Poder predictivo de las pruebas en el rendimiento académico de médicos: un análisis correlacional del rendimiento de médicos en las Pruebas Saber (11 y Pro) y su rendimiento académico en los ciclos básico y clínico de un programa de medicina

Presentado por:

Gerardo Valencia Villa

Trabajo de investigación para optar al grado de Magíster en Educación

Director(a):

Elsa Lucía Escalante Barrios

Universidad del Norte

Maestría en Educación

Barranquilla

2022

Nota de aceptación:

Jurado

Jurado

Dedicatoria

“A mis ángeles, tanto en el cielo como en la tierra”

¡Todos creyeron en mí y fue más fácil…!

Agradecimientos:

Al departamento de Medicina de la Universidad del Norte y sus administrativos

Al comité de evaluación -curricular del Dpto. de Medicina

A Nelly Lecompte, Yolima Altamar Medina y Hernando Baquero L.

A Sonia Suárez Enciso, mi esposa

Tabla de Contenido

Lista de tablas

Lista de figuras

Resumen

Lograr que la trayectoria académica universitaria de los estudiantes de educación superior incida en el desarrollo de sus competencias profesionales es el mayor reto para los educadores. Esta investigación pretende determinar la relación entre el rendimiento en las pruebas Saber (11 y Pro) presentadas por estudiantes y su rendimiento académico en los ciclos básico y clínico de un programa de medicina. Estas últimas, conocidas como pruebas de progreso, evalúan la consecución de los resultados de aprendizaje de la educación médica de pregrado en el plan de estudios. Esta investigación cuantitativa correlacional no causal se realizó con datos secundarios académicos y socioeconómicos de dos cohortes de estudiantes: 150 que tomaron la prueba Saber Pro en 2018 y 197 que lo hicieron en 2019, para una muestra total de 347 participantes. La asociación entre las mediciones de desempeño fue medida con el coeficiente de Pearson, mientras que el valor predictivo se obtuvo por regresiones lineales. La mayoría de los coeficientes de correlación resultaron ser los esperados, con 12 de 15 estadísticamente significativos. La prueba Saber 11 predice significativamente el rendimiento de la mayoría de las pruebas subsiguientes; la de ciclo básico predice significativamente el rendimiento de todas sus subsiguientes pruebas. El examen de fin de ciclo clínico arrojó resultados no esperados. Se determinó que las pruebas de progreso, como el examen de fin de ciclo básico, poseen un poder predictivo significativo sobre las pruebas subsiguientes y su validez predictiva es evidenciable cuando su análisis está acompañado por evaluaciones estandarizadas, como las pruebas Saber aplicadas por el Ministerio de Educación colombiano.

Palabras clave: Pruebas de progreso (ciclo básico y ciclo clínico); Poder predictivo de pruebas; Saber 11; Saber Pro; Rendimiento académico de médicos.

Abstract

Attaining the influence of higher education students' academic trajectory on the development of their professionals' competencies is our greatest challenge as educators. This research aims to determine the relationship between students' performance in Saber tests (11 and Pro) and their academic performance in basic and clinical tests of a medicine program. The latter, known as progress tests, assess the achievement of undergraduate medical education learning outcomes. This non-experimental correlational research was carried out with secondary academic and socioeconomic data from two cohorts of students: 150 who took Saber Pro in 2018 and 197 who did so in 2019, giving a total sample size of 347. The association between test performances was measured using Pearson´s coefficient, while the predictive coefficients were estimated using linear regressions. Most of the correlation coefficients showed to be as expected, with 12 out of 15 statistically significant. Saber 11 significantly predicts performance of most of the subsequent tests; performance in basic test significantly predicts performance in all its subsequent tests. Results involving performance in clinical test were not as expected. Performance in progress assessment such as the basic test has a significant predictive power over the subsequent tests; its predictive validity is evident when the analysis includes standardized tests results such as those of the Saber applicability for the Colombian Ministry of Education.

Keywords: Progress tests (basic cycle and clinical cycle); Predictive power of tests; saber 11; Professional saber; Medical Graduates; Academic performance.

1. Introducción

La educación médica como proceso debe formar médicos resolutivos a nivel asistencial y social. Durante el trasegar del plan de estudios, el aprendiz médico debe desarrollar sus habilidades investigativas y de autogestión del aprendizaje, además de adquirir o perfeccionar su capacidad de interrelación humana (Epstein, 2007). El Estado, en representación de la sociedad que lidera, debe supervisar directa e indirectamente los compromisos enunciados.

La globalización en salud coloca otra arista al deber ser de la educación médica, ya que la sociedad representada por el Estado, las colegiaturas médicas, las asociaciones de especialistas y las instituciones de educación superior deben velar por el proceso de desarrollo de las competencias médicas, para hacerlas holísticas y auditables. Simultáneamente, es su deber participar en la auditoría de dichas habilidades (Patterson et al., 2018). Diferentes pruebas a nivel mundial, como el USMLE en Estados Unidos y sus tres *steps*, el MIR en España o el Examen de calificación del Consejo Médico de Canadá (MCCQE), se encargan de esa auditoría holística que permite a cualquier egresado médico de una facultad en el planeta ser admitido a un programa de especialidad en países referentes en esta materia. Algunas pruebas del mismo tipo se dan en Latinoamérica. El EUNACOM en Chile; un examen estatal en el caso de México; la homologación obligatoria de los títulos o competencias médicas (con o sin pruebas) realizada por universidades estatales, en los casos de Argentina y Paraguay.

Además, las facultades de medicina desde la supervisión interna de sus procesos, realizan pruebas en diferentes momentos del proceso educativo. A estas pruebas les llaman “exámenes comprensivos o comprehensivos”, y todas refieren a pruebas de supervisión de lo aprendido durante la carrera o un segmento de la misma (Taber et al, 2020). En Colombia, además de lo mencionado a nivel internacional, el médico en formación debe realizar sendas

pruebas oficiales de Estado para su admisión al pregrado (Saber 11) y otra al final de la carrera (Saber Pro).

Las instituciones universitarias colombianas también reciben información de sus estudiantes inscritos y la complementan con detalles socioeconómicos generales al matricularse en sus facultades o departamentos. Es el caso del departamento de medicina, donde cada estudiante –desde su ingreso– recibe la información detallada de sus pruebas Saber 11; también de su procedencia escolar, estrato socioeconómico y el origen de su financiación de matrícula. Luego, para optar por el título de médico se debe obtener un promedio ponderado y presentar la prueba Saber Pro. El promedio aparecerá en el acta de grado. Todos los datos mencionados se usarán en la presente investigación, para lo cual se obtuvo permiso de la división de salud y del departamento de medicina del programa de la universidad respectiva.

El análisis estadístico correlacional entre las diferentes pruebas que realiza un médico desde su admisión hasta su egreso de la facultad, mostrará el nivel de coherencia secuencial entre ellas en los correspondientes momentos de la vida académica de dos cohortes de médicos. Las variables de contexto previo, tanto sociodemográfico como académico de cada individuo, serán usadas para una mejor correlación. El objetivo de la investigación es medir la capacidad predictiva del rendimiento global en las diferentes pruebas entre sí y su correlación con los subcomponentes de cada una. Los métodos cuantitativos de investigación fueron la base para el buen uso de los datos obtenidos en búsqueda del objetivo enunciado.

En este estudio se confirmó la coherencia entre las pruebas estandarizadas estatales (Saber 11 y Saber Pro) a las que estuvo expuesta la población de estudio. Igualmente, se obtuvieron correlaciones importantes entre las variables consideradas. Todo esto servirá para apoyar la discusión a partir de investigaciones homólogas, así como las recomendaciones curriculares que se planteen, fundamentadas en las conclusiones del presente trabajo.

2. Justificación

Los organismos multilaterales, como la Organización para la Cooperación y el Desarrollo Económico (OCDE), han tenido un gran impacto en la educación, al punto que las pruebas estandarizadas estatales de cada país (Saber 3°, 5°, 9°, saber 11° y Saber Pro, en el caso de Colombia) están en la misma dinámica y son coherentes con pruebas estandarizadas internacionales como: Programme for International Student Assessment (PISA), Segundo estudio regional comparativo y explicativo (SERCE) y Trends in International Mathematics and Science Study (TIMSS). De las pruebas internacionales aplicadas en Colombia, la de mayor connotación para generar políticas públicas en educación e indicar status para el país son las PISA, por lo que hoy la necesidad del gobierno nacional de hacer parte de la OCDE está cumplida (Sánchez, 2014).

Para el autor del tratado de la formación médica en Colombia, el objetivo general que engloba a las pruebas estandarizadas es medir la calidad de la educación (Sánchez, 2014); sin embargo, está claro que la calidad de la educación trasciende los exámenes estandarizados y sus resultados. Estas pruebas deben ser fundamentalmente un instrumento para los planes de mejoramiento a nivel institucional y una propuesta de mejora en la medición del proceso de enseñanza-aprendizaje.

La comunidad en la que transitamos es universal. Como resultado de los últimos 50 años y los avances tecnológicos ocurridos se ha reforzado el concepto de sociedad del conocimiento. Gracias al cambio veloz en las formas de comunicación todo es fácilmente compartido, desde el conocimiento innovador hasta las problemáticas sociales. Esto último incluye la creación de entes sociales de gran responsabilidad y actuación efectiva desde el proceso educativo, donde el desarrollo de la base cognitiva y las habilidades profesionales deben forjar individuos resolutivos que socialmente sean adaptables a lo rápido y plural de

los cambios (American Educational Research Association et al., 2014; Goldhaber & Özek, 2019).

Actualmente, no es posible imaginar un profesional “conducido” para conocer y solucionar problemas específicos. Debe ser formado, como mínimo, con la respectiva transferibilidad de los conceptos que lo caracterizarán como ciudadano y profesional de la llamada sociedad universal (Epstein, 2007). La educación debe formar seres humanos capaces de autoaprender a través del contexto brindado por la actualidad del mundo (Goldhaber & Özek, 2019; McManus, Woolf et al., 2013), impregnados de los conceptos básicos y las habilidades fundamentales de cada profesión, sumado a un excelente desarrollo de sus habilidades de pensamiento básico y un aceptable nivel de crecimiento en las habilidades de pensamiento avanzadas (Anders et al., 2019).

