor riesgo de parcialidad, menor riesgo de que los estudiantes sean evaluados por diferentes personas, visto positivamente por estudiantes y docentes, amplitud de habilidades evaluadas, se reduce la escogencia al azar y aumenta la igualdad de experiencias entre estudiantes, motivación para el aprendizaje, y alto nivel de confiabilidad y validez. (Rushfort, citado en López, 2017, p. 6-7)

25. **Pugh, D., Bhanji, F., Cole, G., Dupre, J., Hatala, R., Humphrey-Murto, S., Touchie, C. & Wood, T. (2016). Do OSCE progress test scores predict performance in a national high-stakes examination? Medical Education, 20, 351-358. doi: 10.1111/medu.12942**

El propósito del estudio fue establecer pruebas de validez para el uso de una prueba de progreso de la OSCE, en el ámbito de las relaciones con otras variables, examinando la asociación entre las puntuaciones de este examen y las de un examen nacional de alto rendimiento/exigencia.

La población de estudio fue de 244 residentes entre 1-4 año de estudios de posgrado de la Universidad de Ottawa. 115 participaron en la prueba IM-OSCE. Del total de participantes, 189 tenían información sobre la prueba RCPSC IM (Prueba Comprensiva Objetiva de Evaluación Estructurada en Medicina Interna de la Universidad Real de Médicos y Cirujanos de Canadá).

Los residentes fueron evaluados usando checklists pertinentes en cada caso así como por medio de escalas de calificación global. Ambas notas componían una nota total general. Dado que cada año había candidatos, evaluadores y contenido, los puntajes totales IM-OSCE de cada año fueron estandarizados con una media de 500 y una desviación estándar de 100. Los residentes podían tomar la prueba RCPSC IM al final de su cuarto. La prueba escrita fue aplicada 2-3 meses después del IM-OSCE.

Se realizaron análisis correlacionales, cada año la prueba de situación variaba; la confiabilidad conocida (calculada usando alfa de Cronbach) para IM-OSCE estuvo en un rango entre 0.61 y -0.70. Y análisis de regresión lógica, donde se correlacionaron los resultados de IM-OSCE y RCPSC IM utilizando una variable binaria de “alto riesgo de fracaso” estableciendo una nota de corte.

La prueba (IM-OSCE) fue aplicada anualmente a residentes entre su primer y cuarto año de posgrado en donde se les exponían de entre 9-10 casos que se enfocaban en la evaluación de distintas habilidades clínicas. Los residentes fueron asesorados por un médico evaluador. El tipo de retroalimentación que recibieron fue verbal y numérica (detalles del contenido y el formato de la retroalimentación numérica han sido previamente publicados).

La Prueba (RCPSC IM) está compuesta de dos elementos: prueba de selección múltiple (200 preguntas) y una prueba de desempeño de habilidades clínicas (8 estaciones de 15 min. evaluando 14 casos). Los casos eran de 8 tipos: ejercicios orales, chequeo físico, comunicación y ética, revisión física simulada (usando un simulador cardiopulmonar).

Con respecto a las resultados el análisis correlacional entre IM-OSCE y notas de RCPSC IM (incluyendo la parte escrita y la parte de desempeño) fueron positiva y moderadamente altas. El análisis de regresión, conforme las notas de IM-OSCE incrementaban, el riesgo de fallo en cualquier componente de RCPSC disminuía. Por ejemplo, por cada aumento de 10 puntos en la IM-OSCE, la posibilidad de fallo en la prueba escrita de RCPSC de un residente de cuarto año disminuía por un 10.5%. Como se

indica en los radios de posibilidad, conforme los puntajes de IM-OSCE (Examen Clínico de Estructuración Objetiva) incrementan, el riesgo de fallo en la prueba RCPSC IM descendía.

26. **Eva, K. & Macala, C. (2014). Multiple mini-interview test characteristics: 'tis better to ask candidates to recall than to imagine. Medical Eduacation, 48, 604-613. doi:10.1111/medu.12402**

El objetivo del estudio fue investigar las características de la Entrevista Corta Múltiple (MMI) cuando el tipo de estación es manipulada. La población de estudio fue de 41 aplicantes a la escuela de medicina después de haber asistido a la entrevista de admisión a la Universidad de British Columbia (UBC). Se establecieron 4 circuitos de 12 estaciones. Grupo de evaluadores (48) compuesto por profesores de la Escuela de Medicina, miembros de la comunidad y estudiantes actuales de medicina. A los candidatos se les dio 2 minutos para leer la primera estación. 7 minutos después sonaba una alarma para indicar que la entrevista había concluido. Pausas de 3 minutos entre estaciones de donde un minuto era para completar una encuesta y 2 para relajarse. Evaluadores también completaron sus hojas de evaluación durante este lapso.

En las estaciones de juicio de situación (SJ) se les pedía a los participantes que imaginaran cómo reaccionarían ante las situaciones planteadas. Evaluadores recibieron una hoja con 6 preguntas, pero las debían abarcar a modo de diálogo y no de entrevista; recibieron una hoja con información de referencia respecto a las competencias para CanMeds relevantes a la situación descrita. También recibieron una escala de evaluación para valorar al candidato en habilidades de comunicación, habilidades de razonamiento, y profesionalismo.

Por otro lado, en las estaciones de entrevista de comportamiento (BI) se les pedía a los candidatos que pensaran en una situación que ellos o ellas hubieran experimentado y que fuera análoga con la situación presentada en la estación, y debían discutirlo con los examinadores.

En las estaciones de estilo libre (FF) se les permitió a los evaluadores elaborar cualquier pregunta que les ayudara a generar un puntaje con respecto al rol profesional del candidato. Finalmente se utilizó estadísticas descriptivas y análisis de varianza (ANOVA) para comparar el desempeño de los candidatos y las respuestas a las encuestas tanto de los candidatos como de los evaluadores.

En esta investigación los resultados muestran que las notas alcanzadas en el examen de admisión MMI para aquellos que participaron en el estudio (nota promedio: 78.9) y quienes no participaron (nota promedio: 79.1) fue bastante similar ($F(1,617) < 1$, $p > 0.5$). En cuanto al desempeño de los candidatos no se presentó mucha diferencia entre estaciones.

La estación de comportamiento fue la que presentó mayor diferencia al resto. Además, las opiniones de los participantes en las estaciones de tipo libre se caracterizaron por provocarles más ansiedad, menos claridad en sus respuestas, y más difíciles que el resto de las estaciones. Para los interrogadores no hubo diferencia en estos aspectos.

La correlación entre usar el puntaje promedio usando estaciones BI y el puntaje promedio usando estaciones SJ fue $r = 0.65$. La correlación entre el puntaje promedio usando estaciones FF y el puntaje promedio usando estaciones BI o SJ fue $r = 0.34$ en ambas instancias.

En cuanto a la relación con la prueba de admisión MMI: La correlación entre el promedio de las 4 estaciones dentro de cada tipo de el promedio de las 9 estaciones MMI usadas para admisión fue: SJ, $r = 0.45$; BI, $r = 0.57$, y FF, $r = 0.42$. En general los

candidatos consideraron las estaciones FF más desafiantes y las que les provocaban más ansiedad que las SJ o BI.

Es importante destacar que otros estudios han demostrado que una mini entrevista de 8 minutos es suficiente para que los entrevistadores formen su criterio respecto a una persona. Adicionalmente, el hecho de tener un set de preguntas estructuradas hace que el entrevistado no pueda ahondar en temas que no van conforme al tema principal de la entrevista.

27. Knorr, M. & Hissbach, J. (2014). Multiple mini-interviews: same concept, different approaches. Medical Education, 48, 1157-1175. doi: 10.1111/medu.12535

El objetivo de la investigación fue revisar la literatura existente sobre MMI (Entrevista Múltiple Corta) para identificar aspectos del diseño de MMI que influencian la credibilidad, validez y costo-eficacia del formato de entrevista MMI.

La búsqueda de información se condujo hacia finales de julio del 2013. Buscaron en OVID (incluyendo bases de datos PsycARTICLES, PsycINFO, y PSYNDEX) y en PubMed con ciertas palabras clave. Luego se extendió la búsqueda a “entrevista MMI” y “entrevistas MMI”. Segundo paso: se buscaron publicaciones que citaran Eva et al. (via Web of Science) quien fue la primera en explicar los principios de MMI. Se revisó rápidamente los resúmenes de cada uno de los artículos y se categorizaron en: (i) principalmente enfocados en MMI’s, (ii) reporte de hallazgos en MMI, o (iii) descripción de un enfoque intermedio que adopta formato de estación múltiple como la entrevista personal modificada (MPI).

De la información recolectada 66 publicaciones cumplieron con los criterios incluyendo dos artículos acerca de una herramienta de evaluación llamada, ver Ziv et al. el “MOR” en el cual la idea principal de MMI y los principios de evaluación centrales son

combinados. Los estudios fueron ingresados a una tabla con los detalles de diseño de MMI y con los resúmenes de los principales descubrimientos de cada publicación.

Con respecto a los resultados del estudio, se destaca que en el diseño de la MMI, el tipo de estación, los detalles del proceso de la MMI, y el sistema de puntaje final determinan el diseño de un MMI. Por lo tanto, todos estos son factores que proveen posibilidades de variación.

En cuanto a los atributos/habilidades, sólo unos pocos autores han descrito su enfoque. El enfoque hace posible identificar las características principales que se van a medir. En la literatura consultada, las listas de las principales características están entre 3 y 19 atributos. Algunos atributos son más comúnmente usados como habilidades de comunicación, y otros son más específicos de programa como potencial de liderazgo.

El número usual de estaciones está entre 6 y 12. Todas estas MMI's se aplicaron a candidatos en proceso avanzado de estudio que aplicaban para residencia o posgrados. Duración: entre 1-11 días, de hasta 7 sets por día y 7 circuitos por set. Tiempo: entre 30 segundos y 3 minutos para leer el escenario y prepararse. Duración en estación: entre 5 y 15 min. El sistema de puntaje se realiza usualmente a través de escalas Likert con puntajes de 4 a 10 o dos entrevistas en cada estación. El puntaje para la entrevista se puede componer de una sola escala o de la combinación de varias subescalas.

En este tipo de evaluación, los análisis de validez de constructo son exploratorios principalmente y sus resultados son inconclusos. En este sentido, 40 estudios reportaron valores de credibilidad. En general, el incrementar el número de estaciones tiene mayor credibilidad que incrementando el número de entrevistadores por estación. Los factores más relevantes con respecto a costos comparados con entrevistas convencionales son los costos de la implementación de la estación y pago a los actores.

28. **Gafni, N., Moshinsky, A., Eisenberg, O., Zeigler, D. & Ziv, A. (2012). Reliability estimates: behavioural stations and questionnaires in medical school admissions. Medical Education, 46, 277-288. doi:10.1111/j.1365-2923.2011.04155.x**

EL objetivo de la investigación es proveer estimaciones confiables para la MMI (Entrevista Múltiple Corta) que incluyan fuentes de variación como ocasión y diseño así como también coeficientes G para 16 o 17 estaciones de conducta. Además, se espera estimar los coeficientes G dentro de cada centro de evaluación, y de las estaciones combinadas con los dos cuestionarios.

La población invitada a participar en el estudio fueron aquellos candidatos que aplicaron a programas de 6 años para al menos una de las 4 instituciones (3 escuelas de medicina y una de odontología). El total de participantes fue de 2662 participantes en MOR y 2023 en MIRKAM.

El Instituto Nacional de Pruebas y Evaluación de Israel desarrolló y administró dos centros de evaluación para escuelas de medicina y odontología. Los centros son: MOR y MIRKAM.

MOR (estaciones de simulación) tiene 9 estaciones individuales (6-9 min c/u) y 2 estaciones grupales (30 min c/u). 3 estaciones de simulación son de interacción entre candidato y SP (paciente estándar); 2 estaciones de información donde se les aplicó una entrevista acerca de su desempeño en la estación de interacción con paciente estándar; una entrevista personal estandarizada sobre lo que piensa el candidato acerca de la profesión médica y los actuales problemas en políticas médicas.

En cada una de las estaciones grupales, se requirió un grupo de 6 miembros donde los candidatos se enfrentaron a situaciones interpersonales e intra-grupales. Para la evaluación de los candidatos se utilizaron formularios de evaluación estructurados donde en algunos casos incluían escalas de 1-6 y algunas otras veces eran evaluados por un profesor o por un SP. Participan un total de 348-386 profesores y 33-35 SP´s cada año.

MIRKAM tiene 8 estaciones de entrevistas. Desde el 2006, los candidatos que participen en las dos evaluaciones completan un cuestionario una vez al año y sus puntajes son utilizados en ambos sistemas. MIRKAM: Mini Entrevistas múltiples: 10 minutos cada una. Hubo 3 estaciones de semi simulación donde se daba un role-play entre entrevistador y candidato.

En dos estaciones se le presentaba al candidato un dilema médico ético a discutir. En 3 estaciones, se le preguntaba al candidato sobre su historia biográfica. Para cada estación se los evaluó por un profesor usando métodos similares a MOR.

Se aplicó un Cuestionario de juicio y toma de decisiones (JDQ) que consistía en 3 pruebas a forma de ensayo donde los candidatos describían dilemas éticos de la vida real. 45 minutos de tiempo. Los argumentos dados por los candidatos se contaban, y cada argumento recibía de 0-2 puntos basándose en cómo se relacionaba a los valores generales, profesionalismo, y consideraciones morales. Se sacó un promedio ponderado de los ensayos.

También se aplicó un Cuestionario biográfico (BQ) que consistió de 20 preguntas abiertas separadas en dos secciones. La primera sección se enfocó en la experiencia de vida del candidato. La segunda en la concientización emocional del candidato. Tiempo total: 95-120 min. Cada pregunta se evaluaba en una escala de 1-5. El rango de coeficientes G a lo largo de los 4 cohortes fue 0.64-0.71.

Se calcularon coeficientes de generalidad para las estaciones cada año, coeficientes de confiabilidad para todos los centros en general, coeficientes para aquellos que volvían a tomar la prueba, y la correlación de coeficientes entre los centros.

Con respecto a los resultados, la generalizabilidad de las estaciones de comportamiento para todas las pruebas fue de 0.73 para las estaciones de comportamiento de MOR. Los coeficientes G son similares al rango reportado por MMI de 0.65 - 0.81. En general las estimaciones tienden a ser más altas para las estaciones de comportamiento MOR que para las estaciones de comportamiento MIRKAM.

La confiabilidad de todo el centro de prueba: Ambos MOR y MIRKAM comprenden estaciones y dos cuestionarios. La confiabilidad se obtuvo de calcular un puntaje compuesto. Tal y como se esperaba, al aumentar el número de estaciones a 14, se encontró que estaba asociado con un incremento en la confiabilidad de un 0.67 (MIRKAM) o 0.76/0.69 (MOR) a un 0.81/0.77 a lo largo de los años.

La correlación entre MOR y MIRKAM: Las correlaciones de los dos cuestionarios en las dos estaciones diferentes son similares (0.28 y 0.23 entre JDQ y las estaciones MOR y MIRKAM respectivamente; 0.51 y 0.47 entre BQ y las estaciones de comportamiento de MOR y MIRKAM, respectivamente). El coeficiente de la correlación Pearson entre MOR y MIRKAM a lo largo de los 4 años fue de 0.56.

29. **Tetzlaff, J., Dannefer, E. & Fishleder, A. (2009). Competency-based assessment in a medical school: A natural transition to graduate medical education. Journal of Education Research, 2(4), 241-255.**

El objetivo de la investigación fue describir el diseño e implementación de un portafolio de evaluación basada en competencias del Cleveland Clinic Lerner College of Medicine (CCLCM) así como abordar el enfoque de portafolio y los desafíos de implementación más generalmente.

La población de estudio fue estudiantes o residentes de medicina del Cleveland Clinic Lerner College of Medicine. Se definieron las competencias principales que se esperaba que los estudiantes mostraran durante su evaluación. Las competencias son: investigación; conocimiento médico en lo básico, clínico y ciencias sociales; comunicación; profesionalismo; desarrollo personal; habilidades clínicas; razonamiento clínico; sistemas de salud; y práctica reflexiva.

Las y los estudiantes deben desarrollar planes de aprendizaje de acuerdo a sus fortalezas y debilidades. Al final del primer y segundo año, un comité evalúa los portafolios. En el 3 y 4 año, los estudiantes desarrollan portafolios formativos y uno sumativo en el 5to. Estudiantes de medicina y estudiantes haciendo su internamiento recibieron un portafolio en donde se les instaba a dominar y evaluar su desempeño en 9 competencias relacionados al profesionalismo de un estudiante de medicina. Por medio de ensayos, los estudiantes proporcionaban su propia retro alimentación así como a través de evidencia que ellos seleccionarán de su base de su hoja de evaluación. Todo lo anterior supervisado por un profesor consejero.

Evaluación de competencia en bachillerato de medicina: Bajo la supervisión de un médico tutor (PA), los estudiantes desarrollaron 3 portafolios formativos en su primer año y 2 en su segundo año. Se espera que los estudiantes documenten su progreso de competencia a través de la escritura de ensayos reflectivos acerca de sus fortalezas y debilidades.