En las instituciones educativas de alto nivel en el mundo, la estimulación a las capacidades de autoaprendizaje y transferibilidad de lo aprehendido se han vuelto pilares fundamentales de carácter transversal en los planes de estudios (Pinilla & Parra, 2009; Valencia et al., 2020). La actividad de formación investigativa, el emprendimiento, las actividades colaborativas, la simulación (situaciones teóricas y prácticas) y las habilidades comunicativas (orales y escritas), entre otras, se evidencian actualmente como tendencias metodológicas comunes en la educación superior (Wiley, 2018). Lo proyectado y lo realizado llega a ser coherente en muchas facultades de diferentes ubicaciones geográficas. Sin embargo, la supervisión y evaluación de estos objetivos no han sido procesos fáciles; los exámenes oficiales estandarizados para profesionales han sido un muy buen intento, especialmente los que han sido engranados a pruebas homólogas realizadas en la educación primaria y secundaria para los mismos individuos (Aparicio & Valencia, 2020).

En ese sentido, el marco pedagógico requerido para la consecución de lo propuesto a nivel mundial está basado en el anclaje de los resultados de aprendizaje del egresado en

asignaturas, áreas o fases de los planes de estudio. Desde el inicio de la vida universitaria se tributa y/o valora el desarrollo de las competencias esperadas en el profesional. Las diversas estrategias pedagógicas y didácticas utilizadas para tales fines se plasman a través de las tendencias metodológicas actuales (Wiley, 2018).

Es claro entonces, que el médico recién egresado debe insertarse a la sociedad con la capacidad de desempeñarse como un profesional resolutivo (ASCOFAME, 2017). Lo más cercano a este último indicador es la prueba estandarizada que oficialmente aplica el Estado a todos los profesionales próximos a su egreso (Saber Pro). Esta prueba está diseñada para ser comparable con una prueba previa (Saber 11), cuyos resultados se utilizan como base fundamental para que un estudiante sea admitido al programa de salud, respectivamente (Aparicio & Valencia 2020; Pinilla & Parra, 2009).

La supervisión de las modificaciones individuales, colectivas y su repercusión social son indispensables en el ámbito educativo, especialmente en estos tiempos donde la generación de datos es abundante y su manejo es importante. Los entes estatales y privados que lideran, tanto la estandarización como la acreditación de los procesos educativos, facilitan dicha revisión y comparación continua a través de pruebas de suficiencias en diferentes niveles del proceso educativo (educación primaria hasta la inserción social del profesional); en algunos casos, hasta la refrendación de dichas capacidades profesionales u ocupacionales son supervisadas (Pinilla & Parra, 2009). El Observatorio de Educación para el Caribe colombiano de la Universidad del Norte, apoyado por el Ministerio de Educación Nacional de Colombia (MEN) y el Instituto colombiano para la Evaluación de la educación (ICFES), es un ejemplo de comunión estatal y privada en esta búsqueda, al liderar estudios referentes al valor agregado de la educación superior con resultados que han motivado investigaciones como la presente.

Buscando conseguir los Resultados de aprendizaje esperados (RAE) para un médico recién egresado, cada programa de medicina utiliza métodos de verificación, que van desde los clásicos (promedio ponderado de notas durante la carrera y al final de la misma) hasta diferentes mecanismos de evaluación. De esta manera, se intenta evidenciar de manera objetiva una aproximación a la realidad del estudiante durante su pregrado. La autonomía universitaria y la diversidad curricular permitida, dentro de los límites del registro calificado y la acreditación de alta calidad, se encuentran entre los múltiples mecanismos utilizados por los programas de salud en nuestro territorio para la supervisión y comprobación mencionada.

Los exámenes de fin de ciclo en el programa de medicina de la Universidad del Norte (si bien provienen de las ingenierías en su intento de estandarizar la consecución de competencias comunes a todas sus modalidades) se han aplicado y han sido fundamentales para la supervisión curricular de resultados de aprendizaje. Los buenos resultados en los ámbitos de retroalimentación para el programa, los docentes y especialmente los estudiantes, han motivado algunas modificaciones curriculares (nuevas asignaturas electivas para reforzar áreas) y metodológicas (casos clínicos transversales en las asignaturas de ciclo básico para aumentar la integración). Sin embargo, las limitaciones en la estandarización de estas pruebas no han permitido mayor lucro académico e información estadística relevante para la toma de decisiones basadas en la objetividad del análisis de sus resultados (Martínez et al., 2015).

Thorndike et al. (1991) declaraban en su quinta edición de "*Measurement and evaluation in psychology and education*" que en el área de educación es fundamental conocer la estimación de la asociación entre las diversas evaluaciones a las que se expone el individuo dentro de un sistema educativo: "Las pruebas existen para ayudar a tomar decisiones y la calidad de esas decisiones depende del uso informado de la información de las pruebas"(s. p.); especialmente, cuando con dichas pruebas se busca evidenciar el desarrollo de

competencias propuestas en el estudiante a través de la estimulación de procesos cognitivos básicos y complejos.

Es casi un deber, por el énfasis que realiza en educación social el grupo de investigación Cognición y Educación, aprovechar los datos generados por los estudiantes del programa de medicina de la Universidad del Norte. Estos estudiantes, además de la información estatal que traen al ser admitidos y la que deben completar al egresar del programa, realizan pruebas de suficiencia para el paso de los ciclos (básico y clínico) al interior de su carrera (Martínez et al., 2015). Estos exámenes a mitad y final de carrera se convierten en una forma de supervisión, pertinente y paralela, del "plus institucional y programático"; por su posible capacidad predictiva del rendimiento futuro del egresado.

El alcance de un análisis correlacional entre las pruebas estandarizadas antes y después de transcurrido un proceso educativo superior, se podría evidenciar de forma indirecta en la consecución de los denominados *graduate learning outcomes* (Ferrer y Arregui, 2003). Permitiendo además la observación de los "pluses" que la institución educativa le imprime a su egresado, inclusive mostrando una revisión paracurricular de los RAE de curso. Estos últimos, para la mayoría de planes de estudio en las carreras de salud se pueden resumir en los ciclos básico (preclínico) y clínico (Sánchez, 2014).

El departamento de medicina de la Universidad del Norte realiza una prueba obligatoria por cada ciclo del plan de estudios (básico y clínico); allí, se pretende evidenciar su suficiencia en las áreas constituyentes de cada ciclo. Estas pruebas se vienen realizando hace más de una década y han tenido un rol de supervisión curricular interesante. Los resultados de dichas pruebas proporcionaron al comité curricular del departamento instrumentos para estimular avances pedagógicos en el *pool* docente, la revisión de la integración de conceptos y la vigilancia a mediano/largo plazo de la pertinencia de lo aprendido (Martínez et al., 2015).

Las correlaciones a obtener en esta investigación, agregando las pruebas de fin de ciclo y el promedio de fin de carrera, entregarán la posibilidad de ser consideradas como evidencias para validación y mejoramiento de los exámenes de suficiencia desarrollados por los estudiantes del programa de medicina de la Universidad del Norte; obviamente, en caso de ser positiva la correlación. Además, las pruebas estatales que deben presentar estos estudiantes, en los extremos temporales de la correlación, poseen subcomponentes comparables (líneas de horizonte) que podrían vislumbrar el "plus" atribuible a la formación académica institucional a la que se expone al médico durante su pregrado.

La información contextual básica (socioeconómica) de los sujetos de investigación, también supone un boceto del "valor agregado programático"; éste podría evidenciar fortalezas a mantener y debilidades a convertir en oportunidades de supervisión (*assessment*) del proceso de enseñanza aprendizaje, coherente con el proceso actual de Valoración Institucional y Seguimiento del Aprendizaje (VISA), al que ingresará próximamente el departamento de Medicina.

Para esta investigación, tanto los procesos de autoevaluación para la acreditación del programa de medicina como la supervisión curricular continua, han permitido el registro habitual de los datos de académicos (enseñanza-aprendizaje-evaluación). Esto facilita la obtención de datos de fuentes secundarias, garantizando la accesibilidad a la información de manera organizada, anónima y permitiendo el manejo ético del presente trabajo.

3. Planteamiento del Problema

El uso de los puntajes de las pruebas como medida de desempeño en la responsabilidad educativa se volvió cada vez más popular a nivel mundial. En Estados Unidos, desde la promulgación de la Ley "*Que ningún niño se quede atrás"* de 2001 *(NCLB)*, los fondos federales para las escuelas dependían de la implementación de un sistema estatal de responsabilidad escolar que evalúa el progreso del estudiante continuamente (Goldhaber & Ozek, 2019; Jornet, 2017). Aunque, ese tipo de políticas reactivas debilitan la esencia del fin buscado por dichas pruebas (Jornet, 2017).

La validez predictiva del rendimiento en las pruebas se fundamenta en el uso de toda la evidencia empírica relevante (Goldhaber & Ozek, 2019). Sin embargo, para las medidas de éxito de los estudiantes también debe considerarse la validez predictiva de las evidencias académicas no resultantes de test (APA, 2013); aunque, dicha proyección de éxito de los puntajes de las pruebas podría ser más fuerte en algunos contextos que en otros. Ejemplo de criterio contextual a tener en cuenta puede ser la característica del estudiante (McManus, Woolf et al., 2013); también, la carrera elegida y las consecuencias de la prueba (Kennet-Cohen et al., 2016).