Según los resultados los residentes fueron capaces de llegar al mismo nivel de evaluación técnica que sus profesores con un modesto entrenamiento, especialmente si el entrenamiento incluía expectativas específicas. La ventaja agregada de este proceso es el aprendizaje adicional del acto de la auto-evaluación.

Para efectos generales del contexto de la práctica, reflexionar acerca de casos desafiantes combinados con mantener datos en una libreta y la retroalimentación de un tercero mejoró las habilidades de auto-evaluación. El sistema de portafolios enfocado en la evaluación de competencias propias mostró ser efectivo para que los estudiantes auto reflexionen y asuman un sentido de responsabilidad de su propio aprendizaje. Además de que este sistema crea una transición natural para el momento de hacer la residencia.

El portafolio no tenía nota académica para evitar la competitividad entre estudiantes de modo que se enfocaran en el desarrollo de las habilidades. Este sistema de evaluación permite detectar problemas de conducta a un periodo temprano.

30. **Wamsley, M., Julian, K., O'Sullivan, P. & Satterfield, J. (s.f.). Designing Standardized Patient Assessments to Measure SBIRT Skills for Residents: A Literature Review and Case Study.**

El objetivo de investigación fue describir el desarrollo de una evaluación estandarizada de pacientes para medir SBIRT skills (habilidades en detección, intervención breve y remisión a tratamiento, por sus siglas en inglés), la percepción de los residentes respecto a dicho ejercicio de actividad, y la confianza en SBIRT skills.

La población de estudio fue de 15 residentes de atención primaria de medicina interna quienes participaron en la evaluación pre-curriculum SBIRT SP y 12 residentes participaron en la evaluación post-curriculum SBIRT SP. Para la pre evaluación, 66.7% de los participantes eran mujeres y 53.7% estaban en su tercer año de residencia. Para la post-evaluación, 83% fueron mujeres y 33.3% estaban en su tercer año de residencia.

El desarrollo modelo de evaluación SP debía incluir mínimo 10 casos en el lapso de 3-4 horas para alcanzar una nota con confiabilidad de 0.85 a 0.90. Los casos incluían una serie de sustancias, comorbilidades de salud médica y mental, en edades de pacientes.

Para las medidas de desempeño, se utilizaron checklists que contenían entre 28 a 33 ítems, incluyendo de 24 a 27 ítems que se enfocaban en habilidades SBIRT clave, la mayoría de los ítems eran de "sí" o "no." Todos los casos incluían una escala de clasificación de 6 puntos (máximo) de satisfacción del paciente. Ejercicios posteriores al caso: utilizado inmediatamente después del encuentro SP. Se creó una rúbrica para referirse a los puntos que debieron haber sido incluidos en las respuestas escritas de los residentes.

Con respecto a la satisfacción del residente, estos debían contestar una encuesta electrónica anónima, la cual incluía 7 ítems respecto a sus impresiones de la evaluación SP, aunado a 2 evaluaciones generales respecto a su experiencia. Acá los ítems fueron evaluados por escalas Likert de 1 (totalmente en desacuerdo) a 5 (totalmente de acuerdo), además incluían 2 preguntas abiertas acerca de las fortalezas y debilidades de usar SP como una herramienta de aprendizaje para SBIRT. Por otro lado, para evaluar la auto eficacia del residente, estos completaron una encuesta anónima donde ellos mismos evaluaban su propio nivel de confianza en una escala Likert con 4 ítems.

En línea con lo descrito, para el análisis estadístico, se calcularon puntajes promedio para cada ítem de la satisfacción del residente y de las encuestas de autoeficacia para la participación en evaluación SP antes del plan de estudios (rephrase) y posterior al post-curriculum.

Los participantes SP (pacientes estándar) recibieron un total de 7 horas de entrenamiento donde practicaron por medio de role-plays usando los checklists para asegurar confiabilidad y precisión en el estudio.

En relación al procedimiento de Evaluación SP, los residentes participaron en una evaluación SBIRT SP antes y después de completar 10 horas de curriculum SBIRT, que trataba sobre bebidas peligrosas y abuso de sustancias. Las evaluaciones SP pre y post estuvieron separadas por un periodo de 10 meses. Para cada evaluación SP, los residentes completaron una sesión de 1.5 hrs en la cual se desempeñaban en 3 casos SP. Cada SP duró 20 minutos, seguido de un ejercicio post caso de 10 minutos. Los residentes completaron una encuesta de satisfacción después del último ejercicio post-caso y antes de iniciar la sesión de revelación del objetivo del estudio. El equipo experto de evaluación SBIRT SP asistió a la evaluación SP y monitorearon el desempeño SP de los residentes remotamente usando monitores de video.

Los resultados arrojan que los residentes demostraron grandes mejoras en conocimiento SBIRT, toma de historial, detección de abuso de sustancias, y diagnóstico

SUD (desórdenes de abuso de sustancias), desde la evaluación pre hasta la evaluación post-curriculum.

Las y los participantes reportaron que su experiencia SP fue similar a lo que habían visto usualmente en las clínicas y que además tuvieron tiempo suficiente para los encuentros SP y ejercicios post-caso. Comentaron que uno de los puntos fuertes de la evaluación pre-curriculum se enfocó en la posibilidad de practicar sus habilidades en un ambiente seguro con casos realistas. Otros comentaron que no les gustó la evaluación porque se sentían artificiales. Los residentes reportaron que se sentían más seguros detectando desórdenes de abuso de alcohol y menos seguros en el desarrollo de tratamientos para pacientes con abuso de sustancias.

31. Alarcon. A. (2013). Incorporación del Examen Clínico Objetivo Estructurado (ECOE) en la Carrera de Enfermería. Rev Educ Cienc Salud. 10 (1): 18-22

El objetivo del estudio fue analizar la percepción de alumnos de enfermería en la implementación de ECOE en la asignatura Enfermería en Adulto y Senescente, Carrera de Enfermería, Universidad San Sebastián, y relacionarla con otros procedimientos evaluativos utilizados.

Se aplicó a 22 estudiantes en la asignatura Enfermería en Adulto y Senescente durante el segundo semestre en el año 2010. Para su realización se utilizaron ocho estaciones donde el estudiante disponía de siete minutos para resolverlas y una vez que las resolvieran se les solicitó su opinión por medio de una encuesta sobre la metodología empleada

Entre los resultados obtenidos como aporta Alarcón (2013) se concluye de que los estudiantes perciben el ECOE como una metodología que evalúa contenidos teóricos, destrezas y actitudes; “un 40,91% estimó que el tiempo era insuficiente en algunas estaciones; un 31,82% consideró al método como una oportunidad de aprendizaje e igual porcentaje lo percibió estresante tan sólo al comienzo por no tener experiencia previa”.

Hubo una relación significativa entre la evaluación de procedimientos en clínica con las estaciones que se trabajaron, se analizó la relación entre la nota final del ECOE con los diferentes ítems evaluados en clínica para lo que se utilizó el coeficiente de correlación Producto Momento de Pearson en base a un contraste bilateral, en donde se encuentra relación estadísticamente significativa, “el promedio del ECOE con la primera historia clínica y la evaluación de procedimiento ($p < 0,05$) y entre promedio del ECOE y test ($p < 0,01$)”.

Alarcón (2013) en su estudio reafirma la importancia que tiene la retroalimentación de los estudiantes que realizan las pruebas con respecto al ECOE, pues esto permite tener un panorama cercano a la realidad clínica que van a realizar en sus prácticas profesionales.

Entre estas retroalimentaciones se puede rescatar por ejemplo lo que mencionan los estudiantes sobre el poco tiempo para realizar las evaluaciones siendo entonces uno de los puntos que se pueden considerar para futuras aplicaciones, siguiendo esta misma línea además es importante tener en cuenta el nivel del plan de estudios que están cursando los estudiantes.

32. Díaz-Plasencia, J., Moreno-Castillo, P., Calmet-Ipince, J., Yan-Quiroz, E., Díaz-Villazón, M., Iglesias-Obando, A., Zegarra-Castillo, K., & Urquiaga-Ríos, K. (2016b). Validez concurrente del examen clínico objetivo estructurado con el portafolio electrónico, examen teórico y promedio ponderado en estudiantes de cirugía de la Universidad Privada Antenor Orrego. FEM, 19(5), 237-245.

El objetivo del estudio fue demostrar la validez concurrente del examen clínico objetivo estructurado (ECOE) con el promedio ponderado, la nota teórica y el portafolio electrónico en 123 (71 fueron mujeres y 52 varones.) estudiantes de medicina del curso de Cirugía I del IX ciclo, durante el semestre académico 2014-I, de la Facultad de Medicina de la Universidad Privada Antenor Orrego (Trujillo, Perú).

Se realizan cuatro ECOE de 18 estaciones cada uno durante el curso, dos para la rotación de especialidades quirúrgicas y otros dos para cirugía general y abdominal. Se programan las tareas que deben realizar los alumnos en las estaciones correspondientes, incluyendo habilidades clínicas básicas: anamnesis y exploración física, habilidad para obtener información, habilidades técnicas, manejo de las situaciones (diagnósticas, terapéuticas y de seguimiento), habilidades preventivas, comunicación y trato con el paciente, y atención a la familia.

Un día antes del examen se realiza una charla de inducción. La duración por estación fue de 5 minutos. Cada docente aplica una lista de cotejo o escala de verificación, validada previamente, para cada una de las estaciones programadas. Al final se realiza la retroalimentación correspondiente. Además, se solicitó a cada estudiante de manera individual, al inicio de cada capítulo, que introdujera información en un portafolio semiestructurado diseñado en forma electrónica (ponderación total: 15%), cuya construcción y entrega de actividades se realizaron al término de cada rotación en la plataforma Moodle de código abierto. El portafolio consta de los siguientes campos: actividades registrales, autoevaluación personal, práctica reflexiva, y estructura y lenguaje escrito.

Con respecto a los resultados, hubo correlación bivariada aceptable (r = 0,65) entre la nota teórica y el ECOE; correlación moderada (r = 0,52) entre el promedio ponderado y el ECOE; y correlación alta (r = 0,77) entre la nota del portafolio electrónico y el ECOE. Hubo correlación lineal múltiple entre el portafolio, el ECOE y el examen teórico (coeficiente de determinación múltiple R^2 = 0,55).

El promedio de la nota teórica, el ECOE, el portafolio y el promedio ponderado de la serie global fue de 10,65 ± 0,83, 13,28 ± 0,53, 11,88 ± 0,74 y 11,67 ± 0,93, respectivamente (p = 0,0001). Hubo correlación aceptable (r = 0,65) entre la nota teórica y el ECOE estructurado. Hubo correlación moderada (r = 0,52) entre el promedio ponderado previo y el ECOE (Fig. 2). Hubo correlación alta (r = 0,77) entre las notas del portafolio y el ECOE.

33. Donato, A., Pangaro, L., Smith, C., Rencic, J., Diaz, Y., Mensinger, J., & Holmboe, E. (2008). Evaluation of a novel assessment form for observing medical residents: a randomised controlled trial. Medical Education. 42, 1234-1242. doi:10.1111/j.1365-2923.2008.03230.x

El objetivo de la investigación fue evaluar si un nuevo formulario de evaluación puede mejorar la precisión de miembros de facultad en la detección de desempeños insatisfactorios, generar más observaciones del evaluador y mejorar la calidad de la retroalimentación.

La población de estudio fue de 80 profesores de medicina interna de 4 programas de residencia (2 comunitarios y 2 basados en universidad-hospital). La participación en este proceso fue remunerada con un honorario de $100.

Cada participante perteneció al grupo control o al grupo de intervención, el primero de estos tenía un formulario que incluía 7 dominios: intervención médica, habilidades de chequeo físico, cualidades humanísticas, juicio clínico, habilidades de asesoramiento, organización, y competencia clínica en general. Este formulario tenía poco espacio para comentarios y las evaluaciones tienen términos como “insatisfactorio” (1-3), “satisfactorio” (4-6), y “superior” (7-9). Por otro lado, los formularios de evaluación para los grupos de intervención, utilizaban una tarjeta llamada Minicard, que tiene 4 secciones (2 atrás y 2 por el frente) con los títulos “historia”, “chequeo”, “presentación de plan”, y “asesoramiento”. Cada dominio tenía 11 sugerencias. Había 4 categorías evaluativas: excelente, bueno, marginal, pobre. También había espacio libre para comentarios al final de cada dominio y en la parte superior del formulario había una sugerencia de plan de acción.

Para el desarrollo de la Minicard se hizo una compilación de sugerencias encontradas en literatura que representaban habilidades y comportamientos clave en médicos. Cada sección tenía sugerencias en: habilidades interpersonales/comunicativas,

conocimiento médico, y profesionalismo. Tenía escala de puntaje de 4 puntos (2 satisfactorios -excelente y bueno- y 2 insatisfactorios-marginal y pobre) para forzar a los evaluadores a dar una de las dos categorías.

Para llevar a cabo el estudio, se realizó un entrenamiento, donde cada grupo de estudio recibió una hora de entrenamiento el cual consistía en la muestra de un video descriptivo de 9 minutos (ABIM Mini-CEX) u 11 minutos (Minicard), seguido de una práctica. En la práctica se les mostraban 3 casos de entrevistas a estudiantes y luego debían discutir al respecto. Grupos de control y de intervención fueron entrenados por separado. Participantes completaron un cuestionario previo y se les pidió practicar usando la herramienta de evaluación de entre 2-3 semanas.

Después de 2-3 semanas las personas participantes se les pidió evaluar 6 videos mostrando situaciones de interacción entre estudiantes de segundo año y pacientes usualmente observadas en medicina interna. Se dispuso de 3 minutos para generar la nota de evaluación y responder la pregunta “¿Qué retroalimentación le daría a este residente?” No se permitía conversación entre participantes.

La codificación de la evaluación, fue por medio de un set distinto de evaluaciones, se entrenó a los evaluadores en comentarios uniformes y se logró un 90% de consenso para así alcanzar la uniformidad de comentarios y de calidad.

Para el análisis estadístico, se efectuó un análisis de factor dentro de los sujetos para determinar diferencias entre los dos grupos (intervención vs control), entre observaciones totales, y entre retroalimentación de observaciones. El hospital de donde provenían los médicos se introdujo como una covarianza entre sujetos. La calidad de la retroalimentación se codificó descriptivamente al calcular la proporción del total de observaciones y la retroalimentación atribuida a cada tipo codificado (mínimo, observacional u orientado a acción). La exactitud de los puntajes se calculó al recolectar las notas más bajas dadas por los participantes en cualquier dominio dicotómico de tipo “pasa/falla” y se comparó eso con la evaluación de “pasa/falla” de expertos. Puntajes correctos

totales se tabularon como puntajes brutos, y luego se ajustaron para las covarianzas con regresión logística.

Los resultados destacan que los participantes que completaron la Mini-card, clasificaron desempeños correctamente 85% de las veces, comparado con un 73% para quienes utilizaron el formulario ABIM Mini-CEX. Sin embargo, quienes usaron Minicard estuvieron más propensos a errores severos en general, a calificar correctamente desempeños satisfactorios un 73%, y fueron menos precisos (58%) al evaluar correctamente al residente que efectuó un chequeo físico completo satisfactoriamente.

Los usuarios de la Minicard identificaron correctamente 96% de los desempeños insatisfactorios (rango 93-98%) al demostrar errores médicos en conocimiento, profesionalismo y habilidades interpersonales, al seleccionar la categoría más baja de evaluación (“pobre”) 55% de las veces. Mientras que los usuarios ABIM Mini-CEX identificaron consistentemente residentes que pasaron (95% identificados correctamente), pero tuvieron más error de clemencia al identificar desempeños insatisfactorios solamente un 52% de las veces (rango 33-63%).

Los participantes del grupo de control registraron un ítem más de retroalimentación que el grupo de intervención ($F [1.72] = 7.87$; $P = 0.006$). Los participantes del grupo de intervención, sin embargo, tuvieron casi el doble del total de observaciones ($F [1.72] = 62.11$; $P < 0.001$) con un gran efecto (Cohen´s $d = 1.76$).

La concordancia intra evaluadores, determinada por Fleiss’ kappa, fue de 0.299 ($z = 4.69$, $P < 0.001$) para quienes usaron el formulario ABIM Mini-CEX, lo que sugiere baja concordancia, y 0.520 ($z = 15.65$, $P < 0.001$) para la Minicard, lo que sugiere concordancia moderada.

La Mini-card mejoró la exactitud general (85% vs 73%), especialmente para desempeños subestándares (96% vs 52%), incrementó el total de observaciones (10.8 vs 5.7) e incrementó la confiabilidad a nivel de pasar/quedarse (0.520 vs 0.299) comparado con el formulario ABIM Mini-CEX.

EL grupo de intervención obtuvo resultados más precisos distinguiendo entre desempeños satisfactorios de insatisfactorios, pero menos precisos al identificar desempeños satisfactorios. Además este grupo generó más observaciones escritas que el grupo de control. Esta evaluación incrementa la concordancia de resultados entre los evaluadores y la precisión de las evaluaciones en general, pero también puede causar un incremento en error de gravedad.