En Reino Unido se realiza una prueba de admisión cognitiva objetiva estándar para el área de salud. Esta prueba tuvo una alta validez predictiva con respecto al desempeño del primer año de la escuela de medicina, donde se examinan las ciencias básicas médicas (Kennet-Cohen et al., 2016; McManus, Dewberry et al., 2013), que conforman la columna vertebral académica en medicina (McManus, Dewberry et al., 2013). La cuantificación cognitiva a través de una prueba de aptitud objetiva estándar es el aspecto más válido del proceso de selección (Kennet-Cohen et al., 2016). Su estandarización permite la comparabilidad entre solicitantes de diversas escuelas secundarias y diferentes cohortes (Goldhaber & Ozek, 2019; Kennet-Cohen et al., 2016), lo que el Bagrut (examen del sistema

educativo israelí), no hace. Su menor validez predictiva podría explicarse, al menos parcialmente, por el hecho de que ese sistema educativo refleja aspectos variables como la diversidad de la sociedad israelí (Kennet-Cohen et al, 2016); diferentes sectores de la población (laicos, religiosos y árabes) asisten a diferentes escuelas que ofrecen diferentes planes de estudio. Tales variaciones en la fuerza de las correlaciones entre cohortes e instituciones es un hallazgo típico (Coates, 2008; Edwards et al., 2013; Julian, 2005). Estas variaciones enfatizan los peligros de generalizar los resultados de los estudios de validez predictiva sin reconocer las diferencias en los contextos en los que se realizan (Kennet-Cohen et al., 2016). Recolectar la mayor cantidad de información contextual es muy importante para la posterior interpretación paralela de los niveles de correlación observables (Kennet-Cohen et al., 2016).

Es claro que el desempeño del estudiante en el primer año de medicina refleja predominantemente el conocimiento fáctico, privilegiado en pruebas cognitivas estandarizadas (Kennet-Cohen et al., 2016; McManus, Dewberry et al., 2013). Sin embargo, algunas asignaturas incluyen cursos que se centran en las habilidades interpersonales, donde se ha comprobado que el desempeño puede predecirse mediante medidas no cognitivas (Lievens & Sackett, 2009). Estas últimas podrían ser similares a las pruebas psicotécnicas y entrevistas dispuestas por algunas instituciones en sus mecanismos de admisión; inclusive podrían equipararse a las áreas genéricas de las denominadas pruebas Saber 11 y Saber Pro en Colombia.

El Ministerio de Educación Nacional colombiano y las Instituciones de educación superior (IES) invierten recursos y tiempo en el diseño de las pruebas Saber 11 y en pruebas de admisión específicas. La finalidad de dicho esfuerzo es verificar la consecución de competencias mínimas a través de la educación secundaria y, consecuentemente, clasificar a los candidatos a admitir en los programas de educación superior (ICFES, 2018). La mayoría

de facultades de medicina, para sus nuevos aspirantes utilizan la prueba Saber 11 como base del proceso de admisión (requisito casi único) (ICFES, 2018). La aplicación de sendos exámenes oficiales estandarizados trajo consigo la tentadora necesidad de aprovechar su alineación longitudinal para realizar correlaciones entre los resultados obtenidos por un estudiante en ambas pruebas. Su utilidad no declarada, pero evidente, es la clasificación de desempeño entre las instituciones de educación secundaria y superior (ICFES, 2018; Pinilla & Parra, 2009).

En consonancia con lo anterior, el Observatorio de Educación para el Caribe colombiano de la Universidad del Norte realizó en 2019 el análisis del valor agregado de las universidades colombianas, tomando como base la información proporcionada por el ICFES de 293.100 estudiantes universitarios evaluados en Saber Pro entre 2016 y 2018, cuyos resultados en Saber 11 corresponden al periodo de 2006 a 2014. En dicho estudio, las tendencias generales mostraron que 3 de cada 4 estudiantes obtuvieron mejores resultados a los proyectados para su respectiva prueba Saber Pro, evidenciando que el paso por la universidad aportó “valor” a sus competencias de lectura crítica o razonamiento cuantitativo (Pinilla & Parra, 2009). Esto demuestra el aporte significativo de las universidades en la formación de ciudadanos más competentes.

Recientemente, la ASCOFAME (2017) enfatizó sus propósitos en dirección a promulgar la formación de un médico general resolutivo. Aunque de la mano del Ministerio de Educación Nacional ya se venía estimulando y verificando la calidad de los programas de educación médica, esta directriz acelera y reenfoca los procesos de autoevaluación, con miras a la acreditación de los programas de medicina del país, en búsqueda de mejoras que los lleven a la certificación de alta calidad.

En esa misma línea, el programa de medicina de la Universidad del Norte aplica exámenes de suficiencia en los dos ciclos que componen su plan de estudios, básico y clínico.

Estos fueron implementados desde el inicio del penúltimo ajuste curricular realizado. No obstante, en las primeras cohortes se tomó como prueba piloto y solo a partir de la cohorte de estudiantes que iniciaron estudios en 2006 se aplicó estrictamente el reglamento (Martínez et al., 2015).

Los exámenes de suficiencia de ciclo básico y clínico aparecen desde entonces como asignaturas y poseen una normativa específica para su aprobación o desaprobación (Martínez et al., 2015). Estos exámenes fueron evolucionando en su nivel de estandarización y simultáneamente se convirtieron en puntos de retroalimentación curricular. Dicha retroalimentación es válida y explicable por su ubicación en el plan de estudios, evidenciándose dos momentos fundamentales del mismo (Martínez et al., 2015).

Definitivamente, la medicina como programa educativo no ha escapado al interés de evaluar el "valor agregado" por la institución a la formación del estudiante y su inminente inmersión a la sociedad. La retroalimentación curricular en el programa de medicina de la Universidad del Norte está cimentada en la valoración secuencial y continua de resultados de aprendizaje esperados en el médico egresado, que se obtienen a través de los diferentes momentos de conceptualización de la información del rendimiento académico de sus estudiantes (promedios ponderados por semestres, exámenes de suficiencia de fin de ciclo básico y clínico, comité de evaluación y comité curricular).

En Colombia, la verificación de las competencias profesionales de un recién egresado es la prueba estandarizada que oficialmente aplica el Estado a todos los profesionales próximos a su egreso (Saber Pro). Esta hace parte de un sistema de pruebas estandarizadas que supervisan los diferentes niveles de educación (Ley número1324 de 2009). Sin embargo, solo a partir del 2013, el Consejo Nacional de Acreditación (CNA) estableció los resultados de las pruebas Saber Pro como un insumo fundamental para el Aseguramiento de la calidad educativa, desde la evaluación y la mejora continua (CNA, 2013). En la literatura, se conocen

correlaciones, entre el rendimiento en Saber 11 al final de la educación secundaria (usada como prueba admisión universitaria) y el resultado obtenido en Saber Pro (usado como valoración de egreso universitario) (Aparicio & Valencia, 2020; Castro et al., 2018). Sin embargo, la correlación del rendimiento del egresado de medicina en la Universidad del Norte a través de las variables mencionadas, sumándole las pruebas de progreso intra-plan de estudio, no ha sido documentada; el único ejercicio investigativo publicado relacionado con las pruebas de fin de ciclo básico y clínico del programa es del año 2015 y tuvo su enfoque desde la revisión curricular interna (Martínez et al., 2015).

Las características de los exámenes de fin de ciclo básico y clínico del programa de medicina mencionado son únicas y han tenido una evolución de madurez cronológica y académica. Ambas pruebas son asignaturas que hacen parte del plan de estudios y tienen cada una prerequisitos y criterios de acceso como incidencia para poder continuar los estudios, cuando no son aprobadas. Las posibilidades de predicción válida de cada uno de los momentos evaluativos del egresado en el desarrollo del programa de pregrado, sumada al potencial de retroalimentación efectiva del currículo programático para el futuro médico, son muy atractivas especialmente si tenemos en cuenta la estandarización y coherencia entre las pruebas estatales disponibles. También, y no menos importante, la transferibilidad del modelo observado y no evaluado en conjunto es un propósito a mediano plazo que puede convertirse en un proyecto institucional.

Finalmente, la mayoría de investigaciones nacionales e internacionales análogas comparan el rendimiento del egresado en los extremos (inicial y final) de la carrera de medicina (Castro et al., 2018; Coates, 2008; Gil et al., 2013); sin embargo, generalmente no lo correlacionan con pruebas homologables obtenidas en la carrera. En algunas oportunidades se han presentado investigaciones donde realizan la comparación con pruebas externas específicas tipo USMLE Paso 1 (Keltner et al., 2021; Mc Manus, Woolf et al., 2013;

Torre et al., 2020), o no específicas tipo TOEFL u otras pruebas no cognitivas (Kennet-Cohen et al., 2016).

3.1. Pregunta problema

En el contexto del programa de medicina descrito sería importante responder a la siguiente pregunta: ¿*Existe una asociación positiva entre el rendimiento en las pruebas Saber (11 y Pro) de médicos y su rendimiento académico en los ciclos básico y clínico de un programa de medicina*?

3.1.1. Preguntas problema específicas

- ¿Existe una asociación positiva entre el rendimiento en las pruebas Saber (11 y Pro) de médicos de un programa de medicina?
- ¿Existe una asociación positiva entre el rendimiento académico en los ciclos básicos y clínicos de médicos de un programa de medicina?
- ¿Existe una asociación positiva entre el rendimiento en las pruebas Saber 11 de médicos y su rendimiento académico (promedio ponderado a final de carrera) de un programa de medicina?
- ¿Existe una asociación positiva entre el rendimiento en las pruebas Saber 11 de médicos y su rendimiento académico en los ciclos básico y clínico (exámenes de suficiencia de fin de ciclo básico y clínico) de un programa de medicina?
- ¿Existe una asociación positiva entre el rendimiento en las pruebas Saber Pro de médicos y su rendimiento académico (promedio ponderado a final de carrera) de un programa de medicina?
- ¿Existe una asociación positiva entre el rendimiento en las pruebas Saber Pro de médicos y su rendimiento académico en los ciclos básico y clínico (exámenes de suficiencia de fin de ciclo básico y clínico) de un programa de medicina?

4. Marco Teórico

El uso de los puntajes de las pruebas estandarizadas para la evaluación de programas e instituciones educativas se volvió cada vez más popular a nivel mundial. Una muestra palpable de esto es que los fondos federales para las escuelas de los Estados Unidos dependen de la implementación de un sistema estatal de responsabilidad escolar que evalúa el progreso del estudiante continuamente (Goldhaber & Özek, 2019; Jornet, 2017).

Las pruebas estandarizadas son también insumo clave para la evaluación y clasificación de instituciones educativas en los rankings internacionales y nacionales en donde toman en cuenta exámenes de ingreso, exámenes de fin de carrera y promedio académico de los estudiantes, entre otros elementos. La Universidad de Harvard es un buen caso para ilustrar este argumento. Esta reconocida institución educativa es la sexta mejor universidad del mundo y encabeza la lista de las mejores escuelas de medicina del mundo en el año 2021, según los rankings de las principales universidades del Times Higher Education (The World University Rankings, 2019). Esta institución se ubica en la cima de la investigación y las innovaciones médicas mundiales. La escuela de medicina de Harvard acepta 3 estudiantes de cada 100 solicitantes (la tasa de aceptación es del 3,4%) y reporta un promedio de Grade Point Average (GPA) de 3.93, ocupando el primer lugar del ranking de escuelas de medicina de US News para 2021. Mientras que la facultad de la universidad John Hopkins es la número 2 en los Estados Unidos y aunque tiene un GPA de pregrado de 3.94 se posiciona detrás de la Universidad de Harvard en ese ranking. En este sentido, el GPA es un criterio de importancia, pero no es el único factor tenido en cuenta para estas clasificaciones (The World University Rankings, 2019).

De manera más frecuente, los resultados de las pruebas estandarizadas además de ser utilizadas como criterios de evaluación para programas e instituciones educativas en los procesos de financiación y clasificación en rankings, también son una herramienta para el

seguimiento de la formación de estudiantes y rendimiento académico, así como análisis curriculares (Sánchez-Mendiola et al., 2019). Aunque en la literatura se encuentra la valoración de asociaciones entre variables lógicas en el trascurso del plan de estudio (rendimiento en: asignaturas básicas, clínicas, pasantías prácticas y promedios de final de carrera) con variables que evidencian el comportamiento previo y posterior de los mismos individuos (rendimiento en: procesos de admisión, pruebas estandarizadas para verificación de capacidades laborales, pruebas básicas y clínicas para admisión a posgrados), entre las escuelas de medicina ganadoras del premio ASPIRE a la excelencia académica 2018 (Aga Khan University Medical College (AKU-MC), Pakistán; Southern Illinois University School of Medicine (SIUSOM), USA; University of Leeds School of Medicine, UK), se proclama la necesidad de analizar nuevas formas para examinar los aspectos específicos que podrían contribuir al éxito en el rendimiento de los exámenes para la obtención de la licencia profesional (Beason et al., 2019) y el desempeño ocupacional del médico general o especializado (Curtis & Smith, 2020; Greatrix et al., 2021; MacKenzie et al., 2016; Schreurs et al., 2020; Torre et al., 2020).

El nivel de aprendizaje evidente en los estudiantes de programas de medicina pudiera ser predecible por exámenes estandarizados. Esta hipótesis ha tomado fuerza en los diferentes escenarios internacionales de educación médica en los últimos años, acompañada de una reflexión acerca de modelos alternativos de evaluación que complementen estos procesos, tipo la prueba de admisión a la Facultad de Medicina de EE. UU. (MCAT). Inicialmente, dar respuesta a esta inquietud plantearía orientaciones válidas para una continua y efectiva revisión curricular que garantice una actualización de las directrices en educación médica coherente con el contexto temporal y social. No solo debería influir en mejorar el plan de estudios para formar médicos, sino en la orientación, recertificación y actualización de la

educación médica continua del egresado. La era de la analítica de los datos ya está presente y la educación médica se está quedando rezagada (Changiz et al., 2019).

4.1 Asociación entre promedio y evaluaciones de aprendizaje en estudiantes de medicina

La evaluación por competencias en la educación médica ganó terreno a nivel mundial en las últimas décadas, reemplazando la evaluación meramente cognitiva en el campo básico y la evaluación aplicada o de revisión de protocolos y técnicas en el área clínica. La coherencia entre la evaluación de los RAE y los perfiles profesionales u ocupacionales promovidos para el egreso, ha liderado la supervisión al interior de los diversos programas de medicina y sus planes de estudios (Frank et al., 2020; Taber et al., 2020). El resultado esperado de egreso en la educación médica internacional es un médico resolutivo y con capacidad de auto-aprender e investigar en beneficio del ser humano, ojalá apoyado por el meta-aprendizaje como estrategia. (Sánchez-Mendiola et al., 2019).

La evaluación del progreso académico en currículos tan centrados en el estudiante sigue siendo un desafío. Algunos investigadores en Estados Unidos han intentado correlacionar evaluaciones internas de facultades de medicina, tanto las que evidencian progresos por áreas del conocimiento o asignaturas (Greatrix et al., 2021; Wang et al., 2021) como las que sintetizan procesos básicos y clínicos con el aprendizaje esperado al final de carrera y/o su desempeño práctico, al final de la formación habilitante (Taber et al., 2020; Tamblyn et al., 2002), la mayoría con resultados estadísticamente poco significativos. Sin embargo, Wang et al. (2021) identificaron cuatro estados latentes con base en los resultados de las pruebas de progreso de los 358 estudiantes matriculados en la facultad de medicina con puntajes en el USMLE 1: Principiante, Principiante avanzado I, Principiante avanzado II y en estado competente. Al final del primer año, los estudiantes que se predijo permanecerían en el estado de novato tenían puntajes medios del Paso 1 más bajos, en comparación con los del estado competente (209, SD = 14,8 versus 255, SD = 10,8 respectivamente) y tuvieron más

fallas en el primer intento (11,5% frente a 0%). En el análisis de regresión encontraron que al final del primer año sí había un 10 % más de posibilidades de permanecer en el estado de novato. Las puntuaciones del Paso 1 se pronosticarían 2,0 puntos más bajas (IC del 95 %: 0, 85–2, 81 con P < 0, 01); mientras que un 10 % más de probabilidad en estado competente. Las puntuaciones del Paso 1 se pronosticaron 4, 3 puntos más (IC del 95 %: 2,92–5,19 con P < .01). Al final del segundo año de la escuela de medicina se encontraron hallazgos similares. Los investigadores concluyeron que el uso del modelo de cadena de Markov para analizar el rendimiento de la prueba de progreso longitudinal, ofrece un método de estimación flexible y efectivo que podría ayudar a identificar estudiantes que corren el riesgo de reprobar el examen de licencia y pueden beneficiarse del apoyo académico específico. Lo anterior es lo más parecido a un modelo predictivo para el comportamiento profesional de egresados médicos.

El uso eficiente de estrategias de aprendizaje en las facultades de medicina asegura el desarrollo normal de la esperada adquisición de contenidos, lo que debería redundar en mayor desarrollo de sus habilidades de pensamiento básico y avanzado, plus atribuible a los cursos superados en la institución de educación superior (Nimkuntod & Tongdee, 2016). Las diversas investigaciones que relacionan el GPA y los avances en los RAE alcanzados por los estudiantes de medicina muestran mejores niveles de alcance, sin llegar a ser estadísticamente significativas (Krupat et al., 2017; Nimkuntod & Tongdee, 2016).

En la literatura los datos varían, se pueden ubicar desde los que muestran cierta coherencia o validez predictiva hasta los que, en definitiva, no evidencian coherencia alguna. Diversas evaluaciones de aprendizaje ubicadas en diferentes momentos del plan de estudio, inclusive las realizadas para áreas específicas (biología, bioquímica, morfología), han sido comparadas con el GPA (Jiraporncharoen, et al., 2015; Nimkuntod & Tongdee, 2016).

En el mismo orden de ideas, se destacan estudios como el de Krupat et al. (2017) donde se observó que de los 164 estudiantes cuyo desempeño en los exámenes del plan de estudios previo a la pasantía fue sólido (es decir, sin apariciones en el cuartil inferior), el 57 % (94/164) se encontraba en el tercio superior de GPA de pasantía; mientras que, solo el 8% (5/61) de los estudiantes que hizo tres o más apariciones en el cuartil inferior estaban en el tercio más alto de GPA de pasantía, una proporción de 9:1. Por el contrario, solo el 9% (15/164) de los estudiantes que nunca aparecieron en el cuartil inferior estaban en el grupo más bajo en función de sus GPA de pasantía, mientras que 71% (43/61) de los que aparecieron tres o más veces estaban en el grupo de GPA más bajo, una proporción de 1:9. Los autores del estudio concluyeron a partir de esos resultados, que el bajo rendimiento académico en el primer año de la facultad de medicina es un factor de riesgo significativo con validez predictiva y utilidad predictiva para el bajo rendimiento posterior en la facultad de medicina y en momentos posteriores (examen de licencia médica GPA).

Otros estudios, como el de Dabaliz et al. (2017), intentaron relacionar datos de admisión de 737 estudiantes (puntajes de la escuela secundaria y puntajes de pruebas estandarizadas, como las de la prueba nacional de logro, la prueba de aptitud general, el TOEFL y el IELTS) con indicadores de desempeño universitario (GPA y la prueba de progreso), utilizando un análisis de regresión lineal multivariado. Ellos encontraron que ninguna de las variables de preingreso fue predictiva, tanto para el promedio de calificaciones acumuladas de GPA en años clínicos como para el rendimiento en las pruebas de progreso de los estudiantes (conocimiento funcional o aplicado). En general, solo el IELTS ($p=0,04$, $B=0,08$) y el TOEFL ($p=0,017$, $B=0,01$) predijeron significativamente el rendimiento en años preclínicos (básicos), y el GPA predijo fuertemente el rendimiento de la prueba de progreso de los estudiantes ($p<0.001$ y $B=19.02$).

En otro estudio realizado, Adam et al. (2015) mostraron una correlación predictiva significativa entre el puntaje académico de admisión (rendimiento académico escolar previo) realizado por la escuela de medicina de Hull York en Reino Unido, los exámenes escritos del cuarto año (r_1= 0.174; r_2=0.200; $p<0.05$) y el examen práctico de quinto año de la carrera de esos mismos estudiantes (r_1= 0.213; $p<0.05$). Además, algunos puntajes de subpruebas de los Test de Aptitudes Clínicas de Reino Unido (UKCAT) y el puntaje total de UKCAT (r_1=0.244 y r_2=0.250; $p<0.01$), también se correlacionaron positivamente con los exámenes clínicos de los últimos años de facultad ($r=0.256$; $p<0.01$).