34. Díaz-Plasencia, J., Sánchez de Cáceda, E, Guzmán-Gavidia, C., Valencia-Mariñas, H., García-Cabrera, J., Yan-Quiroz, E., & Díaz-Villazón, M. (2016a). Fiabilidad y validez de un portafolio reflexivo en la evaluación de la práctica clínica de los estudiantes del capítulo de Cirugía Oncológica del curso de Cirugía. FEM, 19(4), 175-185.

El objetivo del estudio fue determinar la correlación bivariada entre las puntuaciones del examen teórico, la práctica clínica, el aprendizaje virtual y el examen clínico objetivo estructurado (ECOE) de la serie total con el portafolio; la ecuación de regresión para predecir la puntuación del portafolio a partir del examen teórico y ECOE de Cirugía Oncológica; la correlación entre el aprendizaje autorreflexivo del portafolio con la nota final del curso; las puntuaciones del portafolio con el test Inventory y la fiabilidad interevaluador de la estructura del portafolio y de su nota global.

Los participantes fueron 117 estudiantes del IX ciclo del capítulo de Cirugía Oncológica del curso de Cirugía I de la Facultad de Medicina de la UPAO en 2011.

A cada estudiante se le solicitó la elaboración de un portafolio de evidencia de práctica clínica. Al inicio del capítulo un portafolio semiestructurado, en función de la competencia por demostrar, en donde se definieron previamente las tareas clínicas y los apartados que se debían realizar, y cada alumno decidió qué padecimiento abordar y qué fuentes bibliográficas consultaría para sustentar la toma de decisiones clínicas realizadas. Este portafolio se estructuró en tres partes: actividades registrales que

acreditaron las habilidades trabajadas por el alumno y el nivel de profundidad con que las trabajó, actividades realizadas para la planificación (autoevaluación) y reflexión.

La significación estadística se definió como un valor $p < 0,05$. Para determinar la validez empírica se utilizó el coeficiente de correlación de Pearson. La fiabilidad de los instrumentos de evaluación del portafolio que usaron escalas de Likert se analizó mediante el coeficiente α de Cronbach.

Los resultados muestran que los promedios del examen teórico, nota de práctica clínica, paciente virtual, portafolio y ECOE de la serie total fueron: 9,20 ± 1,33; 13,86 ± 2,72; 13,73 ± 1,17; 16,07 ± 2,30 y 12,95 ± 1,43, respectivamente. Estas diferencias fueron estadísticamente significativas ($p = 0,0001$). Se evidenció que el portafolio se correlacionó directamente con la nota teórica, la práctica clínica, el aprendizaje virtual y el ECOE obtenidos al final del curso de Cirugía I, lo cual muestra que esta estrategia metodológica de evaluación tiene validez concurrente con otros formatos de evaluación que consideran aspectos cognitivos y procedimentales

Hubo correlación significativa entre el portafolio con el examen teórico ($r = 0,410$; $p = 0,0001$), la práctica clínica ($r = 0,258$; $p = 0,003$). Los promedios del examen teórico, caso clínico virtual, ECOE y portafolio del capítulo de Cirugía Oncológica fueron: 9,71 ± 2,65; 14,87 ± 1,43; 13,75 ± 1,35 y 16,07 ± 2,30, respectivamente ($p = 0,0001$). El examen teórico ($r = 0,38$; $p = 0,0001$) y el ECOE ($r = 0,33$; $p = 0,0001$) se correlacionaron con el portafolio; no fue así con el caso clínico virtual ($r = 0,13$; $p = 0,122$).

Hubo correlación significativa entre el aprendizaje autorreflexivo ($r = 0,305$; $p = 0,0001$) y la nota teórica final del curso. La fiabilidad interevaluador del portafolio fue significativa en: caso clínico real ($\alpha = 0,486$; $p = 0,006$), incidente crítico ($\alpha = 0,702$; $p = 0,0001$), aprendizaje autorreflexivo ($\alpha = 0,664$; $p = 0,0001$) y estructura del lenguaje ($\alpha = 0,431$; $p = 0,017$).

35. **Malhotra, S., Hatala, R., & Courneya, C. (2008). Internal medicine residents' perceptions of the Mini-Clinical evaluation exercise. Medical Teacher, 30, 414-419.**

El objetivo de la investigación fue evaluar la percepción de los residentes respecto al mini-CEX usando métodos cualitativos. La población de estudio fueron: estudiantes matriculados en el programa de entrenamiento de residencia de medicina interna de la University of British Columbia (UBC) en Vancouver, Canadá. 12 residentes participaron en el estudio; 5 de PGY-1er año y 7 de PGY-2do año.

Con respecto a la metodología, implementaron el Mini-CEX. Al momento del estudio, se aplicaba una evaluación acumulativa a los residentes por parte de Medicina Interna de UBC. Esto anualmente, y consistía en una prueba de selección múltiple y reportes dentro de la sesión de entrenamiento (ITER's).

Entre abril y junio del 2005, los directores de la sección de medicina interna de la UBC introdujeron un proceso evaluativo estandarizado de observación directa al implementar el mini-CEX como plan piloto en las dos unidades clínicas de enseñanza. Antes de tomar la prueba se les brindó instrucciones a los residentes sobre cómo completar el mini-CEX.

En julio de 2005 la implementación completa del programa inició esperando que cada residente completara un mini-CEX durante cada rotación clínica académica con un mínimo de 6 completadas por año académico.

En agosto de 2005 se invitó a residentes a ser voluntarios para una entrevista uno-a-uno. La pregunta inicial fue: ¿Cuál fue su experiencia al tomar el mini-CEX?" Luego siguieron preguntas abiertas; las respuestas fueron grabadas y transcritas para luego ser analizadas y de ahí formular las preguntas del grupo focal. Otra técnica utilizada fue la de grupo focal, en la cual se invitó voluntariamente para participar en el grupo focal (1 hora) y discutir sobre la "percepción general de residentes de medicina sobre el mini-CEX". Respuestas fueron grabadas en audio y transcritas verbatim (literalmente). Hubo un moderador, quien utilizó un proceso de entrevista semi conducida.

En el análisis de información se utilizó un enfoque fenomenológico para analizar los audios de las discusiones y las notas de campo del moderador.

De acuerdo con los resultados, emergieron 3 temas principales de los grupos de enfoque: la evaluación, educación, y preparación para el examen RCPSC IM (prueba comprensiva objetiva en medicina interna es una prueba a nivel nacional de especialidad que incluye 200 preguntas de selección múltiple y una prueba oral para evaluar competencia clínica). En el área de evaluación, la percepción de los residentes fue influenciada/afectada por ansiedad, además de que dijeron que el sentirse observados pudo haber cambiado su desempeño académico, por otro lado, en el ámbito de educación, los residentes expresaron la importancia del realismo en su experiencia educacional con el mini-CEX. Además, valoraron su exposición con la Facultad de Medicina y el tiempo para revisar sus habilidades de comunicación y trato con el paciente.

En el análisis del grupo focal, se muestra que entre más veces se repetía la experiencia mini-CEX, los residentes se sentían más confiados y tranquilos y a su vez le sacaban mayor provecho a los beneficios educacionales de la prueba. La preparación para el examen RCPSC IM. A pesar de que los residentes no habían tomado la prueba RCPSC IM, ellos sintieron la similitud entre la prueba a nivel nacional y el mini-CEX, lo cual les ayudaría a prepararse para este examen de alto rendimiento.

36. Tiller, D., O'Mara, D., Rothnie, I., Dunn, S., Lee, L., & Roberts, C. (2013). Internet-based multiple mini-interviews for candidate selection for graduate entry programmes. Medical Education, 47, 801-810. doi: 10.1111/medu.12224

El objetivo de la investigación fue determinar si los resultados derivados del proceso iMMI (Mini entrevista múltiple a través de internet) eran equivalentes e igual de confiables que los resultados de la entrevista en persona MMI (Mini entrevista múltiple regular). Segundo, describir la factibilidad, aceptabilidad y costos de efectividad del iMMI.

Se analizó información de 4 fuentes; primero, a partir del 2009 se evaluó en persona tanto estudiantes locales como internacionales. Segundo, a partir del 2011 se aplicó la MMI para estudiantes locales. Tercero, a partir del 2011 se aplicó la entrevista basada en internet iMMI. Finalmente, la información de las entrevistas para la aceptación del iMMI estuvo disponible en el 2011. Todos los análisis fueron conducidos usando IBM SPSS Statistics Versions 20.

En las entrevistas en persona MMI del 2009 participaron 135 estudiantes internacionales en Vancouver por 3 días y en Singapore por 1 día. Para el proceso del 2011 de entrevistas en persona MMI se aplicaron 84 preguntas a 571 candidatos locales por 196 entrevistadores. Para la entrevista iMMI, fueron 76 preguntas aplicadas por 83 entrevistadores a 293 candidatos internacionales. Los 83 entrevistadores iMMI también participaron en la MMI, representando el 42% de los 196 entrevistadores MMI en el 2011. Para ambas formas de entrevista, MMI y iMMI, se calculó un puntaje total para cada candidato sobre las 9 preguntas, dando un puntaje total para el MMI para el posterior análisis.

Para la equivalencia de notas de MMI y iMMI. Se hicieron dos análisis unidireccionales de varianza (ANOVAS). En el primer análisis se comparó los resultados para estudiantes internacionales iMMI del 2011 con los resultados de los candidatos MMI del 2009. Posteriormente, se compararon los resultados iMMI de estudiantes internacionales en el 2011 con los de estudiantes locales que tomaron la entrevista en persona MMI de ese mismo año. Se usó Eta-cuadrado para examinar la proporción de varianza asociada con los principales efectos tratados en el ANOVA.

Se estimó la varianza debido a “candidato,” “pregunta,” “entrevistador,” y “pregunta x entrevistador”. El número de entrevistadores varió de entre 1 a 15 en los circuitos en ambos formatos. Del mismo modo, se analizó la factibilidad y aceptación. Se recopiló retroalimentación del proceso iMMI por medio de entrevistas 3 semanas pos-

teriores a la terminación de la prueba. Costos fueron estimados por miembros administrativos encargados de las entrevistas de admisión internacionales de la sección de Vancouver en el 2009.

Los resultados evidencian que los puntajes de las 293 entrevistas iMMI del 2011 tuvieron una mediana más grande y mayor distribución que los resultados de los 135 estudiantes internacionales entrevistados por MMI en 2009 y que 571 candidatos locales entrevistados en 2011. La prueba de homogeneidad de varianza Levene fue significativa en el nivel $p < 0.01$, indicando que la suposición de igualdad de varianza ANOVA fue violada. No se encontró diferencia significativa entre las notas medias del 2011 entre iMMI y MMI ($p = 0.196$). El coeficiente de confiabilidad fue de 0.76 para iMMI y de 0.70 para MMI.

Hubo 119 respuestas que representan el 41%. Hubo una representación mayoritaria de encuestados de odontología y de mujeres. 1 de 10 candidatos se mostró insatisfecho con la calidad visual o audio de la iMMI. 89 (76%) de los encuestados concordó que la entrevista en línea fue una buena forma de selección, sólo 9 (8%) dijeron que no, y 19 (16%) no estaban seguros. La correlación entre haber usado Skype previamente y otros aspectos de la entrevista fue débil, lo más significativo fue la edad ($r = -0.29$) donde los aplicantes más jóvenes tenían más probabilidad de haber usado Skype previamente.

37. **Brazil, V., Ratchiffe, L., Zhang, J., & Davin, L. (2012). Mini-CEX as a workplace-based assessment tool for interns in an emergency department - Does cost outweigh value?. Medical Teacher, 34, 1017-1023. doi: 10.3109/0142159X.2012.719653**

El objetivo de la investigación fue determinar la factibilidad y valor de agregar evaluaciones mini-CEX a los procesos de evaluación existentes para un cohorte de internos en el Departamento de Emergencias (ED). El estudio fue conducido por el Departamento de Emergencias (ED) del Royal Brisbane y Hospital de Mujeres (RBWH),

que consta de 750 camillas y es hospital de enseñanza metropolitano sólo para adultos. Internos (n = 20) en rotación en el ED haciendo su 4to año de internamiento.

Se evaluaron a los sujetos en dos procesos distintos: la evaluación en sitio de trabajo que consistía de series de 4 casos mini-CEX, y evaluación en el entrenamiento hecha por supervisores clínicos usando el formulario de evaluación RMO (Resident Medical Officer)

Con relación a la evaluación del mini-CEX, cada interno fue evaluado con un paciente de resucitación, dos pacientes de agudeza moderada, y un paciente de vía rápida. Hubo 7 asesores, quienes tomaron una sesión de entrenamiento de dos horas previo a la evaluación mini-CEX. Posteriormente los internos recibieron retroalimentación: un puntaje entre 1 y 8 para cada uno de los 6 dominios que componen el instrumento de evaluación del mini-CEX. Una semana después de haber completado la evaluación, tanto internos como asesores llenaron una encuesta en línea de percepción educacional de la prueba y de factibilidad, la cual incluía escalas Likert.

La evaluación en la formación se realizó por medio de un formulario de evaluación RMO, en el cual se evaluaron los dominios especificados en el Formulario de Evaluación RMO y en la Guía de Evaluación (PMCQ 2011). Los formularios se completaron por 25 supervisores y especialistas que se reunieron 2 veces por semestre para discutir el desempeño de los internos. Para cada interno, los puntajes iban entre 1 y 5 en cada uno de los 11 dominios de desempeño del formulario. A cada interno se le dio retroalimentación cara a cara.

Se utilizó estadística descriptiva, para lo cual se extrajo temas predominantes en las respuestas escritas de los cuestionarios. Los puntajes de los internos se ingresaron a una hoja de excel. Para los puntajes mini-CEX, se calculó una media y una desviación estándar para los 4 casos tomados por los internos. Finalmente se comparó los resultados del mini-CEX y el RMO usando t-tests de muestras emparejadas.

Los resultados muestran que la satisfacción en general de internos y asesores fue alta. Los internos y asesores identificaron dificultades prácticas al organizar y conducir las evaluaciones mini-CEX en el sitio de trabajo. 17 internos (85%) tomaron todas las 4 pruebas mini-CEX, 2 (10%) tomaron 3, y 1 tomó 2 (5%). Las razones principales fueron: imposibilidad de agendar un horario entre el asesor y los internos, y ausencia no planeada del asesor.

Los participantes indicaron que la prueba mini-CEX tenía validez aparente, sin embargo, el formato de evaluación que se utilizó no abarcó todos los dominios de evaluación. La mayoría sintió que la prueba era buena pero como un entrenamiento adjunto. El punto fuerte de la evaluación fue la mejora en la observación directa como medio de transmisión de la evaluación. La mayoría de los desempeños estuvieron igual o por encima de los niveles esperados, solo un interno estuvo por debajo de lo esperado en los dominios de juicio/toma de decisiones clínicas y en habilidades de administración de tiempo para RMO. Para mini-CEX, se aplicaron 76 evaluaciones, de las cuales 74 tuvieron un resultado satisfactorio o superior.

Mini-CEX fue más efectivo en identificar los dominios en los cuales los internos tenían deficiencias. Además incluir el Mini-CEX representa una inversión monetaria significativa principalmente por el tiempo que los asesores invierten.

38. **Illesca. M, Cabezas. M, Romo. M y Díaz, P. (2012). Opinión de estudiantes de enfermería sobre El Examen Clínico Objetivo Estructurado. Ciencia y enfermeria. XVIII (1): 99-109**

El objetivo del estudio fue conocer el significado que tiene para los estudiantes de segundo año de la Carrera de Enfermería el sistema de evaluación ECOE al finalizar la práctica clínica, Módulo Enfermería Básica del Niño, Adolescente y Adulto, del cuarto nivel académico, año 2007".

Esta es una Investigación educativa desde el paradigma cualitativo a través de un estudio de caso, respaldado por la perspectiva hermenéutica. La organización del

ECOE estuvo a cargo de dos docentes; se estructuraron 13 estaciones (8 para demostración de habilidades clínicas y actitudinales y 5 orientadas al dominio cognoscitivo), además cada una con su respectiva pauta de evaluación, las que fueron elaboradas con el consenso del equipo examinador, compuesto por 10 personas.

Se menciona la importancia de que los estudiantes conozcan la metodología de lo que van a realizar es relevante para tener resultados positivos. Analizando las sugerencias de los estudiantes, aparte del desconocimiento de este tipo de evaluación a futuro, los docentes deben considerar un tiempo previo de formación para que se enfrenten al examen.

Como parte de los resultados se encuentra como desventaja de aplicar la prueba, se menciona por parte de los estudiantes el tiempo de espera para que los participantes sean examinados, el tiempo establecido en cada una de las estaciones y el momento del semestre en que se realiza.

El ECOE permite el desarrollo de habilidad mental, sistematizar procesos, identificar debilidades y fortalezas, mejorar errores. De ese mismo modo algo positivo es que se destaca la retroalimentación recibida posteriormente dada por el evaluador, hecho que si bien es cierto se produce en las experiencias clínicas, aquí en cinco minutos igual se realiza, demostrando la experticia del docente y la objetividad del mismo en el proceso.

39. **Wadee, A., & Cliff, A. (2016). Pre-admission tests of learning potential as predictors of academic success of first-year medical students. South African Journal of Higher Education, 30, 264-278. Recuperado de: http://dx.doi.org/10.20853/30-2-619**

El objetivo de la investigación fue investigar el grado de asociación entre los puntajes en componentes específicos de HSPT’s (Pruebas de nivel/ubicación en ciencias

de la salud) y las notas en cursos a medio año y pruebas finales de estudiantes de medicina de primer año.

Hubo un total de 120 estudiantes de medicina de primer año de la Universidad de Witwatersrand, Sudáfrica, durante el 2010. La mayoría (62%) fueron mujeres y de etnicidad negra (42.5%), 33% blancos, 15% hindúes, y 9.5% de color.

El estudio se basó en un contexto en donde existe desigualdad educativa entre lo público y lo privado. Una entrevista y una prueba de admisión no garantizan el logro académico futuro de estudiantes de medicina. Tomó lugar con estudiantes de primer año de medicina porque el primer año es usualmente el filtro de la carrera. Se busca determinar si estudiantes con desventaja educativa (educación pública) pudieron haberse convertido en estudiantes exitosos de medicina, esto al evaluar los mecanismos de admisión a la carrera.

Se realizó un estudio retrospectivo donde se evaluó el desempeño en los cursos para ciencias humanas y ciencias básicas en conjunto con los resultados de la prueba pre-admisión de los estudiantes admitidos en el 2010. Los 4 predictores de desempeño en la pre-admisión fueron:

- PTEE: Prueba de ubicación en Inglés con fines educativos.
- MACH: Prueba de ejecución de matemática.
- MCOM: Prueba comprensiva de matemática.
- SRT: Prueba de razonamiento científico.

Se evaluó el grado de asociación entre los puntajes en componentes específicos de HSPT's y las notas en cursos a medio año y pruebas finales de estudiantes de medicina de primer año en las materias de: física, química, biología, fundamentos de ciencias médicas y clínicas (SCMD), sociología, y psicología.

Las variables consideradas fueron el número de estudiantes, la media de cada componente de la prueba, así como también los promedios de primer año de los sujetos. La administración de bases de datos y análisis estadísticos se realizó a través del software SAS, versión 9.1. Los resultados de cada componente del índice compuesto (CI) se reportan como una media ±SD. Los promedios no ajustados fueron comparados usando el t-test o pruebas Wilcoxon-Mann Whitney según correspondiera. Un valor p de < 0.05 fue considerado estadísticamente significativo.

Con respecto a los resultados, se destaca que hombres y mujeres tuvieron un desempeño similar en las pruebas pre-admisión, con $p > 0.05$. Los estudiantes de colegios privados obtuvieron mejores notas en todas las pruebas además del SRT que sugiere que el pasado académico impacta en la admisión de estudiantes. Se notaron correlaciones fuertes y positivas entre el índice compuesto y las calificaciones obtenidas en junio y noviembre entre 0.42 y 0.79, con $p < 0.0001$.

Las calificaciones más altas a medio año y final de año fueron explicadas por mejores resultados en las pruebas MCOM (parcial $r2=0.20$ a 0.25, $p <0.0001$) y en PTEEP (parcial $r2=0.04$ a 0.25, $p<0.05$) en junio. Resultados similares se encontraron en noviembre.

La prueba pre-admisión combinada (CI) parece ser el predictor más importante para las notas de medio año y final de año para la mayoría de los sujetos. Para pruebas de junio, el CI explicó el 32% de notas en biología, 20% en química, y 32% en psicología de variables de notas ($p<0.0001$). MCOM exhibió una asociación independiente de notas de biología en noviembre (parcial $r2$ de 0.02, $p=0.03$) en conjunto con CI. Notas PTEEP fueron los únicos predictores para notas en sociología (parcial $r2=0.22$, $p<0.0001$) en junio y explicó sólo el 2% de incremento en notas de psicología ($p=0.03$).

40. **Hall, J., O'Connell, A., & Cook, J. (2017). Predictors of student productivity in biomedical graduate school applications. PLos ONE, 12(1), 1-14. doi:10.1371/journal.pone.0169121**

El objetivo de la investigación fue examinar los factores considerados por los comités en la Universidad de Carolina del Norte en Chapel Hill (UNC) al evaluar aplicantes para el Programa de Ciencias Biológicas y Biomédicas (BBSP), un programa de admisión general compuesto de 14 programas PhD en las Escuelas de Medicina, Farmacia, y Odontología de la UNC y el colegio de Ciencias y Artes de la UNC en Chapel Hill.

La población de estudio fue: 280 estudiantes de grado de la Universidad de Carolina del Norte en Chapel Hill del 2008 al 2010. 61.4% mujeres, y 22.9% de grupos raciales que son representados poco en las ciencias (Américo Africano, Hispano/Latino (a), Nativo americano, Hawaiano o de las Islas del Pacífico.

Con relación a la metodología, de la aplicación BBSP de cada estudiante se recolectó: el examen de registro de postgrado (GRE), notas de bachillerato, cartas de recomendación, y experiencia previa en investigación. También se calculó la nota de las entrevistas como un aproximado.

Con respecto a las notas GRE y GPA (Promedio Ponderado Total). GRE está dividido en 3 partes: razonamiento cuantitativo (matemática) y razonamiento verbal, y escritura, lo cual incluye escribir 2 ensayos en tiempo limitado. Cada aplicación BBSP incluyó 3 cartas de recomendación, usualmente de asesores de investigación previos. Los autores de las cartas calificaron a los estudiantes entre 1 y 5 ("Excepcional", "Sobresaliente", "Muy bueno", "Promedio", o "Por debajo del promedio"). Total de cartas: 251.

Se aplicó a 142 estudiantes 5 entrevistas de 30 minutos c/u con profesores de la escuela. Parte de la retroalimentación era la recomendación del estudiante para admisión en una escala de 1 a 5. Se cuantificó las publicaciones de cada estudiante por medio de Python.

Para los resultados de los estudiantes y análisis estadístico, se agrupó a los estudiantes conforme al número de publicaciones que tuvieran: “3+” = estudiantes con ≥ 3 publicaciones de primer autor, “1-2” = 1 o 2 publicaciones como primer autor, “0+” = no publicaciones como primer autor pero al menos como autor medio, y “0” = sin ninguna publicación.

Al momento del estudio, >85% de los participantes entre 2008-2010 ya tenían PhD o estaban en proceso de graduación. El tiempo aproximado era de 5.5 años. No hubo diferencia estadística entre los grupos con respecto a puntajes GRE, puntajes de prueba verbal GRE, prueba escrita GRE, o GPA. En las notas GRE, los hombres tuvieron puntajes más altos que mujeres. Asiáticos y blancos tuvieron puntajes más altos que los grupos minoritarios.

Un número sustancial de estudiantes con un GRE bajo fue bastante productivo, mientras que estudiantes con nota GRE casi perfecta fueron mínimamente productivos durante sus cursos de grado, lo cual pone en duda la utilidad de los puntajes GRE para admisión a los programas PhD de Biomedicina. No hubo diferencia entre la experiencia previa en investigación y la productividad de los estudiantes.

Con respecto a las cartas de recomendación, los estudiantes con puntaje “3+” tuvieron puntajes en su carta de recomendación más altos (1.60+/-0.40) que aquellos en el grupo “0+” (1.82+/-0.44). Las cartas de recomendación fueron un predictor de desempeño de publicación de papers como primer autor por parte de los estudiantes.

Luego de la entrevista, el encargado llenaba una encuesta en línea donde seleccionaba uno de los 4 puntajes para el estudiante: 1-“aceptado sin ninguna reserva”, 2-

"aceptado", 3- "aceptado si hay espacio", 4- "rechazado". Estudiantes con notas altas de la entrevista no publicaron más papers que aquellos con notas de entrevista bajas.

41. **Speyer, R., Pilz, W., Van Der Kruis, J., & Wouter-Brunings, J. (2011). Reliability and validity of student peer assessment in medical education: A systematic review. Medical Teacher, 33, e572-e585. doi: 103109/0142159x.2011.610835**

El objetivo de la investigación fue dar una visión general de todos los instrumentos o cuestionarios de evaluación usados en contextos de educación médica y todos aquellos profesionales en salud, y las características psicométricas tal cual son descritas en la literatura.

Se realizó búsqueda de literatura usando 5 bases de datos: Pubmed, Embase, ERIC, PsycInfo, y Web of Science, limitado a idiomas: Inglés, Alemán, Francés, Español y Holandés. Se excluyó nominación de compañeros o evaluación de desempeño. No se incluyó estudios que describen evaluación de desempeño de grupo que no presentaran información de desempeño individual así como tampoco estudios que presentaban diferentes niveles educacionales entre compañeros evaluadores y quienes iban a ser evaluados (por ejemplo, estudiantes de último año o a punto de graduarse versus estudiantes de primer año).

El total de resultados en la búsqueda: 2899, de los cuales 28 cumplieron con los requisitos establecidos. Todas las publicaciones incluidas, excepto una en español, estaban en inglés. 18 estudios fueron escritos en la presente década, lo cual sugiere un aumento en el interés por el tema. No se encontró cuestionarios en contextos paramédicos, casi todos los estudios se enfocaron en estudiantes, uno en evaluación de compañeros en farmacia, y otro en una combinación de estudiantes de Medicina y Odontología.

El número de participantes en todos los estudios anduvo entre 16 y 349 estudiantes; 7 estudios incluyeron poblaciones con menos de 50 sujetos: 8 estudios tuvieron entre 50 y 100 y 13 tuvo más de 100. El número medio de sujetos fue 98 (percentil 25 = 51; percentil 75 = 160).

La mayoría de los estudios se enfocaron en comportamiento profesional. Otros se enfocaron en capacidades de liderazgo, habilidades en entrevista, o desempeño en solución de problemas. El número de ítems por cuestionario varía enormemente. El más corto tuvo 2 ítems. El más largo fue un instrumento de 22 ítems.

Seis estudios no proporcionaron características psicométricas. Sólo unos cuantos estudios describieron el concepto de validez de contenido: el punto al cual el principal punto de interés fue evaluado exhaustivamente por los ítems en el cuestionario. Unos pocos estudios detallaron información sobre validez de constructo. A pesar de que no existe un criterio perfecto, usualmente, las evaluaciones de miembros de la facultad son considerados de estándar dorado. Se identificaron 22 instrumentos. La mayoría de estudios usó evaluación entre compañeros como una herramienta de evaluación principalmente.

Con respecto a la teoría presentada en el artículo se destaca la evaluación por pares, la cual se puede utilizar para estimular a estudiantes en la participación de actividades educativas y para mejorar el rendimiento del equipo o determinar el esfuerzo individual. Este tipo de evaluación promueve en los estudiantes una actitud crítica, ya que al juzgar a sus compañeros pueden obtener una idea en su propia actuación. De acuerdo con Gielen (2007), evaluación de pares tiene cinco objetivos principales: el uso de la evaluación por pares como herramienta de evaluación y herramienta de aprendizaje, la instalación de control en el entorno de aprendizaje, la preparación de estudiantes para el autocontrol y la autorregulación en toda su vida aprendizaje, y la participación activa de los estudiantes en el aula.

42. **Hillebrand, K., Leinum, C., Desai, S., Pettit, N., & Fuller P. (2015). Residency application screening tools: A survey of academic medical centers. American Society of Health-System Pharmacists, 72(1), s16-s19. doi: 10.2146/ajhp150093**

El objetivo de la investigación fue evaluar el actual uso y contenido de las herramientas de evaluación utilizadas por programas de residencia de farmacia ASHP acreditados, con una población de estudio de 105 directores de programas de residencia de hospitales UHC. Para la selección de candidatos, los factores más importantes fueron la impresión del CV y las cartas de recomendación.

UHC es un comité conformado por centros médicos académicos y hospitales afiliados de Estados Unidos. Este centro se encarga de promover y conducir resultados de investigación basados en la práctica relacionados al desarrollo profesional de la fuerza laboral en farmacia. Este mismo comité desarrolló, distribuyó, y analizó los resultados de esta encuesta.

Esta encuesta consistió en 19 preguntas, de las cuales 9 preguntas fueron sobre información demográfica así como información relacionada al crecimiento del aplicante en el programa de residencia en los años 2010-11 a 2011-12. Las otras 9 preguntas eran sobre los procesos evaluativos en las actuales instituciones de los respondentes. Al final de la encuesta había preguntas abiertas.

En enero del 2012 se envió un correo electrónico a todos los 362 directores de programas de residencias PGY1 y PGY2 de 105 hospitales UHC miembros. En el e-mail venía un enlace que los conectaba a una página segura que contenía la encuesta, la cual estaba disponible por dos semanas. Se utilizó estadística descriptiva para analizar la información recolectada.

Con respecto a los resultados se recibió respuestas de 73 sitios de programas de residencia (69.5%) de las 105 instituciones a las cuales se les envió el correo electrónico. Estas encuestas (78) representan el 21.5% de los 362 programas de residencia en farmacia.

La media de ± S.D. capacidad de pacientes en camas fue 673 ± 262. Disponibilidad de posiciones de residencia, la media ± S.D fue de 5 ± 2.9. De los 73 programas respondentes, 55 (75%) tenían algún tipo de programa de residencia PGY2, incluyendo cuidados intensivos (84%, n = 46), oncología (67%, n = 37), enfermedades infecciosas (45%, n = 25), y transplantes (29%, n =16).

Se reportaron 5675 aplicaciones para solo 355 disponibles entre 2011 y 2012. En este sentido los sitios reportaron un incremento en el número de aplicaciones para residencias en farmacia entre 2010-11 y 2011-12, pero sólo un 8% de incremento del 8% en el número de entrevistas ofrecidas a candidatos.

Con respecto a las herramientas de evaluación, el 78% (n=57) usan alguna herramienta de detección o rúbrica para seleccionar a los candidatos a entrevista. 88% (n=64) pusieron a más de una persona a revisar el documento de aplicación del sujeto (en promedio, 3 revisores).

El método más común de evaluación fue revisión individualizada de aplicación antes de reunirse como grupo a discutir la selección de algún candidato. El 63% (n=46) indicó haber cambiado su proceso de evaluación en los últimos años, y 71% (n=52) ha evaluado sus procesos de evaluación anualmente.

43. **Schripsema, N., Van Trigt, A., Borleffs, J., & Cohen-Schotanus, J. (2014). Selection and study performance: comparing three admission processes within one medical school. Medical Education, 48, 1201-1210. doi: 10.1111/medu.12537**

El objetivo de la investigación fue en primera instancia analizar si los estudiantes admitidos a la Escuela de Medicina basándose en las mejores notas previo a la universidad, proceso de selección multifacético voluntario, o lotería, respectivamente, difieren en el desempeño del estudio; segundo examinar si los estudiantes aceptados en el

proceso multifacético fueron mejores que sus compañeros rechazados, y tercero analizar si la participación en el proceso multifacético estaba relacionado al desempeño.

La población de estudio fue de 1055 estudiantes matriculados en el programa de Bachillerato en Medicina holandés de la Universidad de Groningen en el 2009, 2010, y 2011. 69% mujeres, edad promedio en primer año: 18.6 años, GPA pre-universitario promedio: 7.3.

El proceso de admisión en los Países Bajos se da por medio de tres pasos: 1er paso, estudiantes con un GPA (promedio ponderado total) previo a la universidad de ≥8 obtienen admisión inmediata. Solamente 4% de los estudiantes tienen nota ≥8, lo cual indica un rendimiento excelente. 2do paso: proceso multifacético (se miden variables como conocimiento y comportamiento). 3er paso: candidatos son aceptados por medio de una lotería.

Los participantes se dividieron en 4 grupos. (i) Admitidos inmediatamente con nota ≥8 (n = 143). (ii) Aceptados en proceso multifacético (n = 295). (iii) Admitidos por lotería que habían sido rechazados en paso previo (n = 315). (iv) Admitidos por lotería que no habían participado del proceso de selección (n =302).

El proceso de selección multifacético, consistió de 2 rondas. En la primera los participantes debían enviar un portafolio que contenía 3 secciones (educación pre-universitaria, actividades extracurriculares, y reflexión). Los 255 participantes con los puntajes más altos en la primera ronda eran invitados a participar de la segunda ronda, la cual duraba todo un día. El día se dividió en cuatro bloques que comprendían, respectivamente, una asignación de escritura, una conferencia con paciente con asignaciones posteriores, un bloque de razonamiento científico, y una serie de entrevistas breves y dramatizaciones. Se calculó un promedio de todas las 4 secciones, pero la última tenía un puntaje doble. A quienes tuvieran puntajes más altos se les ofrecía admisión a la universidad.

Las medidas de resultado, fue por medio de prueba escrita, la cual se calificó de 1 a 10. El proceso de estudio fue evaluado de acuerdo con el número de créditos aprobados en los primeros 3 años en la Escuela de Medicina. Puntajes de profesionalismo: calculado de acuerdo al porcentaje de estudiantes que recibieron el puntaje más alto en el curso de profesionalismo.