4.2. Evaluación de Resultados de aprendizaje a nivel estatal en estudiantes de pregrado de medicina

El Estado debe preocuparse por la calidad de la educación superior y de los profesionales resultantes de la misma. De allí que, dentro de las políticas de certificación ocupacional del egresado, especialmente en áreas de la salud, estén establecidos diferentes mecanismos de acreditación de habilidades médicas. Desde cualquiera de los enfoques de los sistemas de salud alrededor del mundo se requiere certificar, y mantener dicha certificación, con la actualización de los conocimientos y habilidades resolutivas del médico.

Para el Consejo colombiano de acreditación y recertificación médica de especialistas y profesiones afines (CAMEC), el modelo más desarrollado de recertificación que existe es el de Estados Unidos conocido por todos como el sistema voluntario de los “Board”, el cual incluye las diferentes ramas de la medicina y especialidades. El “American Board of Medical Specialties” (AMBS) es copiado en todo el mundo por ser quizá el de más trayectoria y organización con un resultado exitoso. Otro ejemplo es el The European Accreditation Council for CME (EACCME), que ejerce las funciones de acreditación de actividades académicas en el área médica para la comunidad europea desde 1999. También en Canadá y Australia operan sistemas similares al americano, con gran posicionamiento y respeto.

En Latinoamérica, México cuenta con un sistema muy desarrollado de certificación y recertificación denominado Comité Normativo Nacional de Consejos de Especialidades Médicas (CONACEM) desde 1995. En Chile se aplica el Examen único nacional de conocimientos de medicina (EUNACOM); en Brasil y Argentina existen sistemas similares, pero menos estructurados que el de México.

En Colombia, con la presentación oficial a la comunidad colombiana del Consejo colombiano de acreditación y recertificación médica, de especialistas y profesiones afines (CAMEC), se buscó garantizar a futuro una rectoría de los procesos voluntarios de mejoramiento de la calidad a partir de la formación continua en el área de la salud y se convirtiera a través del tiempo en el verdadero "Board colombiano".

4.2.1 Evaluación de Resultados de aprendizaje a nivel estatal de estudiantes de pregrado de medicina en el contexto internacional

El rendimiento académico de un estudiante universitario es el resultado de una multiplicidad de factores, que van desde los personales, los relacionados con el entorno familiar y social en el que se desenvuelve, los dependientes de la institución y los factores que dependen de los docentes (Ruiz, 2010). Podemos definir el rendimiento académico como un constructo que puede adoptar valores cuantitativos y cualitativos, a través de los cuales existe una aproximación a la evidencia y a la dimensión del perfil de habilidades, conocimientos, actitudes y valores desarrollados por el alumno en el proceso de enseñanza aprendizaje (Edel, 2003). Sin embargo, las múltiples interpretaciones contextuales del concepto, visto como variable de estudio en la investigación educativa, dan pie a llamativas interpretaciones cuantitativas que se acercan a evidenciar las habilidades adquiridas, a través de la evaluación de los resultados de aprendizaje esperados. Las pruebas que evalúan en un estudiante las habilidades cognitivas aplicadas en bloques conceptuales, de forma coherente y en secuencia, evidencian el "valor agregado" impregnado hasta en un 30 % de los programas

académicos (Keltner et al., 2021). El rendimiento académico ha sido objeto de estudio reiteradamente en la investigación educativa y las calificaciones representan el indicador más empleado (MacKenzie et al., 2016). El rendimiento académico, medido a través de la calificación promedio o del avance en la carrera, está afectado mayoritariamente por el rendimiento previo, que resulta ser la variable predictora dominante (MacKenzie et al., 2016).

Para la Agencia nacional de evaluación de la calidad y acreditación de España (ANECA), organismo autónomo que vela por la promoción y el aseguramiento de la calidad del Sistema de Educación Superior español, la orientación evidenciable de las capacidades que debe tener un egresado de un programa de educación superior se puede obtener a través de los Resultados de aprendizaje esperados (RAE) para el profesional resultante del proceso de enseñanza aprendizaje y el respectivo componente curricular programático. Los RAE muestran el comportamiento deseado del individuo en el paisaje en el que se desenvolverá, por lo tanto, a partir de ellos se deben diseñar las guías metodológicas y su momento de aplicación que hagan factible la consecución de un egresado capaz, dueño de su evolución posterior (ANECA, 2103). En las facultades del área de la salud no se escatiman esfuerzos para que dichos RAE de egreso sean faros en el horizonte curricular. De allí que tanto las revisiones al plan de estudios como la supervisión de las evaluaciones y su retroalimentación, han tomado fuerza de la mano de las pruebas estandarizadas para egresados y las homologaciones de habilidades profesionales. Es así como las trayectorias hacia la excelencia en la evaluación integral de las tres facultades de medicina ganadoras del premio ASPIRE (The AMEE School Programme for International Recognition of Excellence in medical education), cuyo objetivo es reconocer la excelencia de programas educativos en las facultades de medicina de la AMEE (Asociación europea para la educación médica) poseen componentes básicos similares y ponen de manifiesto los sólidos fundamentos de la

evaluación de los estudiantes, a pesar de haberse emprendido en contextos culturales diferentes (Beason et al., 2019; Ferris & O'Flynn, 2015).

Las tendencias a correlacionar el valor predictivo entre las pruebas o mecanismos utilizados para la selección en la admisión de estudiantes al pregrado de medicina y el promedio final de la carrera, se convierten en una de las formas de supervisar lo que los estudiantes traen desde su educación secundaria y su idoneidad para hacer una mejor selección cognitiva aptitudinal. Es así como en una universidad pública argentina, Tomatis et al. (2016) investigaron 225 estudiantes de la carrera de medicina que ingresaron en el mes de marzo del año 2006 y egresaron en el mes de diciembre del año 2012, cumpliendo en tiempo y forma la carrera. Registrando las notas de los estudiantes en el examen de ingreso y contrastándolas con sus promedios generales de la carrera, sus notas finales de la práctica final obligatoria (PFO) y mediante el análisis de regresión lineal, demostraron que hay asociación positiva entre la nota de ingreso y el promedio de la carrera ($p<0,0001$). De esta manera, la nota de ingreso resultó ser predictor del promedio general de la carrera; donde, por cada punto que aumenta la nota del ingreso, aumenta 0,38 el promedio de la carrera.

En otro estudio liderado por Wilkinson et al. (2011), en una muestra de 339 estudiantes que ingresaron a la carrera de medicina en la Facultad de Medicina de la Universidad de Queensland, directamente desde la escuela secundaria, entre los años 2005 y 2009, el test de admisión a la licenciatura de medicina en Australia (UMAT) tuvo una validez predictiva limitada para el rendimiento académico. La puntuación media general de UMAT al ingreso fue 60/100 y el GPA medio durante los estudios universitarios fue 6,1 (rango, 1-7), con un coeficiente de correlación de 0,15 ($P = 0,005$). Esta relación existió sólo en el primer año de estudios universitarios. La UMAT tenía tres secciones: Sección 1: Razonamiento lógico y resolución de problemas (48 preguntas); Sección 2: Comprensión de las personas (44 preguntas); Sección 3: Razonamiento no verbal (42 preguntas). Para el puntaje de la Sección

1 de UMAT, el coeficiente de correlación fue de 0.14 (P = 0.01); para la UMAT Sección 2 el coeficiente de correlación fue de 0,06 (P = 0,29); y para la UMAT Sección 3 el coeficiente de correlación fue de 0.09 (P = 0.11). En el análisis multivariado, solo la correlación entre el GPA y el puntaje de la Sección 1 de UMAT permaneció significativa, pero fue débil y duró 1 año de estudios universitarios. Cabe resaltar que la prueba UMAT fue reemplazada en el 2019 por la UCAT (homologada con la prueba análoga de Reino Unido).

4.2.2 Evaluación de Resultados de aprendizaje de estudiantes de pregrado de medicina en Colombia

El Instituto colombiano para el Fomento de la educación superior (ICFES) ejerce inspección y vigilancia de la calidad del servicio público educativo, a través de una serie de instrumentos, acciones y procesos. Por ejemplo, las pruebas que el ICFES realiza para evaluar la educación a nivel país se pueden dividir, según Castro y Ruiz (2019), en tres grandes grupos: las de educación media, denominadas Saber 11; las de educación superior, conocidas como Saber Pro (antes ECAES) y el Programa para la evaluación internacional de estudiantes (*Programme for International Student Assessment – PISA*) de la Organización para la Cooperación y el Desarrollo Económicos (OCDE). Estas últimas comparan el desempeño colombiano frente al de otros países y ayudan a mejorar la calidad y eficiencia de las pruebas estatales (Ferrer & Arregui, 2003).

Aplicar pruebas estandarizadas como Saber 5, 9, 11 o Saber Pro (sistema integrado de evaluación colombiano) es beneficioso por su alta validez externa; es decir, poder medir el desempeño educativo en distintos lugares, para luego compararlos y establecer conductores y tendencias, entre otros aspectos. Los resultados arrojados por estas pruebas son de utilidad para que muchos actores puedan tomar decisiones informadas, entre los que cabe mencionar el Estado, las instituciones educativas y los propios estudiantes. Dichas pruebas pueden llegar

a afectar la asignación de la cantidad de recursos con base en deficiencias o logros obtenidos (Castro & Ruiz, 2019).

La prueba Saber 11 define, a partir de sus cinco componentes, aspectos que evidencian los resultados esperados en un estudiante de educación media en Colombia. Los cinco componentes aparecen alineados con las restantes pruebas del sistema integrado de evaluación que lidera el ICFES. El componente de lectura crítica aparece evaluado en las pruebas Saber 3, 5, 9, 11 y Pro; igualmente, matemáticas es evaluada como componente de todos los niveles de las pruebas Saber 3, 5, 9, 11 y Pro; ciencias sociales y competencias ciudadanas, al igual que ciencias naturales, como componentes son supervisados en los cuatro niveles superiores de las pruebas Saber (5, 9, 11 y Pro). Ciencias naturales aparece como pensamiento científico en la prueba Saber Pro, haciendo parte del componente específico de dicha prueba. La evaluación de comunicación escrita, hasta ahora, es únicamente evaluada en la prueba Saber Pro (ICFES, 2018). Cabe destacar que, según Castro et al. (2018), Colombia es el único país del mundo donde se puede hacer un seguimiento del mismo estudiante con las pruebas de Estado estandarizadas al terminar la secundaria y la universidad.