En el análisis estadístico se realizó el análisis de covarianza (ANCOVA) para evaluar las diferencias grupales en las pruebas escritas y el número de créditos aprobados en los primeros 3 años. Y se realizó regresión logística. Se excluyó a quienes abandonaron sus estudios.

Con respecto al número de estudiantes aceptados en la etapa multifacética de selección, incluyó más mujeres que el grupo de lotería admitido en el proceso de selección ($p< 0.05$). Los GPA de las top-universidades tuvieron una media más baja que los otros grupos ($p < 0.01$), mientras que el grupo admitido por lotería que no había participado del proceso de selección tuvo una nota media más alta que los otros grupos ($p <0.001$).

Las notas de la prueba escrita, difirió entre grupos ($F3,1025 = 63.20$; $p <0.001$). Grupo top de GPA pre-universitario tuvo un puntaje medio más alto que el resto de grupos (diferencia media [MD]: 1.0 - 1.3, error estándar [SE]: 0.10; $p<0.001$). El grupo aceptado en el proceso multifacético obtuvo notas más altas que el grupo de lotería admitido que no participó en el proceso (MD: 0.30, SE: 0.08).

Con respecto a los puntajes en profesionalismo, el porcentaje de estudiantes que obtuvo la posible nota más alta en el curso de profesionalismo difirió entre grupos ($X 2(3) = 21035$; $p <0.001$). Grupo top de GPA pre-universitario recibió el puntaje más alto la mayoría de las veces, seguido del grupo aceptado en el proceso multifacético.Las diferencias entre este último y los dos de lotería fue significativa. La diferencia entre el grupo top de GPA pre-universitario y el de lotería admitido rechazado en el proceso multifacético no alcanzó significación debido a que el grupo GPA top pre universitario

era pequeño. Finalmente, con respecto al criterio de abandono de universidad, no se obtuvo resultados significativos.

44. Martínez-González, A., Lifshitz-Guinzberg, A., Trejo-Mejía, J., Torruco-García, U., Fortoul-van der Goes, T., Flores-Hernández, F., Peña-Balderas, J., Sánchez-Mendiola, M. (2017). Evaluación diagnóstica y formativa de competencias en estudiantes de medicina a su ingreso al internado médico de pregrado. Gaceta Médica de México, 153, 6-15.

El objetivo de la investigación fue evaluar el nivel de competencia de los estudiantes a su ingreso al internado médico de la Licenciatura de Medicina en un nuevo currículo. Para llevarlo a cabo se eligió una muestra de 577 alumnos, los cuales tomaron el Examen Teórico.

El ECOE se aplicó a 523 estudiantes en 8 sedes clínicas simultáneas de primer nivel de atención médica con un promedio de 65 estudiantes por sede. Para esto hubo 176 evaluadores. El estudio tomó lugar en la UNAM y los estudiantes que participaron estaban a punto de ingresar al internado.

Se realizó prueba escrita con preguntas de opción múltiple para evaluar el Perfil Intermedio II del plan de estudios del 2010. Incluyó conocimientos en áreas Clínica y Sociomédica. El examen se conformó por 232 reactivos de opción múltiple y con 4 opciones de respuesta. La evaluación se realizó en computadora con tiempo de respuesta por reactivo de 0.96 minutos. Se realizaron 2 turnos de 4 horas porque se contaba con 450 computadoras. Se empleó el software ITEMAN V.4 para el análisis psicométrico.

Por otro lado, la evaluación práctica, corresponde al examen clínico objetivo estructurado (ECOE) elaborado considerando las habilidades que debe tener un estudiante al término del 9° semestre de la licenciatura. Expertos en evaluación

realizaron los resúmenes de los escenarios clínicos, las rúbricas con escalas globales y los libretos para los pacientes estandarizados. Se realizó prueba piloto para ajustar detalles. En el ECOE participaron personas no enfermas capacitadas para hacerse pasar por pacientes estandarizados. Hubo 5 estaciones de 12 minutos c/u en dos circuitos simultáneos. El resultado por estación se midió en porcentajes de aciertos con una escala de 9 indicadores con 4 puntajes cada una que en total suman 100%.

Al final se creó un reporte para los estudiantes con los resultados de las pruebas teórica y práctica que entre otras cosas incluyó una escala global de entre 500 y 1500.

Al final de los Perfiles Intermedios I y II se realiza una prueba diagnóstica formativa donde se evalúan los conocimientos adquiridos al final de cada fase del currículo. Dichos perfiles son competencias genéricas del plan de estudios que van de niveles de complejidad 1 al 8.

Con respecto a los resultados del examen teórico, el promedio de aciertos (índice de dificultad) en la fase teórica fue de 61 ± 19.6 (promedio ± desviación estándar), la discriminación (coeficiente de punto biserial) fue de 0.18, la confiabilidad con alfa de Cronbach fue de 0.89, el error estándar de medición fue de 6.4, y los valores mínimo y máximo de aciertos fueron de 73 y 194, respectivamente. Las áreas de conocimiento de mayor puntuación fueron Ginecología y Obstetricia, Pediatría y Propedéutica en contraste a los más bajos que fueron Sociomédica y procesos de diagnóstico/ paraclínicos.

En los resultados de la fase práctica se obtuvo un promedio porcentual de aciertos global de 62.2, con una desviación estándar de 16.8. El nivel de confiabilidad con el instrumento alfa de Cronbach fue de 0.51 considerando ambos circuitos (el circuito 1 de 0.51 y el circuito 2 de 0.49). El coeficiente G del modelo de la teoría de generalizabilidad fue de 0.60.

En general los resultados de atributos más altos fueron habilidades de comunicación y de interrogatorio. Los más bajos fueron interpretación de exámenes de

laboratorio y gabinete. “La estación de embarazo no deseado tuvo la mayor puntuación (67.5); en contraste, la estación Fiebre-cefalea fue la de menor puntuación (54.8).”

45. Rivera, J., Flores, F., Alpuche, A., & Martínez, A. (2016). Evaluación de reactivos de opción múltiple en medicina. Evidencia de validez de un instrumento. Investigación en Educación Médica, 6(21), 8-15. Recuperado de https:///dx.doi.org/10.1016/j.riem.2016.04.005

El propósito del estudio fue proponer un instrumento en español que permite evaluar las características cualitativas de un reactivo de opción múltiple, de acuerdo con las recomendaciones propuestas en la literatura y se describe el proceso de obtención de evidencia de validez.

Fue un estudio no experimental de tipo descriptivo realizado en la Facultad de Medicina de la UNAM, utilizando 308 reactivos de opción múltiple aplicados en exámenes parciales sumativos durante los años escolares 2012, 2013 y 2014, para ello se llevaron a cabo 3 etapas:

- Etapa 1. Evidencia de validez relacionada con el contenido. Se elaboró tabla comparativa basada en Haladyna, Downing y Rodriguez para elegir recomendaciones más claras y precisas que fueron convertidas en afirmaciones y que pasaron a formar parte del instrumento, generando una propuesta inicial de 24 ítems. Este instrumento se sometió a evaluación de contenido por jueces.

- Evidencia de validez relacionada con el proceso de respuesta. Instrumento resultante de etapa previa consistió de 22 criterios y fue el que se aplicó para evaluar la calidad de reactivos de selección múltiple de exámenes sumativos de la asignatura biomédica. Se incluyeron reactivos de formato convencional (se plantea enunciado con pregunta y una serie de opciones con sólo una

respuesta). Se excluyeron reactivos con formato de verdadero/falso y de emparejamiento. La evaluación fue llevada a cabo por 9 profesores, siendo capacitados previamente sobre la utilización del instrumento.

- Etapa 3. Evidencia de validez relacionada con la estructura interna. Para evaluar ítems se utilizó la correlación punto-biserial de Pearson (Rpbis) y una prueba t de Student. Para análisis de confiabilidad: alfa de Cronbach. Para definir la estructura final del instrumento: análisis factorial exploratorio.

Los resultados arrojaron que las aportaciones de los revisores expertos permitió la generación de un instrumento final a partir de las recomendaciones sobre la redacción de los criterios, la unificación de algunos de ellos, y la inclusión de la clasificación taxonómica del reactivo como un indicador de la calidad del mismo. El nuevo instrumento quedó conformado por 21 criterios o ítems más la clasificación del reactivo a evaluar de acuerdo con la taxonomía modificada (evidencia de validez relacionada con el contenido).

Se evaluaron un total de 308 reactivos de opción múltiple convencionales. El índice kappa para el nivel taxonómico fue de 0.19, indicando poco acuerdo entre los jueces. El criterio 11 (“¿Es posible responder la pregunta sin necesidad de observar las respuestas?”) obtuvo un índice kappa negativo (-0.23) (evidencia de validez relacionada con el proceso de respuesta).

La mayoría de correlaciones entre criterio reflejaron poca fuerza de asociación (menores de 0.1 y mayores de -0.1). El análisis de confiabilidad resultó en un alfa de 0.615 para los 20 elementos cuantificados; al eliminar el criterio 11, la confiabilidad subió hasta 0.641. Se excluyó los criterios 4, 6, 12, 17 y 21 de la versión final del instrumento. A partir del análisis factorial exploratorio se identificaron 5 factores, cuatro con al menos 3 indicadores en su estructura y uno con solo un indicador. La varianza

explicada con 4 factores fue de 49. 979 y con los de 5 fue de 57.561 (evidencia de validez relacionada con estructura interna).

46. Sharma, N., Cui, Y., Leighton J., & White, J. (2012). Team-based assessment of medical students in a clinical clerkship is feasible and acceptable. Medical Teacher, 34, 555-561. doi: 10.3109/0142159X.2012.669083

El objetivo del estudio fue desarrollar y evaluar el método de evaluación clínico por equipos utilizado en la práctica clínica por un año académico, enfocándose en la factibilidad y aceptabilidad de la metodología para estudiantes y asesores. La investigación se realizó con 127 estudiantes (2009-2010) con 6 rotaciones en un total de 4 hospitales de enseñanza en un entorno de cirugía general, anestesiología, y medicina del dolor en su tercer año de medicina.

El instrumento de evaluación principal (una prueba de selección múltiple) fue elaborado al consultar los ítems evaluativos de 4 fuentes. Varios grupos de expertos se reunieron a discutir los ítems que mejor evaluaran las áreas deseadas. Varios ítems seleccionados fueron recurrentes en algunos grupos.

Luego, los ítems fueron revisados por representantes de cada grupo de asesores. Se le pidió a médicos, residentes, administradores, pacientes y enfermeros responder las siguientes preguntas para así modificar, agregar u omitir ítems: “En su rol, ¿observa usted a los estudiantes de medicina desempeñar la tarea o comportamiento descrita en estos ítems?”, “En su papel en el equipo, ¿piensa usted que puede evaluar qué tan bien se desempeña el estudiante con respecto a las tareas y comportamientos descritos en los ítems?”, y “¿Existen algunas otras tareas o comportamientos en estudiantes de medicina que usted observe y que no estén descritas en estos ítems?”

La versión final de evaluación contenía enunciados con respuesta de “de acuerdo/en desacuerdo” de 5 puntos con un espacio para comentarios al final. La mayoría de formularios tenían entre 6-12 ítems, el de residentes tenía 26 ítems y la autoevaluación tenía 24.

Se unieron todas las evaluaciones y se les entregó un folleto a los estudiantes. El folleto contenía 18 formularios evaluativos: médico-cirujano (6), médico-anestesiólogo (2), jefe de residentes (2), enfermero de sala de operaciones (2), y paciente (6). Los formularios fueron completados por enfermeros de guardia (2), compañeros (anónimos 4-6), administrativos (1), y autoevaluativo (1). (todo esto a lo largo de un año de rotaciones). De no cumplir con el número de formularios de evaluación completados, el estudiante reprobaría. Se transcribió todos los comentarios y se asignó una nota aprobatoria/reprobatoria a este componente.

Al final se realizó una entrevista uno-a-uno entre asesores y estudiantes, así como grupos focales.

Los resultados mostraron que: el total de asesores que completaron formularios para estudiantes fue de 1068 asesores, lo que significó 37007 ítems de evaluación individual y 3051 comentarios escritos. La consistencia interna de cada evaluación fue aceptable, la cual estuvo en un rango de 0.856 a 0.948. Ningún estudiante tuvo dos puntajes de 2 o menos (‘fuertemente en desacuerdo’ o ‘en desacuerdo’ de la escala Likert de 5 puntos).

Los estudiantes reportaron trabajar con una media de 9 cirujanos y 4 anestesiólogos por un total de 4 semanas. Los cirujanos asesoraron en promedio 18 estudiantes, y 86% de los encuentros de enseñanza registrados por los estudiantes estuvieron asociados con alguna evaluación. Un total de 104 anestesiólogos completaron un total de 273 evaluaciones de estudiantes con un promedio de asesoría de 2 estudiantes por anestesiólogo durante el año académico.

Cada estudiante recibió un aproximado de 188 palabras en la retroalimentación escrita dividido en un aproximado de 26 comentarios de “áreas de excelencia” y 5 comentarios de “áreas por mejorar.” Las entrevistas fueron conducidas por 5 médicos, 2 jefes de residentes, 2 enfermeras, 2 administrativos, y 3 estudiantes, en donde las opiniones expresadas fueron similares a los resultados obtenidos en el folleto a través de todo el año. Todos los médicos entrevistados excepto uno consideró este método de evaluación aceptable.

Un estudiante expuso que algunos residentes “fueron más allá” con pacientes que luego iban a completar sus formularios. Enfermeros reportaron que los residentes usualmente fallan en presentarse a sí mismos al equipo de enfermería con los cuales iban a estar trabajando.

47. Schweitzer, R., Khawaja, N., Strodl, E., Lodge, J., Coyne, J., & King, R. (2014). Towards a model for student selection in clinical psychology. Clinical Psychologist, 18, 125-132. doi:10.1111/cp.12025

El objetivo de la investigación fue investigar los factores de selección de estudiantes que aplican al programa de posgrado de psicología clínica. El objetivo de la medida de selección es identificar personas acordes para entrenamiento en vez de identificar personas quienes ya posean los atributos de graduandos exitosos. La población de estudio fue de: 88 aspirantes al Programa de Posgrado de Psicología clínica. Edad promedio de 28.49 años. 72 mujeres y 16 hombres.

En cuanto al instrumento de evaluación para selección del estudiante, los miembros del programa identificaron áreas críticas basadas en teoría integrativa, la literatura, y experiencia tales como: buenas habilidades de comunicación, toma de decisiones éticas, razonamiento conceptual, empatía, entre otras, así como el promedio ponderado total (GPA).

Se desarrolló un modelo de selección enfocado en 9 componentes principales respecto a las habilidades personales (consistentes con el dominio de Bennett-Levy: esquema de uno mismo, sistema reflexivo, y habilidades perceptivas interpersonales) y académicas (GPA) del estudiante. Se repartió los factores de evaluación por las diferentes estaciones que en conjunto constituían la entrevista para los aplicantes.

En la primera estación, se pidió a los aplicantes entrevistar por 10 minutos a un cliente simulado. Luego se les pidió reflexionar sobre qué les había parecido la experiencia y cómo pensaban que había sido para el "cliente." Usando checklist se medía las habilidades de comunicación del aplicante así como también microhabilidades de asesoramiento y la capacidad de auto-concientización. La estación también evaluaba un dilema ético.

En la segunda estación, se llevó a cabo la concientización de pensamiento y auto observación que fueron los aspectos cubiertos. Se les mostró un video corto con una escena emocionalmente intensa para observar su capacidad de empatía y también para explorar sus propias reacciones emocionales.

Y en la tercera estación, se cubrió razonamiento conceptual y habilidades guiadas de reflexión y comprimió habilidades de escritura para demostrar habilidades de razonamiento abstracto. Se les hizo preguntas como "¿En qué forma son la psicoterapia y la religión similares?" Al final de cada estación se asignó un puntaje.

Los resultados evidenciaron que las correlaciones variaron de leve a altas. Se correlacionó GPA con ética (.237). Reflexión guiada, toma de decisiones éticas, y habilidades de comunicación estuvieron relacionadas la una con la otra. Concientización de pensamiento estuvo asociado con auto observación y empatía.

Se analizó el principal componente con rotaciones Varimax y Oblimin. La medida de muestreo de adecuación de Keiser-Meyer-Olkin confirmó la escala de factorable al mostrar α=.69, una indicación de una buena interrelación de ítem. Las correlaciones entre ítems fueron lo suficientemente altos para análisis de factor, X^2 (36)= 304.578, p >

0.001. La solución de 3 factores. Las comunalidades entre los 9 ítems anduvo entre .282 a .863. El valor de GPA tuvo el valor más bajo y el de escritura el más alto. Las habilidades de escritura parecen estar altamente relacionadas al instrumento desarrollado.

Las cargas factoriales indican que el factor 1 está constituido por 3 habilidades: empatía, concientización de pensamiento, y auto observación, el cual constituyó 35.71% del total de varianza. El factor 2 consiste de 3 habilidades: reflexión guiada, y toma ética de decisiones, éste aportó 20.56% del total de varianza. El tercer factor consiste en: habilidades de escritura, razonamiento conceptual, y GPA. Este factor cuenta con 18.24% del total de varianza.

48. Emery, J., Bell, J., & Vidal-Rodeiro, C. (2011). The biomedical admissions test for medical student selection: issues of fairness and bias. Medical Teacher, 33, 62-71. doi: 10.3109/0142159X.2010.528811

El objetivo del estudio fue investigar las relaciones entre las variables de antecedentes de los aplicantes y los puntajes BMAT, y ver si fueron rechazados u ofrecidos un lugar, para aquellos admitidos, investigar el desempeño en las evaluaciones de los cursos de primer año.

La población de estudio fue de 2577 aplicantes, a los cuales 659 se les ofreció un lugar. Aplicantes al Programa de Pregrado de Medicina de menos de 21 años y que fueran de Inglaterra. Estudio llevado a cabo por la Universidad de Cambridge del 2003-2005 (para ingreso entre 2004-2006).

BMAT (Examen de Admisión a BioMedicina), consiste en una prueba en papel de 2 horas conteniendo 3 secciones: (i) Aptitud y habilidades: ítems de selección múltiple o que requieren respuestas numéricas de solución de problemas, comprensión de argumentos, datos e interpretación gráfica e inferencia. (ii) Conocimiento científico y

aplicaciones: ítems de selección múltiple o que requieren respuestas numéricas de solución de problemas basados en las bases de biología, química, física y matemática. (iii) Prueba de escritura: un ensayo corto escogido de 3 opciones donde se evalúa claridad en la comunicación y habilidades argumentativas.

Las secciones 1 y 2 son calibradas en una escala de 1-9. Ensayos son evaluados holísticamente de 0-15. Las notas de las 3 secciones son presentadas por separado y no combinadas. La Universidad de Cambridge no usa el puntaje de la sección 3 en el proceso de admisión pero sí utiliza este ensayo como evidencia cualitativa. Cambridge Assessment es la oficina encargada de aplicar la BMAT y de suministrar los datos de los aplicantes. La Oficina de Admisiones de la Universidad de Cambridge proporcionó el tipo de escuela de donde provenían los candidatos y la Universidad a la que aplicaron. Se calculó una variable de frecuencia para indicar el número de aplicantes ingresados por centro de aplicación de BMAT en un periodo de 3 años para investigar los efectos de la experiencia de los centros aplicantes y la probabilidad de que al aplicante se le ofreciera un lugar. Centros ingresando 1-2 aplicantes fueron clasificados como “inexpertos,” de 3-9 como “moderadamente expertos” y 10 o más como “muy expertos.”

Se obtuvo también información socioeconómica sobre los candidatos. Se construyó 3 modelos de regresión multinivel para investigar las 3 preguntas de investigación: los factores asociados a las notas BMAT, factores asociados a que se les ofreciera un cupo, y a aquellos admitidos, los factores asociados con los resultados de los cursos de primer año.

Con respecto a los resultados, los estudiantes de tipo “A” (nota “A” en matemática/ciencias, química y alguna otra materia excepto estudios generales o pensamiento crítico) representaron el 3.4% (7717 de 223.981 candidatos en 2006). Las escuelas independientes tuvieron una representación de 14% en estudiantes de tipo A. Una mayor proporción de aplicantes provenían de educación más avanzada a nivel formal de sistema de enseñanza y de escuelas independientes. En este sentido, los factores asociados con los puntajes BMAT de los aplicantes, dieron una regresión

positiva de coeficientes, lo que indica asociación positiva con las notas BMAT, coeficientes negativos indican una asociación negativa.

49. Dowell, J., Lynch, B., Till, H., Kumwenda, B., & Husbands, A. (2012). The multiple mini-interview in the UK context: 3 years of experience at Dundee. Medical Education, 34, 297-304. doi: 10.3109/0142159X.2012.652706

El objetivo del estudio fue presentar información metodológica, cuestionaria y psicométrica en el proceso de transición de entrevistas tradicionales a MMI's (Mini entrevista múltiple) en un periodo de 3 años y discutir las implicaciones para aquellos considerando esa transición. Para lo cual, hubo 452 candidatos entrevistados en 2009 y 477 entrevistados en 2010.

El proceso de transición. Se empleó una fase transicional entre el sistema antiguo de entrevistas y la implementación completa de la MMI para 2009. Por lo tanto, se hizo un plan piloto que fue un híbrido con contenido de la entrevista vieja utilizando un enfoque rotacional. El plan piloto consistía de 4 estaciones de 10 minutos cada una, 3 de las cuales eran estaciones de entrevistas tradicionales uno-a-uno. La 4ta consistía de una evaluación interactiva donde se hacía un role-play. Esta estación se enfocó en empatía y trabajo en equipo.

Los asesores recibieron de 15-20 minutos de entrenamiento general, y 15 minutos de entrenamiento específico por estación. Los asesores evaluaron candidatos en 3 atributos por estación usando escalas de Likert de 5 puntos. Entonces el máximo de puntos por estación era de 15 y de 60 en total. El alpha de Cronbach del total de las estaciones MMI fue de 0.66 con un rango de estación/puntaje total de correlaciones de entre 0.259 y 0.387.

Se establecieron 10 estaciones con 1 asesor por estación. Se generó un conjunto superior de posibles atributos vía revisión de literatura, el cual luego fue evaluado y

puesto a discusión por el comité de admisiones de la Escuela de Medicina. En este proceso se definieron 6 atributos o “dominios” (habilidades interpersonales y de comunicación, razonamiento lógico y pensamiento crítico, razonamiento moral y ético, preparación y motivación para estudiar medicina, liderazgo y trabajo en equipo, y honestidad e integridad). El contenido específico y la rúbrica de evaluación de cada estación se generó por: un psicólogo cognitivo y un conferencista en Educación Médica, el director de admisiones de pregrado, y un estudiante de doctorado relacionado con investigación MMI. De cada estación se obtenía información de 3 de los 6 dominios y cada una tenía una escala Likert de 5 puntos.

Se implementó un sistema de “bandera roja” para candidatos en que asesores presentaran grandes preocupaciones respecto a su desempeño. Al obtener dos banderas rojas el participante sería excluido y se aseguraba de que no fueran excluidos por la percepción de uno solo de los asesores. Cada estación tenía una duración de 7 minutos.

Se aplicó análisis de varianza (ANOVAs), cálculos con SPSS 17.0 para análisis estadísticos. Comparaciones ANOVA post hoc utilizaron la prueba de Fisher menos significativa. La r de Pearson se utilizó para correlacionar los puntajes MMI con otras medidas de admisión previas a la entrevista. Y finalmente se aplicó el modelo multifacético de Rasch.

Los resultados mostraron que los puntajes promedios no fueron significativos a través de los días de las MMI, o incluso entre los años de aplicación. Un ANOVA mostró diferencias significativas entre los graduados/maduros, estudiantes desertores del Reino Unido y candidatos del extranjero en ambos años; 2009: $F(2) = 7.75$, $p < 0.01$; 2010: $F(2) = 13.23$, $p<0.01$. Comparaciones post hoc muestran que candidatos graduados/maduros alcanzan promedios significativamente más altos que el resto de grupos. El Alpha de Cronbach para las 10 estaciones; 2009: 0.70; 2010: 0.69.

La correlación de MMI con otro tipo de medidas, no obtuvo significancia estadística. El rango de correlaciones fue de -0.075 a 0.123 en 2009, y de -0.077 a

0.067 en 2010. Candidato y punto de vista del asesor. Se analizaron 324 (75%) cuestionarios de entrevistas y 116 (58%) cuestionarios de asesores. Una gran mayoría (94% y 90% respectivamente) seleccionó “estrictamente de acuerdo” con respecto a si la MMI fue justa. 33% indicó que la prueba MMI fue más estresante que la entrevista tradicional.

Los cuestionarios fueron devueltos y analizados de 324 (75%) de entrevistados y 116 (58%) de los evaluadores. Una abrumadora mayoría (94% y 90%, respectivamente), se obtuvo un totalmente de acuerdo en respuesta a la pregunta de si el MMI fue justo. Una minoría considerable de entrevistados (33%) consideró que el MMI fue más estresante que la entrevista tradicional.

El 98% indicó que el uso de la MMI no tuvo ningún efecto o en realidad mejoró su punto de vista de la Escuela de Medicina de Dundee. Una gran mayoría (74%) de los que ya fueron entrevistados en otra Escuela de Medicina indicó una preferencia por la MMI. Una pequeña mayoría (60%) dijeron que recomendarían Dundee a un amigo debido a la MMI. Cuando preguntó qué tan bien las estaciones lograron lo que se propusieron hacer, siete estaciones recibieron calificaciones de "muy bien" o "moderadamente" bien 'de al menos el 75% de los encuestados.

50. Monroe, K. (2016). The relationship between assessment methods and self-directed learning readiness in medical education. International Journal of Medical Education, 7, 75-80. doi:10.5116/ijme.56bd.b282

El objetivo del estudio fue explorar la preparación en aprendizaje autodirigido dentro de una evaluación comprensiva del conocimiento y habilidades de estudiantes de medicina así como también hasta qué punto cuáles variables predicen la preparación en el proceso de aprendizaje autodirigido en los participantes previo a su graduación. Los participantes fueron estudiantes de 4to año de medicina de una universidad de

medicina al sureste de los Estados Unidos, el número de participantes fue de 124 y el número de entrevistas recolectadas fue 91.

Se realizó múltiples análisis estadísticos de regresión. Los participantes completaron voluntariamente una escala de evaluación de preparación autodirigida. Se consideró 5 variables independientes (GPA [prom. Pond. total], notas OSCE, prueba de licencia médica de los Estados Unidos [UMSLE], pasos 1 y 2 de conocimiento clínico [CK], y evaluaciones de pasantías y variables descriptivas y de inconveniencia).

Se obtuvo las notas de los participantes en MCAT (Prueba de Admisión a las Universidades de Medicina) -una posible variable de inconveniencia- como parte de su proceso de aplicación a la universidad (previo a su matrícula). Las demás variables de predicción se generaron conforme los estudiantes avanzaban en su plan de estudios.

Se empleó la escala de Hendry & Ginns para el aprendizaje autodirigido (SDLRS) para medir las 4 variables dependientes del estudio. La escala consiste de 36 ítems evaluados en escala Likert de 5 puntos. Para el propósito de este estudio, cada subvariable de SDLRS presentó una variable dependiente: autoevaluación crítica, eficacia propia de aprendizaje, autodeterminación, y organización efectiva para el aprendizaje.

Con respecto a los resultados, al final se eliminó la variable OSCE del análisis porque sólo estaba registrada como aprobado/reprobado. Se encontró que los participantes que respondieron se desempeñan bien en los cursos y evaluaciones, lo cual no es de sorpresa ya que han sido estudiantes que han avanzado a través de una escuela de medicina competitiva. Sus respuestas indican que tienen una aptitud para el aprendizaje autodirigido.

El componente de predicción más fuerte en el aprendizaje autodirigido fue las evaluaciones de profesores de la facultad en diferentes rotaciones en pasantías. La autoevaluación crítica y eficacia propia de aprendizaje tuvieron un nivel estadístico significativo, y las correlaciones de Pearson fueron moderadas ($r = -.30$, $p = .01$) para

cada una de las subescalas. La organización efectiva para el aprendizaje se relacionó nominalmente con el desempeño de una tercera rotación (r = .21, p = .05)

GPA, puntaje del paso1 de USMLE, y nota del paso 2 de USMLE de CK no fueron factores de predicción. Los puntajes MCAT no predijeron ninguna subescala de SDLRS, y el MCAT no mejoró la calidad de predicción de ningún predictor potencial. Entonces, teoría estadística sugiere que el rango de restricción de puntajes de MCAT en la muestra (27-39 de un máximo de puntaje de 40) pudo haber contribuido a una correlación menor que la observada en una distribución de puntajes normal y no restringida.

51. Allerup, P., et al. 2007). Use of 360-degree assessment of residents in internal medicine in a Danish setting: a feasibility study. Medical Teacher, 29, 166-170. doi: 10.1080/01421590701299256

Para este estudio participaron 44 residentes de medicina interna de 6 hospitales. 23 mujeres, 19 hombres. Se eligió el periodo de residencia de medicina interna para el periodo de estudio. El estudio fue desarrollado independientemente de otras evaluaciones y no influyó en la aprobación o rechazo de la pasantía. Participaron 6 departamentos de medicina interna. Se escogieron 22 de 65 objetivos de aprendizaje para la evaluación de 360 grados, los cuales se encuentran dentro de los dominios: experto médico (6 objetivos), comunicador (3 objetivos), colaborador (4 objetivos), líder/administrador (2 objetivos), académico (3 objetivos), y profesional (4 objetivos).

Se desarrolló un formulario donde se citaba cada uno de los objetivos de la versión original de objetivos junto con una escala Likert para cada objetivo. Cada residente fue evaluado por una secretaria, 4 enfermeros, y 5 médicos mayores. Cada vez que algún asesor marcaba “no tan satisfactorio” o “insatisfactorio” debían escribir comentarios. Después de 7-10 días los estudiantes eran reevaluados al igual que la

primera vez. Además, se les pedía a los estudiantes evaluarse ellos mismos con los mismos formularios.

Para el análisis de datos se utilizó el programa estadístico SAS. La concordancia intra asesor de la toma por segunda vez de la evaluación entre los 7 y 10 días fue evaluada utilizando el coeficiente kappa de Cohen. Consistencia interna fue evaluada por medio del alfa de Cronbach. También se utilizó ANOVA. Los resultados de los enfermeros y doctores fueron comparados utilizando la correlación de Spearman.

Con respecto a los resultados, participaron en la evaluación 16 secretarias, 94 enfermeras, y 85 doctores con experiencia. El tiempo aproximado para llenar los formularios fue 14.5 minutos. Uno de los cuadros en la escala Likert tenía la opción “no es posible evaluar este criterio” y se logró determinar cuáles ítems podrían ser evaluados por secretarias, enfermeras y doctores, respectivamente. Secretarias pudieron evaluar 2, enfermeras 7 y doctores 14 objetivos.

Los valores Kappa más altos fueron obtenidos por el dominio “experto médico”. Hubo una buena correlación entre las evaluaciones de médicos y las de enfermeras, la cual fue estimada de 0.60 de las correlaciones directas entre diferencias (Correlación de Spearman, $P < 0.001$).

No hubo diferencia significativa sobre cómo las mujeres se evaluaban a sí mismas en relación con los hombres residentes. De los 42 residentes, 29 fueron evaluados como “satisfactorio” o “muy satisfactorio” en todos los 15 objetivos de aprendizaje. Los restantes 13 residentes obtuvieron en el cuadro “no tan satisfactorio” para uno o más de los objetivos. (“Insatisfactorio” fue utilizado sólo una vez).

Se encuentra que los residentes sistemáticamente sobreestiman su competencia clínica comparada con las evaluaciones de las enfermeras y médicos. Las tasas de respuesta en los cuestionarios fueron de 38% en los residentes, 55% de enfermeras y 94% de médicos.

Es importante tener en cuenta que la credibilidad de este método varía entre dominios. La evaluación de 360 grados es considerada especialmente apta para objetivos de aprendizaje “humanísticos”**.** Las respuestas de los residentes en los cuestionarios fueron bajas. Una de las posibles explicaciones fue que los cuestionarios se les entregaron dos semanas antes de terminar la rotación en Medicina Interna cuando los residentes estaban a punto de ser asignados a sus siguientes rotaciones. Desde esta investigación se recomienda este método para ser utilizado a una etapa temprana de entrenamiento en especialidad.

52. Zaidi, N., Swoboda, C., Wang, L., & Manuel, S. (2014). Variance in attributes assessed by the multiple mini-interview. Medical Teacher, 36, 794-798. doi: 10.3109/0142159X.2014.909587

El estudio tuvo como objetivo explorar si los ítems, definidos como atributos específicos en un formulario de evaluación MMI (Mini entrevista múltiple), son evaluados consistentemente a través de estaciones MMI sin importar el escenario de la estación. Los datos analizados, van desde 2009 a 2013. Análisis de datos en proceso de admisión en una universidad de Estados Unidos.

Se identificó una característica principal para evaluar a través de la MMI: Comunicación. Para tomar esta decisión y elegir comunicación se basó en estudios de literatura. Se utilizó una tabla de evaluación MMI como sub puntaje que tenía 6 atributos específicos y un “puntaje general”. Los 6 atributos incluían: perspectivas múltiples, reflexión sobre escenario, articulación, interés en el dilema, comunicación no verbal, y habilidades interpersonales. Estos 7 [dice 7 y no 6] puntos fueron evaluados en una escala Likert de 7 puntos que asume intervalos iguales entre puntajes principales (Insatisfactorio-1, Por debajo del promedio-2, Cerca del promedio-3,

Promedio-4, Poco más por encima del promedio-5, Por encima del promedio-6, Sobresaliente-7).