Hipotéticamente, entonces el rendimiento en la prueba Saber 11 en Colombia podría ser tomado como evidencia de rendimiento previo para ser comparado con las pruebas e indicadores de progreso y final del proceso académico del programa de medicina. A su vez, cada una de las demás pruebas analizadas por la presente propuesta de investigación (examen de fin de ciclo básico, examen de fin de ciclo clínico y GPA) se convierten en evidencia de rendimiento previo de las evaluaciones posteriores en la secuencia académica (Keltner et al., 2021). En nuestro caso, la prueba de egreso final de la secuencia ascendente de evaluaciones oficiales a la que está expuesto un médico en formación es la prueba Saber Pro.

La prueba Saber Pro hace parte del proceso de evaluación de la calidad académica de la educación superior en Colombia; esta, contiene un conjunto de pasos estandarizados y está

fundamentada en el DCE (Diseño centrado en evidencias) (ICFES, 2018). Este examen comenzó a aplicarse desde el año 2003 para la evaluación de 22 programas de educación superior, recibiendo inicialmente la denominación de Examen de la calidad de educación superior (ECAES). En el 2007, la base de programas evaluados aumentó a 55, centrándose en la evaluación de competencias específicas por programa. Entre 2009 y 2010 se incluyeron dos componentes evaluativos comunes a todos los programas de formación: comprensión lectora y comprensión del idioma inglés (ICFES, 2018).

La obligatoriedad en la realización de la prueba Saber Pro como requisito para la obtención del título del nivel de pregrado se dio con la publicación de la Ley 1324, y su decreto reglamentario 3963 de 2009. Así se dio una nueva orientación a los exámenes estatales de la educación superior (ICFES, 2018). En general, son tres los objetivos declarados por el ICFES para el examen Saber Pro:

1. Comprobar el desarrollo de competencias de los estudiantes que han aprobado el 75% de los créditos en un programa de formación.
2. Producir indicadores de valor agregado.
3. Servir de fuente de información para la generación de indicadores de evaluación de la calidad de la educación superior. (ICFES, 2018).

Para el tercer objetivo mencionado, los resultados obtenidos en las pruebas Saber Pro son utilizados en los indicadores evaluados para la acreditación de programas. Estos resultados tienen relevancia en la característica de integralidad del currículo, correspondiente a uno de los factores de acreditación denominado procesos académicos. El desempeño de los estudiantes de un programa en las pruebas de Estado de educación superior de los últimos cinco años y sus calificaciones promedio, comparadas con el promedio nacional, se tiene en cuenta para evaluar la característica mencionada (Consejo Nacional de Acreditación-CNA-, 2013).

Desde 2010 y en coherencia con la Ley 1324, la nueva prueba está compuesta por una evaluación de competencias genéricas, entendidas como "aquellas que todos los estudiantes deben desarrollar independiente del énfasis de formación" (ICFES, 2018), y la evaluación de competencias comunes a grupos de programas con características de formación similares.

Para la evaluación de competencias genéricas se realizan cinco pruebas: Lectura Crítica, Razonamiento Cuantitativo, Comunicación Escrita, Inglés y Competencias Ciudadanas (ICFES, 2018). Por otro lado, desde el 2021 las pruebas de competencias específicas comunes para distintos grupos de referencia son 49. En el grupo de profesiones del área de la salud se incluye a la medicina y se evalúa con tres subcomponentes específicos: fundamentación en diagnóstico y tratamiento, atención primaria a la salud y promoción y prevención de la enfermedad (ICFES, 2022).

Actualmente, la información obtenida a través del examen se reporta a tres niveles: individual, por programa y por instituto de educación superior. En esos tres niveles los datos incluyen la información del puntaje global y el percentil general en el que se encuentra el evaluado, así como también el percentil y el nivel de desempeño alcanzado en cada una de las competencias evaluadas y la descripción de las habilidades asociadas a cada nivel. Esto, permite su correspondiente análisis teniendo en cuenta determinadas características (carácter, sector o acreditación) de los programas y universidades (ICFES, 2022). Los lineamientos para el diseño del examen Saber Pro se definieron de acuerdo con la política de formación por competencias del Ministerio de Educación Nacional, tanto en el nivel universitario como en el nivel tecnológico y técnico profesional (Pruebas TyT), y en su desarrollo han participado las comunidades académicas, asociaciones y redes de facultades y programas (Delgado, 2011).

La Asociación colombiana de facultades de medicina (ASCOFAME) ha liderado la educación médica en Colombia en momentos coyunturales. Factores como la globalización,

la tecnología y la veloz evolución del conocimiento en la última década han planteado retos a la educación médica en Colombia. Esos retos se resumen en la formación coherente de un médico competentemente resolutivo y la supervisión de los procesos de enseñanza-aprendizaje en los programas de medicina del país. ASCOFAME tiene asociadas a más del 90 % de las facultades de medicina y en 2017 se reunieron en el llamado Consenso de Montería sobre educación médica para emitir directrices que garantizaran la calidad en el afrontamiento de los mencionados retos (ASCOFAME, 2017).

El Ministerio de Educación nacional ha sido otro pilar fundamental en la supervisión de la calidad de los programas profesionales. Los procesos de registro calificado y acreditación de alta calidad para programas han sido una hoja de ruta para el logro de la excelencia académica en educación superior. La educación en salud no ha sido ajena a esta influencia (Delgado, 2011). Los programas de medicina apuntan sus esfuerzos hacia la acreditación de Alta calidad, a sabiendas que el prestigio y la garantía de procesos académicos auditados por pares externos son invaluables para las oficinas de mercadeo de las instituciones educativas y para la homologabilidad programática interinstitucional, nacional e internacional (Gil et al., 2013).

Las universidades con programas de medicina acreditados y en los diez primeros lugares de los diferentes rankings nacionales han hecho esfuerzos importantes para evidenciar en sus egresados su capacidad resolutiva, promocionada desde ASCOFAME. Generalmente, dichos esfuerzos se han basado en modelos de modernización curricular por competencias. Una muestra de esto es la facultad de medicina de la Universidad del Rosario que utiliza de forma explícita la resolución de problemas en sus asignaturas del componente básico, a través del “aprendizaje basado en problemas y enfocado en la comprensión de conceptos, creando un núcleo curricular denominado Actividades integradoras del aprendizaje por sistemas (AIAS).” (Vergel, 2019). Otro ejemplo de las modificaciones aplicadas para este fin se puede

evidenciar en los planes de estudio de la Universidad de los Andes y la Universidad Javeriana; en sus programas de medicina ambas optaron por el uso de los sistemas corporales para integrar de manera coherente y secuencial las áreas disciplinares básicas (Morfología, Fisiología, Microbiología, Patología, Farmacología) (Pontificia Universidad Javeriana, s.f.; Universidad de los Andes, s.f.).

El programa de medicina de la Universidad del Norte, desde el año 2003, le apostó a una modernización curricular basada en la integración de áreas disciplinares a través de sistemas corporales. Su buen funcionamiento fue evidenciado a partir de las acreditaciones de alta calidad otorgadas al programa (2012 y 2021). Dentro de los cambios posteriores, adoptados como resultado de los procesos de autoevaluación con miras a la acreditación, estuvieron la creación de las pruebas de suficiencia de sus ciclos básico-profesional y profesional. Las sendas evaluaciones se aplican en momentos del fin de los correspondientes ciclos de conceptos básicos y clínicos, respectivamente (Martínez et al., 2015).

Las investigaciones sobre los resultados de las pruebas Saber Pro son pocas y principalmente se basan en la descripción de los resultados por componentes y sus comparaciones con los resultados obtenidos a nivel nacional. Sin embargo, Gil et al. (2013) estudiaron a 4.498 estudiantes de medicina en 40 universidades del país que presentaron el examen Saber pro en el año 2009. Ellos encontraron que la edad promedio de los estudiantes era de 24,8 años (desviación estándar [DE], 2,6 años); el 55,0% fueron mujeres y en general correspondían a estratos socioeconómicos 3 o 4; en su gran mayoría su estado civil era solteros (95,9%), no tenían personas a cargo (94,4%), y el hogar de residencia en el 63,8% de los casos era el habitual (permanente). La actividad laboral de sus madres mayoritariamente era el hogar (30,1%), mientras que sus padres eran principalmente empresarios (30,8%) o trabajadores independientes (25,2%). En cuanto a los ingresos económicos a nivel familiar, los hogares en mayor frecuencia se encontraban entre 3 y menos de 5 salarios mínimos

legales vigentes (SMLV); el 15,8% de los hogares tenían menos de 2 SMLV como ingresos y solamente el 10,5% reportó ingresos superiores a 10 SMLV.

En lo relacionado con los promedios obtenidos por los estudiantes, Gil et al. (2013) observaron que los mayores puntajes recayeron en los hombres comparados con las mujeres, los estratos socioeconómicos 5 y 6 con estudiantes de padres con formación de posgrado, y estudiantes en hogares de ingresos mayores a 10 SMLV. Estos hallazgos a nivel de estudiantes son consistentes con otros estudios realizados sobre educación en Colombia a nivel de primaria y bachillerato; así mismo, son concordantes con otros estudios realizados sobre las pruebas Saber Pro en otros programas de formación universitaria (Sánchez -Bello et al., 2016).