Se utilizó el software G-String IV para estimar la varianza entre componentes atribuible a las facetas de medición. En la teoría G, el objeto de medida no es considerado una faceta. Por lo tanto, el diseño duo facético de este estudio incluye el objeto de medición -aplicantes (p), y dos facetas de generalización – escenario (s) e ítem (i). La faceta de diferenciación es la persona, el aplicante (p), el cual representa verdadero para el puntaje MMI del aplicante. Las varianzas de escenario y evaluador son completamente ajenas al control del estudio. Por lo tanto, se considera una limitante en este caso y será atribuible al escenario. Se realizó un estudio G cruzado de prueba del escenario (s) con el aplicante (p). Entonces se hizo un muestreo de los subconjuntos de los conjuntos de datos para aplicantes evaluados dentro del mismo escenario usando los mismos ítems para asegurarse de un diseño completamente cruzado.

A pesar de que el puntaje real (p) representa una cantidad significativa de varianza, los aplicantes (p) representan solamente un 6% del total de varianza. Los componentes de varianza estimados del estudio G sugieren que la mayor cantidad de varianza es atribuible al efecto principal de la faceta de escenario y aplicante (ps). Colectivamente, estos componentes componen un total del 77% del total de la varianza. La faceta de ítem (i) representa la faceta más baja de las varianzas estimadas, con sólo 0.6%. La interacción ítem-escenario corresponde al 1.4% del total de varianza. La baja estimación de varianza atribuible a la faceta de ítem es reforzada por el alfa de Cronbach (0.97) para los 7 ítems, lo cual sugiere consistencia interna muy alta entre los atributos medidos por esta MMI, esta alta consistencia interna puede admitir suposiciones de que el proceso actual de MMI es midiendo un atributo unidimensional; esto es más apoyado por solo el 2% de la varianza atribuible a la faceta del artículo (i), (pi) y (si).

Asimismo, estos hallazgos apoyan la teoría G, suposición de que las condiciones de la faceta del artículo se pueden considerar intercambiable o puede sugerir que los

evaluadores no entienden cómo usar los elementos asociados con la herramienta de evaluación de MMI y simplemente asigne el mismo valor para cada artículo.

53. **Kaliyadan, F., Khan, A., Kuruvilla, J., Feroze, K. (2014). Validation of a computer based objective structured clinical examination in the assessment of undergraduate dermatology courses. Indian Journal of Dermatology, Venereology, and Leprology, 80 (2), 134-136. doi: 10.4103/0378-6323.129386**

El objetivo del estudio fue validar la OSCE's (prueba objetiva clínica estructurada) como un método de evaluación en dermatología y la facilidad relativa de la administración del mismo formulario OSCE basado en computadora. Para ello participaron 129 estudiantes de quinto año de medicina en un curso de dermatología. El estudio se realizó por 3 semanas que es el tiempo que dura la rotación para el curso de dermatología.

A cada estudiante se le mostró 16 imágenes que representaban casos clínicos comunes. Las preguntas que se les hacían eran sobre la descripción de las lesiones de la piel, diagnóstico y diagnóstico diferencial, investigación y tratamiento. El tiempo entre cada filmina fue de 5 minutos y las imágenes fueron presentadas por medio de Microsoft PowerPoint.

Los puntajes del OSCE se correlacionaron estadísticamente con los puntajes obtenidos en la presentación del caso clínico y los puntajes totales, incluyendo la evaluación escrita. Se evaluó la retroalimentación de los estudiantes por medio de un cuestionario de 7 preguntas, evaluadas en una escala Likert de 5 puntos, este instrumento también contenía preguntas abiertas. La correlación estadística entre puntajes fue realizada usando el coeficiente de correlación de Pearson.

Los resultados muestran que los puntajes de la prueba OSCE en computadora mostraron una correlación positiva con los puntajes de la presentación clínica

(Coeficiente de Pearson -0.923, valor P <0.000, significativo al nivel 0.01) y buena correlación con los puntajes en general (coeficiente de Pearson -0.728, valor P < 0.000, significativo al nivel .01), lo cual indica que este método es confiable para evaluaciones en dermatología.

El cuestionario de retroalimentación indicó buena confiabilidad (alfa de Cronbach -0.78). 62.7% de los estudiantes dijo estar de acuerdo o totalmente de acuerdo en que preferían este método de evaluación sobre la evaluación tradicional. 99 estudiantes (76.7%) estuvo de acuerdo o totalmente de acuerdo en que los temas abarcados en el OSCE fueron relevantes. También se observó desacuerdo con respecto al tiempo dado para cada filmina. 63 participantes (48.8%) estuvieron en desacuerdo o muy en desacuerdo en que el tiempo dado por cada filmina fue suficiente.

54. Rahim, A., & Yusoff, M. (2016). Validity evidence of a multiple mini interview for selection of medical students: Universiti Sains Malaysia experience. Education in Medicine Journal, 8 (2), 49-63. doi: 10.5959/eimj.v8i2.437

El objetivo del estudio fue describir la implementación de la MMI (mini entrevista múltiple) y reportar la evaluación de información preliminar en la validez de su evidencia. La población de estudio fue: 447 participantes y 30 entrevistadores por sesión. Estudio se hace en el 2015 cuando se hace el ejercicio de selección de estudiantes.

Se implementaron 9 estaciones: 5 operadas por personas y 4 de descanso. En cada estación se dura 7 minutos (5 minutos + 2 minutos para que evaluadores escriban y para la preparación de los candidatos. Para atender el mayor número de candidatos, se emplearon 6 circuitos idénticos por 2 días. Cada sesión con un tiempo de aproximadamente 1hora con cerca de 6 sesiones por día.

Cada estación se enfocó en los siguientes dominios: razonamiento crítico, habilidades de comunicación, concientización ética, y conocimiento del sistema de salud, preguntas estándar, desempeño en el idioma, y conducta en general. Para disipar el aburrimiento se les permitió cambiar de estación al haber completado una

sesión. También participaron entrevistadores no académicos como miembros del cuerpo médico de hospitales y enfermeros ya que son ellos quienes van a recibir a los estudiantes en los hospitales.

Cada estación evaluó un máximo de 4 dominios. A pesar de que evaluadores calificaron cada dominio por separado, se les pidió también proveer una calificación general del desempeño. Se realizó un taller de entrenamiento. También se realizó una simulación de entrevista la cual fue grabada. Dicho taller se realizó unas semanas previas a la entrevista. La mañana de la entrevista también se les solicitó llegar antes para una pre-entrevista y se les entregó una guía ya que se esperaba el mayor nivel de similitud entre estaciones.

Después de registrarse, participaron en una mini inducción. La evaluación de MMI en SMS (Escuela de Ciencias Médicas). Para evaluar aceptabilidad, se aplicaron cuestionarios con escalas Likert de 7 puntos a los candidatos y entrevistadores. Para medir factibilidad, se realizó un análisis comparativo con entrevistas anteriores. Para medir la calidad de las estaciones MMI, se midió los índices de dificultad y discriminación en las preguntas. Para medir la construcción de validez, se realizó confirmación de factor de análisis para medir el modelo de evaluación del constructo latente. La AVE (Varianza aproximada extraída) fue de 0.77, lo cual significa validez convergente

Con respecto a los resultados, en cuanto a la **v**alidez y confiabilidad. Se usó la Guía de Clasificación de Ítem. Si se considera cada estación por separado entonces 40% de las estaciones son Nivel I, 43.3% Nivel II, 3.3% Nivel III, y 13.4% Nivel IV. El análisis de confiabilidad reveló que el valor CR fue 0.94, lo cual indica un alto nivel de consistencia interna.

Los candidatos fueron bastante positivos respecto a la calidad e implementación de la prueba, donde la media de las evaluaciones estuvieron todas por encima de 5. Con respecto a conocimiento especial para responder las preguntas, todas las medias estuvieron por encima de 3.5 y cerca de 4. Se encontró resultados similares con

respecto al tiempo disponible por estación. Las medias estuvieron por encima de 3.5 excepto por estación 1. La confiabilidad general del ejercicio MMI fue 0.94, un aceptable nivel para tomar decisiones de alto riesgo.

55. Baños, J., Gomar-Sancho, C., Grau-Junyent., Palés-Argullós, J., & Sentí, M. (2015). El mini-CEX como instrumento de evaluación de la competencia clínica. Estudio piloto en estudiantes de medicina. FEM, 18 (2), 155-160.

El objetivo del estudio fue presentar los resultados de un estudio piloto para evaluar la factibilidad del empleo del Mini-Clinical Evaluation Exercise (mini-CEX) en estudiantes de Grado de Medicina. La investigación se realizó entre abril y junio del 2014, participaron 13 tutores y 27 estudiantes.

El estudio se aplicó a estudiantes en asignaturas del Grado de Medicina como: Semiología General y Propedéutica Clínica, de tercer curso; aparato respiratorio/cirugía torácica, de cuarto curso, y nefrología/urología, de quinto curso.

Se reunió estudiantes con tutores para evaluar el desempeño de los estudiantes. El estudiante realizó la tarea que el tutor le designó de acuerdo con la asignatura. El tutor permaneció con el estudiante durante la tarea para observar su desempeño (fase de observación). Al final se daba retroalimentación por parte del tutor sobre los aspectos más relevantes del desempeño, tanto positivo como negativo (fase de feedback).

Posteriormente, el tutor anotaba el tiempo en la fase de observación y en la de feedback y le pedía al estudiante su grado de satisfacción. Anotado todo eso, se entrega el documento al responsable de la asignatura para posterior análisis. La hoja de recogida de datos es un cuadro que contiene escala Likert. Finalmente, los datos fueron

procesados para obtener la media y el rango. Se utilizó la prueba t de Student y la prueba U de Mann-Whitney para la comparación entre grupos.

Con respecto a los resultados, el mayor número de estudiantes y tutores pertenecía a la asignatura de Semiología y Propedéutica. Los diagnósticos de los pacientes que fueron evaluados difirieron según el servicio donde los estudiantes realizaban sus prácticas. El diagnóstico más frecuente en cirugía torácica fue el cáncer de pulmón (n = 5), y en urología, los trastornos del tracto urinario y los tumores vesicales (n = 5). En medicina interna: enfermedades respiratorias (n = 10), cardiopatías (n = 7) y cuadros infecciosos (n =8)

La valoración media de los estudiantes por parte de los tutores en escala de 10. Aparato respiratorio/cirugía torácica: 8.5, Nefrología/Urología: 8.6, y Semiología y Propedéutica: 7. Se observó diferencias entre tutores en la tercera, en donde las puntuaciones medias de los tutores variaron entre 6.4 (rango: 5-7) y 8.5 (rango: 7-9).

Con relación a los tiempos de evaluación, en Semiología y Propedéutica eran aproximadamente 12 minutos. El tiempo de feedback en Semiología y Propedéutica y en Nefrología/Urología fue similar, pero inferior al empleado en aparato respiratorio/cirugía torácica. La media en tiempos de evaluación y feedback fue de 14 y 8.4 minutos. La media de la satisfacción de la experiencia de tutores y estudiantes fue siempre superior a 8.5.

En todas las asignaturas, las puntuaciones de los estudiantes fueron superiores a las de los profesores, aunque con diferencias no significativas. Los aspectos destacables y mejorables variaron según la madurez del estudiante. Semiología y Propedéutica: (aspectos destacables) predisposición a la prueba, la exploración sistemática, las habilidades de comunicación y la profesionalidad. (Mejoras) Recogida de la anamnesis y la realización de una exploración física sistemática. En las asignaturas de 4to y 5to curso: (destacable) habilidad de comunicación, empatía, y toma de anamnesis. (Mejoras) Exploración física.

56. Fernández, G. (2011). Evaluación de las competencias clínicas en una residencia de pediatría con el Mini-CEX (Mini-Clinical Evaluation Exercise). Archivos argentinos de pediatría, 109(4), 314-320.

El objetivo del estudio fue evaluar las competencias clínicas de los residentes de pediatría con la implementación del Mini-CEX, determinando su validez, confiabilidad, factibilidad y la satisfacción de docentes y de residentes.

Se utilizó el Mini-CEX, método basado en la observación directa del desempeño del residente durante su práctica diaria, por parte de un docente. Las observaciones del Mini-CEX se realizaron durante el segundo semestre de 2010, abarcando situaciones clínicas en consultorios externos, sala de internación pediátrica, Unidad de Cuidados Intensivos Neonatales (UCIN), sala de recepción del recién nacido y habitaciones de internación conjunta madre-recién nacido. El grupo de evaluadores estuvo compuesto por diez docentes de la residencia y cuatro docentes de la Facultad de Medicina de la Universidad Nacional del Comahue. Las evaluaciones se realizaron durante la práctica profesional habitual de los residentes, en los distintos ámbitos de atención donde les correspondía actuar. Los docentes observaron y evaluaron la actuación del residente y anotaron en las fichas estandarizadas (Anexo 1) sus apreciaciones, considerando el grado de complejidad de la situación clínica, de acuerdo a su criterio en: bajo, moderado o alto. Se consignó si el enfoque de la observación estuvo destinado a la anamnesis, el diagnóstico, el tratamiento o el asesoramiento del paciente y la familia, o cualquier combinación de éstos.

El puntaje obtenido por los residentes en su competencia global varió de acuerdo al escenario clínico donde se desarrolló la observación. El puntaje más elevado fue en neonatología, con un puntaje medio de 7,45 (± 0,77), el más bajo en sala de internación pediátrica con 6,29 (± 1,33), p= 0,002. El tiempo de duración del Mini-CEX, la satisfacción de los docentes y la de los residentes también tuvieron una variación estadísticamente significativa de acuerdo con el escenario clínico (Tabla 3). Se calculó el alfa de Cronbach para cada competencia considerando todas las evaluaciones de los docentes. El coeficiente de correlación entre ítems varió entre 0,91 y 0,97, lo que indica muy buena consistencia interna entre los ítems del instrumento. El coeficiente alfa de Cronbach fue de 0,97, lo que indica elevada confiabilidad del método de evaluación. Se realizó un ANOVA de un factor entre los puntajes de competencia global de cada uno de los docentes y se observó una diferencia estadísticamente significativa entre los mismos, p <0,0001.

El Mini-CEX conjuga bastante bien los criterios descriptos, tiene validez ya que permite diferenciar entre diferentes niveles de experiencia entre los residentes.5-8 Tiene confiabilidad, ya que múltiples encuentros, con diferentes pacientes y situaciones clínicas y distintos observadores, permiten obtener conclusiones sobre la competencia clínica global. Diferentes autores afirman que son necesarios entre 10 y 15 encuentros para que la evaluación sea confiable y los resultados reproducibles.7,5,12,13 tienen un gran impacto educativo dado por lo significativo de la devolución constructiva en el aprendizaje futuro del residente.14,15 Es aceptado favorablemente por docentes y residentes, de acuerdo a los niveles de satisfacción mostrados por diferentes estudios incluyendo el presente.5-8,16 La implementación del Mini-CEX durante este estudio, permitió evaluar a los residentes en todos los escenarios clínicos de la práctica diaria de un pediatra. Fue factible de implementar ya que se desarrolló en un solo programa formativo y se hicieron más de 22 observaciones por residente en promedio, muchas más que las informadas por otros autores que variaron entre 2 y 4 evaluaciones.

57. Eva, K., Reiter, H., Trinh, K., Wasi, P., Rosenfeld, J., & Norman, J. (2009). Predictive validity of the multiple mini-interview for selecting medical trainees. Medical Education, 43, 767-775. doi:10.1111/j.1365-2923.2009.03407.x

El objetivo del estudio fue reportar pruebas más allá de la validez del proceso de selección de las mini entrevistas múltiples (MMI), comparando puntajes MMI con aquellos obtenidos en pruebas de habilidades clínicas a nivel nacional.

Con respecto a la metodología se destacan tres procesos importantes, el punto 1, en el cual se evaluó la estabilidad del desempeño, para lo cual se tomó una muestra de 29 estudiantes residentes (de segundo y tercer año, con edad promedio 28.7 años) que fueron reclutados de los programas de medicina interna (n=2) y de oncología radiológica (n=27). Esta investigación tomó lugar en la Universidad McMaster. Estudio se hace con 9 estaciones MMI. En este proceso los participantes fueron asignados a uno de los 3 circuitos que corrían paralelos (2 en un día, el otro una semana después). Cada circuito estaba compuesto por estaciones de 10 minutos cada una, en donde había un evaluador en cada una.

Posterior a cada desempeño, los participantes eran evaluados en una hoja que contenía 4 ítems con 7 puntos a evaluar. En total participaron 18 evaluadores a lo largo de los 3 circuitos (9 por día). Con respecto a los resultados del punto 1, la confiabilidad de cada estación fue G=0.24, en donde la confiabilidad del puntaje total generada por la ponderación de las 9 estaciones fue de 0.76. Se sugiere que una evaluación con 12 estaciones podría subir el nivel de confiabilidad a 0.80. La confiabilidad de cualquiera

de las estaciones por separado es baja, generalmente <0.25, en donde el promedio para 12 estaciones en total se ha encontrado ser de 0.73.

Por otro lado, con relación al punto 2, se evaluó la validez predictiva del MMI. En esta línea, los estudios previos han demostrado una relación complementaria entre la capacidad de predicción de GPA (promedio ponderado total) y la MMI.