4.3. Validez Predictiva

La validez se refiere al grado en que la evidencia y la teoría soportan la interpretación de los puntajes derivados de las pruebas (test o instrumento) para los usos propuestos de dicha prueba (Asociación americana de psicología, APA por sus siglas en inglés). Otra definición tradicional de validez establece que ésta se refiere al grado en que un instrumento mide con precisión el constructo que busca medir. La validez es una cualidad que se refiere a la inferencia que se hace de los datos recolectados con un instrumento. De acuerdo con Sireci y Sukin (2013), el concepto de validez es comprensivo y se refiere no solo a las características de la prueba, sino a la oportunidad del uso de la prueba y a la precisión de las inferencias realizadas a partir de los puntajes (p. 61). Shadish et al., (2002) se refieren a la validez como la verdad aproximada de una inferencia (p. 34), es decir, la medida en que los resultados son fiables o dignos de confianza.

La validez es especialmente importante en las ciencias sociales, en la psicología y la educación, donde lo que se busca medir no es tan concreto como lo es en las ciencias físicas, y por lo tanto las definiciones son complejas. Como consecuencia, lo que se busca medir

generalmente se define operacionalmente; es decir, de forma tal que pueda ser "medido" con el instrumento que se dispone o desarrolla. La teoría referente a la validez ha pasado de percibirla como diferentes tipos de validez (por ejemplo, validez del constructo, validez basada en criterio, validez facial) a definirla como único concepto que se verifica a través de distintos tipos de evidencias definidas en la literatura (Kane, 1992; 2006). Se tiene entonces que los distintos "tipos de validez" que se han venido desarrollando constituyen actualmente diferentes tipos de "evidencias" que confirman (o no) la existencia de validez de las inferencias obtenidas a través de los datos. Es por esto que generalmente se recomienda proveer más de una evidencia de validez para confiar en las inferencias derivadas de los datos revelados a partir de un instrumento.

Los estándares para pruebas educativas y psicológicas de la APA (2014) identifican cinco fuentes de evidencias de validez: (i) contenido de la prueba, básicamente refiere a la relación que existe entre el contenido de la prueba y el constructo que busca medir dicha prueba, con esto se determina el alcance del significado de los puntajes obtenidos en la prueba y en consecuencia las conclusiones o inferencias a las que se puede llegar; (ii) relaciones con otras variables, dado que es altamente probable que existan otras pruebas que persiguen el mismo objetivo (o el objetivo exactamente opuesto) que el de la prueba en cuestión, la relación con variables externas a la prueba también provee evidencias de validez; (iii) estructura interna, que indica el grado en el que la relación entre los ítems y los componentes de la prueba se conforman con el constructo que la prueba buscó medir; (iv) procesos de respuesta, que considera no el resultado obtenido en una prueba, sino el proceso cognitivo subyacente a la resolución de los problemas (o selección de respuestas) presentados en la prueba; y finalmente (v) consecuencias de las pruebas, refiere a la interpretación de los puntajes en relación al propósito que se buscaba (por ejemplo, determinar la competencia de las personas en una dimensión específica) así como sus consecuencias indirectas (por

ejemplo, ranquear universidades en función a puntajes que no buscaban eso) (Sireci & Sukin, 2013).

Muchas de las técnicas empleadas para recopilar evidencias incluyen análisis estadísticos. Por lo tanto, las pruebas de significancia y las estimaciones del tamaño del efecto también forman parte de todo el proceso. La segunda fuente de evidencia arriba mencionada incluye la evidencia convergente y divergente. En el primer caso, se refiere a la relación entre los puntajes de la prueba bajo estudio y otras que de alguna forma miden lo mismo o miden constructos teóricamente relacionados entre sí (amor y amistad). La evidencia de validez divergente busca lo opuesto, determinar el grado en el que los puntajes de la prueba bajo estudio se correlacionan inversamente con pruebas que miden constructos completamente opuestos (alegría y tristeza).

La segunda fuente de evidencias también incluye la evidencia de validez de la prueba en relación con un criterio. En otras palabras, refiere a la capacidad de la prueba de predecir un criterio (otra variable). El criterio es una medida de algún atributo que difiere operacionalmente al de la prueba bajo estudio; es decir, la prueba no es una medición o intento de medición del criterio. Por lo tanto, lo que se busca aquí es determinar el potencial predictivo de la prueba con relación al criterio.

La prueba puede predecir un criterio que se ubica en el mismo momento temporal o en otro (APA, 2014). Dos tipos de evidencia en relación con el criterio han sido históricamente estudiados: evidencia concurrente y predictiva, que en la literatura también se conocen como “validez” concurrente y predictiva, atendiendo la perspectiva inicial que se tenía de la validez (es decir, eran varias y no una sola como ahora se asume). La evidencia concurrente refiere a la predicción que hace la prueba de un criterio que se genera en el mismo tiempo que dicha prueba (APA, 2014).Un ejemplo de lo anteriormente señalado sería la predicción que tiene el rendimiento en la prueba de lectura en el de la prueba de

matemática, ambos administrados al mismo tiempo a los estudiantes de sexto grado (esta es la situación de muchas pruebas estandarizadas, nacionales e internacionales, como por ejemplo las Saber o las del Laboratorio de la UNESCO).

Por su parte, la evidencia predictiva se refiere a la predicción que hace la prueba de un criterio que se genera en un tiempo posterior al de la prueba bajo estudio, por ejemplo, la capacidad que tiene el rendimiento en las pruebas Saber 11 de predecir el rendimiento de los estudiantes en la universidad. La evidencia de validez en relación con el criterio es importante dado que muchas veces se utilizan puntajes en una prueba para tomar decisiones sobre los individuos en relación con su desempeño en otra área, pues se asume que el rendimiento en una prueba predice su desempeño en otros espacios o contextos; un ejemplo de esto sería en la asignación de puestos en una empresa, en la adjudicación de cargos en un concurso laboral, en la admisión de un estudiante a una universidad. En consecuencia, es importante comprobar que los puntajes obtenidos en una prueba efectivamente soportan las decisiones efectuadas a partir de ella. Esto se logra determinando el poder predictivo de la prueba sobre la dimensión que se afecta a partir de ella y demostrando que el rendimiento en la prueba es un recurso útil y efectivo para la toma de decisiones en las dimensiones de interés (APA, 2014; Thorndike et al., 1991).

Puede haber más de una variable criterio en un estudio de validez predictiva. En educación, por ejemplo, el rendimiento en la prueba de admisión podría estudiarse como predictor de criterios como el rendimiento académico a lo largo de la carrera de pregrado (calificaciones obtenidas) o el rendimiento en alguna prueba comprensiva que forma parte del programa de estudio. Cabe señalar que todas las variables que sirven de criterio sólo miden parcialmente el éxito académico de los estudiantes, puesto que existen otros factores determinantes de este fenómeno. De igual forma, el éxito académico de los estudiantes no predice el éxito profesional (variable criterio en este caso) de estos estudiantes, por la misma

razón (Thorndike et al., 1991). Por lo tanto, la utilidad de una prueba como predictora depende de lo bien o no que se relacione con la variable criterio y qué tan informativa sea la prueba. Es decir, si la prueba bajo estudio contribuye a la predicción de manera innovadora en comparación con otras pruebas predictivas que quizá ya existan.

La medida de la relación o de las evidencias entre variables predictivas y de criterio se derivan de procesos de análisis estadísticos de datos como la correlación y los modelos de regresión (de un solo nivel o multinivel), cuyos coeficientes incorporan la correlación (matemáticamente hablando). Se asume que, a mayor correlación, mayor capacidad predictiva de la prueba; sin embargo, no existe un parámetro predeterminado que indique que una variable es "buena" como predictora. Cabe señalar que la predicción del criterio no está libre de errores, que a su vez tienen varias fuentes (medición, estimación). El error de medición (que a su vez está relacionado al coeficiente de fiabilidad de una prueba) se refiere al error en la medición del constructo con un instrumento (como cuando la balanza reporta varios pesos para la misma persona en un rango de tiempo bastante corto en el que no se espera que la persona haya cambiado de peso), mientras que el error de estimación se refiere al error que se produce al estimar el valor del criterio a partir de los valores de la prueba predictora. En la medida que estos errores sean altos, la calidad de la evidencia de validez es baja. Es más, los coeficientes de relación estimados (por correlación o regresión) podrían estar siendo subestimados (Thorndike et al., 1991).

Resumiendo, la evidencia predictiva permite determinar el potencial de predicción de una variable de interés a partir de otra temporalmente disponible y tomar decisiones basadas en esta capacidad predictiva. Por su parte, una evidencia predictiva pobre no necesariamente indica problemas en la prueba bajo estudio, pues este bajo poder predictivo podría también estar siendo afectado por otros factores no contemplados en el análisis, por la calidad de la variable criterio, y/o por el nivel de solapamiento que existe entre la variable predictora y el

criterio desde la teoría (por ejemplo, es esperable que no exista mucha correlación entre el rendimiento en una prueba de lectura y el rendimiento en educación física).

Revisando las pocas investigaciones específicas que hacen referencia a la validez predictiva en la secuencia entre las pruebas Saber 11 y Saber Pro en egresados de programas de salud y humanidades, esta fue variable. Algunas mostraban resultados esperables de acuerdo a su contexto, sin la confirmación de un plus en lo adquirido por el estudiante durante el paso por las diferentes etapas de su programa de estudio. Una muestra de lo anteriormente señalado es el estudio presentado por Bahamón y Reyes (2014), ellas encontraron que de 68 estudiantes de Psicología participantes de su investigación, el 57% obtuvo puntuaciones bajas en el examen Saber 11, en tanto solo el 26.5% mostró puntuaciones altas; contrastando con sus resultados en dos dimensiones de las pruebas Saber Pro (escritura y lectura crítica), donde ese mismo grupo de estudiantes obtuvo puntajes altos. Ellas asumen que las modificaciones en el proceso del estudiante durante su programa académico les permitieron mejorar su desempeño, como resultado de las características propias de la carrera. La comparación con los resultados de las pruebas Saber 11 se realizó de manera individual a los puntajes de las competencias evaluadas, por medio de una Prueba T Student, hallando diferencias significativas en los resultados de las áreas mencionadas.