La prueba de dos partes MCCQE es una prueba a nivel nacional que todos los estudiantes de medicina deben pasar para licenciarse en el Colegio de Médicos de Canadá (MCC). Para que los estudiantes puedan pasar a la II parte de la prueba MCCQE, deben pasar la primera más tener mínimo un año de experiencia en el entrenamiento posterior a su graduación.

Para evaluar la predicción se correlacionó los puntajes obtenidos en la MMI del estudio de posgrado realizado en la sección anterior con aquellos obtenidos en la parte II del MCCQE. Participaron 117 aplicantes en el 2002 quienes habían participado en la MMI además de otros 4 procesos de admisión. Estos estudiantes fueron entrevistados ante un panel de 3 jueces y luego participaron en una MMI de 10 estaciones.

Los resultados del punto 2, en la muestra de posgrado, la MMI fue estadísticamente predecible respecto al porcentaje de estaciones que los candidatos pasaron en la segunda parte de la prueba MCCQE ($r = 0.43$, $P < 0.05$) y la cual tendió a ser estadísticamente predictiva del puntaje total ($r = 0.36$, $P < 0.1$). Sin embargo, se encontró que la MMI fue el único predictor estadístico del porcentaje de estaciones pasadas por los candidatos en la parte II de MCCQE.

Finalmente, en el punto 3, se revisó la relación entre las medidas de competencias “cognitivas” y las “no cognitivas” En los 3 puntos de tiempo, los puntajes que estaban destinados generalmente a reflejar competencias “cognitivas” estaban correlacionados para reflejar generalmente competencias “no cognitivas.” Primer punto de tiempo: la aplicación a la Escuela de Medicina, relacionado al puntaje MMI con el GPA para el grupo del 2002 de 117 aplicantes descrito arriba. Segundo punto de tiempo:

la conclusión de la Escuela de Medicina y entrenamiento. Tercer punto de tiempo: entrenamiento de residencia.

Los resultados del punto 3, arrojaron que la correlación entre el GPA y el puntaje MMI, para 117 aplicantes, fue r = - 0.23 (P < 0.01). La correlación entre los dominios cognitivos predominantes en la Parte II y los dominios no cognitivos predominantes de la misma prueba para los 34 estudiantes que coinciden con los datos fue de r = 0.27 (P < 0.10). La misma comparación para los 22 residentes con información disponible que tomó la MMI fue r = 0.43 (P < 0.05).

58. Axelson, R. & Kreiter, C. (2009). Rater and occasion impacts on the reliability of pre-admission assessments. Medical Education, 43, 1198-1202. doi:10.1111/j.1365-2923.2009.03537.x

El objetivo del estudio fue examinar las contribuciones relativas de las dos facetas, y la confiabilidad del puntaje de la entrevista. Con el afán de valorar el aporte de la prueba MSPI (Entrevista de admisión a la Escuela de Medicina). La población de estudio fue de 3071 estudiantes que aplicaron para una Escuela de Medicina en el medio oeste de los Estados Unidos entre 2003 y 2007.

Con respecto a la metodología, de los 3071 aplicantes, 1448 fueron aceptados a la Escuela de Medicina, y de esos, 711 consolidaron matrícula. Cada entrevistado participó en una entrevista de 25 minutos conducida por dos profesores. Las entrevistas iniciaban con un componente estructurado, en donde se les leía a los candidatos y ellos respondían a una serie de 4 preguntas predeterminadas. Las respuestas eran puntuadas inmediata e independientemente en una escala de 1 - 5. Posteriormente se abrió espacio para una entrevista libre, la cual fue también calificada en una escala de 1 a 5.

A lo largo de los 5 años, 168 aplicantes fueron entrevistados dos veces en años consecutivos, haciendo que también hubiera un cruce entre persona (p) y ocasión (o), y persona cruzada con ocasión (r: [p x o]) fue utilizada para estimar los componentes de varianza (VCs) para aquellos aplicantes entrevistados en dos ocasiones. Se llevó a cabo un estudio D para examinar el impacto diferencial en la confiabilidad del aumento en evaluadores versus el número en aumento de las ocasiones de entrevista.

En este estudio, los componentes de varianza (VC) para el evaluador contienen influencia de los efectos no deseados/mezclados. El componente de varianza refleja el hecho de que el 23% del total de varianza es atribuible a las diferencias sistemáticas entre los entrevistados (varianza de “puntaje real”) en el constructo a medir. El restante 1% del total de varianza es atribuible a diferencias sistemáticas entre las ocasiones. Sin embargo, la influencia de la ocasión está mejor definida en la ecuación 1, la cual muestra el cálculo de la confiabilidad. Esta ecuación fue utilizada para calcular todos los coeficientes G, en donde se muestra que el estudio G aumenta a 0.73 por 9 ocasiones cada uno con un evaluador. Sin embargo, cuando el número de evaluadores aumenta por una sola ocasión, la credibilidad, estimada a 0.23 para un evaluador, incrementa a solamente 0.36 por nueve evaluadores.

59. Rosenfeld. J., Reiter, H. Trinh, K. y W. Eva, K. (2008). A Cost Efficiency Comparison Between The Multiple Mini-Interview and Traditional Admissions Interviews. Advances in Health Sciences Education 13, 43-58. DOI 10.1007/s10459-006-9029-z

El objetivo del estudio fue comparar los costos de la MMI solamente con aquellas técnicas de una entrevista de panel clásica dado que estas son las comúnmente más utilizadas en programas de medicina.

Hay 3 tipos generales de estaciones: discusión, habilidades interpersonales y de cooperación. La herramienta de evaluación se expandió a 12 estaciones. Típicamente,

cada circuito se recorre dos veces (uno en la mañana y otro en la tarde). Además, se separó a los candidatos de la mañana de los de la tarde para evitar fugas de información y que se alteren los resultados.

En la Mini entrevista múltiple, por cada estación hay 12 aplicantes. Suena una bocina para indicarles que inició la prueba (para que lean las instrucciones a la entrada) y dos minutos después vuelve a sonar para indicar que deben ingresar a la sala de evaluación. 8 minutos después suena otra bocina para indicar que se completó la prueba. Luego se dirigen a la otra estación y así sucesivamente hasta completar las 12. Los evaluadores constan de 2 minutos para evaluar en una rúbrica a los candidatos. La evaluación de 400 candidatos requiere de 31 MMI, circuito de 12 estaciones y 1 estación de descanso. Con 8 circuitos paralelos, una MMI de este tamaño se completa en 2 días con dos rondas por día.

Por otro lado, en la entrevista tradicional, el proceso requirió 3 entrevistadores por aplicante y una hora de tiempo de entrevista. La misma duró 40 minutos y 20 para que los entrevistadores comentaran al respecto. La evaluación de 400 candidatos requeriría 400 eventos, los mismos se distribuyeron a lo largo de 4 fines de semana. En los días entre semana, 16 equipos de entrevistadores de 3 entrevistadores hubieran alcanzado evaluar 6 candidatos.

Los resultados arrojaron que la creación de los materiales como las hojas de evaluación, checklists, instrucciones, etc, requieren de mucho tiempo y experiencia, ya que de esto depende el enfoque que se le dé a la evaluación. La creación y selección requiere como mínimo 3 horas. La creación de 24 estaciones para dos días requiere de 72 horas por persona. Lo relativo a las 800 horas del observador para montar una MMI para 400 candidatos, este costo representa un incremento de un 9% sobre el proceso de entrevista mismo. Se recomienda construir y mantener un banco de estaciones.

La MMI es considerablemente más eficiente en el número de horas que toma evaluar 400 candidatos. Relativo al costo de las entrevistas tradicionales utilizadas en McMaster, la MMI requiere 67% (800:1200) igual de horas de observación por aplicante

y 16% (66.7:400) igual de tiempo. A pesar de que la MMI evalúa más candidatos por hora, la preparación de la misma revela otro panorama.

Para la MMI a diferencia de la entrevista tradicional, se requirió de 5 personas por día para todo el proceso. La entrevista tradicional requeriría de 4 días para evaluar 400 candidatos, y la MMI solamente de 2. Para el equipo de preparación de la MMI se ahorra un 14% más que en la entrevista tradicional. Se debe considerar que por lo tiempos, existen gastos como meriendas, almuerzo, y bebidas, además de costos de infraestructura.

Un gasto importante para la mayoría de los programas de capacitación profesional, tanto financiera como términos de recursos humanos, es el proceso de entrevista utilizado para tomar decisiones de admisión. Aún así, la mayoría los programas ven esto como un costo necesario dado que la entrevista personal brinda una oportunidad para reclutar candidatos potenciales, mostrándoles lo que el programa tiene que ofrecer.

La MMI requiere mayores esfuerzos preparatorios y un mayor número de salas para llevar a cabo las entrevistas relativas a las entrevistas. Las desventajas de costo son establecidas por las MMI que requieren menos horas-persona de esfuerzo. En absoluto los costos variarán dependiendo de la institución, pero se espera que el marco presentado en este documento proporcione una mayor orientación con respecto a los requisitos logísticos y el presupuesto previsto.

60. Chávez, M. & Barrantes, M. (2014). Confiabilidad y validez de las listas de cotejos del Examen Clínico Objetivo Estructurado para el aprendizaje por competencias de Cirugía. Ciencia y Tecnología, 10(3), 115-128.

El objetivo del estudio fue evaluar el aprendizaje por competencias de cirugía de los estudiantes del sexto año de medicina, así como valorar el nivel de validez y

confiabilidad de las listas de cotejos de las estaciones del ECOE (examen clínico objetivo estructurado). Para esto participaron 86 estudiantes de medicina de 6to año y 18 docentes.

Se preparó dos estaciones por cada especialidad. Cada alumno fue evaluado mediante diez listas de cotejos correspondientes a diez estaciones de evaluación del ECOE. Había 5 estaciones activas, es decir, con profesor y paciente. Y 5 pasivas, es decir, con profesor y algún equipo mecánico de evaluación. Los estudiantes ingresaban uno por cada estación, y al final llenaban unas preguntas escritas en algunas estaciones.

El tiempo de rotación fue de 5 minutos, incluido el tiempo para leer instrucciones. Las listas de cotejos contenían opciones únicamente de “sí” o “no” según lo que observaba el profesor.

Con respecto a los resultados, el rendimiento por competencias quirúrgicas promedio es de 16.2 puntos en escala vigesimal. La confiabilidad varía de 0,33 hasta 0,90. En cuanto a la validez de los diferentes instrumentos, determinada mediante el coeficiente de correlación de Pearson, estos varían desde 0.65 hasta 1.00. Cuatro estaciones no alcanzaron niveles de confiabilidad. Las mismas fueron una de cirugía general la que se refiere a anamnesis del dolor abdominal; dos de otorrinolaringología, la prueba de Rinne y la prueba de senos para nasales y una de urología, la referida a la prueba de cólico renal. Sin embargo, se encuentra también que existen instituciones que aplican ECOE y reportan diferentes niveles de confiabilidad y validez de sus pruebas

61. Trejo, J., Martínez, A., Méndez, I., Morales, S., Ruiz, L., & Sánchez, M. (2014). Evaluación de la competencia clínica con el examen clínico objetivo estructurado en el internado médico de la Universidad Nacional Autónoma de México. Gaceta Médica de México, 150, 8-17.

El objetivo de la investigación fue evaluar la competencia clínica de estudiantes de medicina mediante el ECOE, antes y después del internado médico.

Con respecto a la metodología, el estudio comprendió dos etapas, pretest y postest. Se presentaron 315 internos al pretest y 10 meses después se aplicó el postest; de esos 315 estudiantes iniciales, 37 no se presentaron al postest, por no haber aprobado el internado o el examen profesional teórico.

La muestra final analizada fue de 278 estudiantes del internado médico de la Facultad de Medicina de la UNAM 202 (72.7%) internos mujeres y 76 (27.3%) internos hombres, que realizaron ambas evaluaciones y representaron el 39.2% de la población blanco. El instrumento de evaluación fue un ECOE de 18 estaciones, conformadas y validadas por un comité de expertos profesores de la Facultad de Medicina. Estos 18 profesores elaboraron las estaciones, que fueron validadas además por un grupo de seis médicos familiares y que además fuera posible evaluar en los 6 min de cada estación. Para incrementar la confiabilidad las listas de cotejo utilizadas fueron elaboradas por médicos especialistas en cada una de las áreas. Se seleccionaron pacientes reales y se conformaron estaciones de dos tipos: estaciones de procedimientos y estaciones de interpretación de exámenes de laboratorio y/o gabinete.

El ECOE se planeó para una duración de 2 h y cada una de las estaciones fue de 6 min, incluyendo dos estaciones de descanso. En las estaciones con pacientes, el examinador, además de calificar en la lista de cotejo, emitía una calificación global sobre las habilidades de comunicación interpersonal, mediante una escala global de 1 a 9, en la que 1 a 3 = insatisfactorio, 4 a 6 = satisfactorio y 7 a 9 = superior.

La calificación final del examen fue el resultado del promedio de calificaciones de todas las estaciones; la escala iba del 0 al 100% de aciertos. Se realizó un análisis estadístico con el paquete JMP (versión 8), estadísticas descriptivas, prueba t de Student para dos muestras correlacionadas, prueba ANOVA para tres o más muestras y prueba ANOVA anidada para las instituciones. Se calculó la d de Cohen como medida del tamaño del efecto en la puntuación global del ECOE.

Los resultados muestran que la confiabilidad con el alfa de Cronbach fue de, pretest de 0.62 y postest, de 0.64. La media global del ECOE pretest al principio del internado fue de 55.6 ± 6.6 y la media de la medición postest al final del internado, de 63.2 ± 5.7, con una diferencia absoluta del 7.6% ($p < 0.001$). El índice d de Cohen, como medición del tamaño del efecto en la puntuación global del ECOE, fue de 1.2.

El cambio absoluto de competencia clínica medido con el ECOE fue del 7.6%, lo cual, expresado en forma de tamaño del efecto con el índice d de Cohen, fue de 1.2. Esta magnitud de diferencia observada en términos porcentuales no parece ser muy elevada; sin embargo, el tamaño del efecto se puede interpretar como grande, de acuerdo con la clasificación de Cohen (> 0.8).

La mayor diferencia de medias pre y postest, que fue de 12.1, seguida por el área de cirugía. En contraste, el área de menor diferencia fue la de medicina interna. El componente con mayor diferencia en puntuación fue el de la interpretación radiológica, seguido de la exploración física, y el componente con menor diferencia fue el de la interpretación de los estudios de laboratorio.

62. Gamboa-Salcedo, T., Martínez-Viniegra, N., Peña-Alonso, Y., Pacheco-Ríos, A., García-Durán, R., & Sánchez-Medina, J. (2011). Examen Clínico Objetivo Estructurado como instrumento para evaluar la competencia clínica en Pediatría. Estudio piloto. Bol Med Hosp Mex, 68(3), 184-192.

El objetivo del estudio fue contar con un instrumento de evaluación de competencias que pudiera emplearse con fines de evaluación formativa y acumulativa y, al mismo tiempo, brindar a los profesores el adiestramiento en la elaboración de estaciones para evaluar competencias. La investigación se realizó en un hospital pediátrico de tercer nivel de atención que se aplicó a 20 (2 residentes de primer año, 7 de segundo y 11 de tercero) residentes en un estudio piloto. Fue evaluado por 20 especialistas

Previo a la realización del estudio se realizó un taller de capacitación a los evaluadores con una duración de 24 horas. Los mismos eran 20 especialistas. Cada estación duró 5 minutos y las listas de cotejos fueron creadas con opciones binarias para reducir la posibilidad de sesgo de los evaluadores.

Algunos aspectos relevantes para cada estación fueron el interrogatorio, la exploración física, la interpretación de estudios paraclínicos, la integración de diagnósticos, el plan de manejo integral y las habilidades para la comunicación interpersonal.

Los estudiantes rotaron por todas las estaciones, 20 en total, en donde había un examinador. Al finalizar la prueba, se le solicitó a los estudiantes sus opiniones respecto a la evaluación, y los observadores entregaban sus comentarios por escrito. Los profesores ponderaron los ítems de las listas de cotejo para que cada estación sumara 10 puntos. El análisis estadístico se realizó con el programa SPSS versión 15.0.

Con respecto a la duración total del examen fue de 2 h con 20 minutos, tiempo que había sido planeado. Por otro lado, dado que se trataba de evaluación pediátrica fue difícil contar con pacientes reales que se adaptaran a las necesidades del estudio y se contó sólo con un niño sano de 9 años. Las estaciones aprobadas de manera global fueron doce y las no aprobadas ocho. El promedio global de todas las estaciones fue de 6.53 (DE 0.62).

Los promedios por año de residencia fueron: R1 = 6.13 (DE 0.43), R2 = 6.26 (DE 0.60) y R3 = 6.76 (DE 0.59). La encuesta de salida aplicada a profesores y alumnos reportó que 60% de los participantes consideró que esta manera de evaluarlos era justa, 65% que era práctica y 65% la consideró útil para su vida profesional. De todos los alumnos 25% consideró que las instrucciones fueron insuficientes y 45% consideró que era una forma recomendable para medir sus conocimientos.