También, Domingue (2012), en un estudio sobre la efectividad de las universidades, evalúa la relación entre los resultados del examen Saber Pro y Saber 11, encontrando que el examen Saber 11 es un importante predictor del Saber Pro. En otro estudio, Melo et al. (2014), en un análisis de eficiencia de ciertos factores de la educación colombiana, hallaron una correlación positiva entre los resultados de las pruebas Saber 11 con el resultado promedio del grupo de estudiantes que presentaron la prueba Saber Pro en el segundo semestre de 2011, de 0,88. Inclusive, por grupos de referencia, dicha correlación supera el 0,9 para las carreras de Medicina, Derecho y Ciencias económicas y administrativas.

En otra investigación con mayor especificidad predictiva, Castro et al. (2018), a través de una regresión exploratoria, caracterizaron el desempeño de los resultados de 1806 estudiantes del departamento de Antioquia, Colombia. Este grupo de estudiantes tomaron la prueba Saber 11 para los años 2005 y 2006, y luego la prueba Saber Pro 2009-2010; como era de esperarse, se dio una relación positiva entre los puntajes de la prueba Saber 11 y Saber Pro, esto es: a mayor puntaje en la prueba Saber 11, mayor puntaje en Saber Pro (ß=9.698; $p<0.05$); es decir, los estudiantes que fueron buenos en el colegio continuaron siéndolo en la IES. En otros casos no se registró significancia estadística en las comparaciones realizadas, por ejemplo: los resultados obtenidos para una tesis de pregrado de enfermería de una universidad del oriente colombiano, donde Acaut et al. (2021) contrastaron en 14 estudiantes de enfermería, los resultados totales y por área, tanto en las pruebas Saber 11 como en su correspondiente Saber Pro, no mostraron diferencias significativas. Igualmente, para Gabalan-Coell y Vásquez-Rizo (2016), en una cohorte específica de una IES colombiana (con énfasis en ciencias exactas y de administración), no existió asociación significativa entre los puntajes obtenidos por los estudiantes en las pruebas de Estado (componentes de matemática y lenguaje) y los posteriores desempeños académicos en la universidad en asignaturas relacionadas directamente con estos dos componentes. Ellos observaron que, al correr el coeficiente de correlación de Pearson para las dos poblaciones con el rendimiento en asignaturas del primer año, se puede notar que la asociación muestra diferencias entre los resultados de Saber 11 de los estudiantes nivel A (más de 60 en puntaje general) con los rendimientos universitarios en asignaturas relacionadas con matemáticas, siendo de −0.03; mientras que, en los del nivel B (menos de 60 en puntaje general) dicho coeficiente es de 0.01. Para el caso de lenguaje, la asociación de los de nivel A es de 0.11, mientras que la asociación de los de nivel B es 0.13. La evidenciable ausencia de una fuerte proporcionalidad, que confirma lo no significativo de estas diferencias estadísticas, evidencia,

según dichos autores, el nulo poder predictivo de las pruebas Saber 11 sobre el rendimiento progresivo de este grupo de estudiantes universitarios.

Otro es el panorama que se visualiza en relación a la validez predictiva de las pruebas de progreso en los programas de medicina, homologables a las registradas en la presente investigación (pruebas de fin de ciclo básico y clínico), por lo menos a nivel internacional. Especialmente, si relacionamos su aparición en la educación médica como la forma más fácil de evidenciar la transformación del estudiante de medicina, que era producto de los planes de estudios basados en el aprendizaje a partir de problemas. Los programas de medicina en el mundo que se apropiaron de estas estrategias curriculares necesitaron asegurarse de la consecución de logros de forma espiral en el plan de estudios propio. Así que las evaluaciones periódicas de progreso fueron allí importantes (Plessas, 2015).

La prueba de progreso es una prueba integral que muestra el conocimiento en todas las áreas de contenido de la medicina, que refleja los objetivos finales del plan de estudios o una sección de él. La prueba se aplica periódicamente a todos los estudiantes de medicina del plan de estudios independientemente de su año de formación (Van Der Vleuten, 1996). Es plausible que este enfoque de evaluación integrada, por ser longitudinal en el tiempo, tenga un efecto positivo en el comportamiento de aprendizaje de los estudiantes al desalentar el aprendizaje compulsivo-memorístico (Van Der Vleuten et al., 2012), por lo que su valoración a través de la psicometría ha tomado mucha importancia para el uso y la estandarización de las mencionadas pruebas.

No obstante, la prueba de progreso puede ser un método de evaluación exitoso independientemente de si las escuelas emplean el aprendizaje basado en problemas o no. Verhoeven et al. (2005) no encontraron diferencias sistemáticas en los puntajes totales de las pruebas entre estudiantes de dos escuelas de medicina holandesas, una con aprendizaje basado en problemas y otra que no lo aplicaba en su plan de estudios. Una guía reciente

publicada por la Asociación para la educación médica en Europa (AMEE) describe un marco sistémico para las pruebas de progreso, fomentando el establecimiento potencial o nuevo de estas pruebas en los programas de educación médica, haciendo factible otra manera de valoración internacional de los procesos académicos en la educación médica (Wrigley et al., 2012).

En la misma línea, en el estudio de Plessas (2015), "Validity of Progress Testing in Healthcare Education", se concluye que: 1. La validez del contenido de las pruebas de progreso está asegurada por un plan cuidadosamente diseñado, elementos de alta calidad escritos y revisados por expertos para el control de calidad; también, la retroalimentación de y para los estudiantes apoya el proceso que otorga validez a la prueba. Sin embargo, las características psicométricas integrales de la prueba proporcionan incluso más evidencia relacionada con la estructura interna; 2. La validez de constructo también está respaldada por el aumento de las puntuaciones medias de la prueba, según el año y la relación con otras pruebas pertinentes y exámenes de grado.

A pesar de la relevancia mencionada, existen algunas amenazas a la validez cuando se intentan colaboraciones y comparaciones interinstitucionales. En vista de esto, la AMEE intenta, a partir de un marco de estandarización, proporcionar un análisis de los requisitos básicos para minimizar esas amenazas, se enfatiza: la necesidad de calidad en los procedimientos de control, formación y evaluación continua; esfuerzo y recursos de inversión, coherentes entre sí; y el compromiso para mantener una alta confiabilidad y validez de la prueba (Wrigley et al., 2012).

La validez de una prueba de progreso se podría valorar desde el tipo de pregunta, sus ítems y la forma de aplicación de dicha evaluación; por ejemplo, una prueba holandesa de progreso colaborativo originalmente constaba de 250 preguntas con opción de verdadero-falso, pero a partir de 2005 se cambió a 200 preguntas de opción múltiple con la mejor opción

única (Schuwirth et al., 2010). Un ejemplo diferente es el observado por Rademakers et al. (2005) en la Universidad de Utrecht, donde empleando un menor número de estudiantes y preguntas (de respuesta corta basadas en 40 casos clínicos), se demostró confiabilidad y factibilidad de la prueba (alfa de Cronbach 0.85-0.87). Sin embargo, con las preguntas con opción de única (mejor) respuesta se proporcionan puntajes más confiables y una menor probabilidad de adivinar, mejorando la validez de la evaluación (Wrigley et al., 2012). En su estudio, Plessas (2015) observó que la mayoría de las pruebas de progreso a nivel internacional emplean preguntas de opción múltiple con única respuesta.

La validación del contenido de las pruebas a través de la elaboración y/o revisión por expertos es una forma de conseguir mejores resultados. Una muestra de lo ya mencionado es que en la cooperación interinstitucional holandesa participan en el proceso de control de calidad y redacción de ítems ocho miembros del comité con experiencia en ciencias básicas, clínicas y del comportamiento (Van der Vleuten et al., 2004). De manera similar, en la Península Medical School todos los elementos de la prueba deben ir acompañados de una referencia bibliográfica, lo que respalda aún más la validez relacionada con el contenido de la evaluación (Freeman et al., 2010; Van der Vleuten et al., 2004). Es claro que la calidad de los ítems se mejora con el entrenamiento en la redacción de ítems seguido de una revisión exhaustiva de los mismos, tanto con respecto a la precisión del contenido del ítem como a una evaluación de la medida en que cada ítem se ajusta al propósito de la prueba (Albanese & Case, 2016).

Evidentemente, los procedimientos de control de calidad para revisar el desempeño de los elementos de la prueba y eliminar los que tienen un desempeño deficiente son una fuente de evidencia de validez relacionada con el proceso de respuesta (Wrigley et al., 2012). La cruzada de colaboración institucional de Maastricht (Van der Vleuten et al., 2004) y Península Dental School (Ali et al., 2015) emplean dos ciclos rigurosos de control de calidad

antes y después de la administración de la prueba. Además, investigar la dificultad y la discriminación de los ítems de la prueba comprende evidencia de validez relacionada con la estructura interna (Albanese & Case, 2016; Wrigley et al., 2012).

La mayoría de las pruebas ofrecen la oportunidad de retroalimentación a los estudiantes: Maastricht (Van der Vleuten et al., 2004), Universidad McMaster (Canadá) (Blake et al., 1996), Alemania (Nouns & Georg, 2010), Península (Ali et al., 2015; Freeman et al., 2010; Freeman & Ricketts, 2010) y Mozambique (Aarts et al., 2010). Esta participación de los estudiantes en la toma de decisiones y el control de calidad de los ítems de la prueba, antes y después, se convierte en una fuente significativa de validez de constructo, relacionada con la consecuencia (Albanese & Case, 2016; Findyartini et al., 2014).

Para Plessas et al. (2015), es claro que a nivel mundial también hay una variación considerable en el número de ítems y la frecuencia de la administración de la prueba. Los diseñadores de pruebas de progreso usan diferentes frecuencias de prueba (típicamente dos, tres o cuatro pruebas por año) y diferentes tamaños de prueba (cantidad de elementos que varían entre 100 y 250) (Ricketts et al., 2010).

La validez de las pruebas de progreso relacionada con el proceso de respuesta puede ser afectada por la sincronicidad en la administración de la evaluación (Wrigley et al., 2012). Cuando la sincronicidad en la realización de la prueba no es factible, un software como un "navegador seguro" puede garantizar la validez de la prueba. Dicho software es utilizado por la prueba de progreso colaborativa de múltiples escuelas de EE. UU./Reino Unido, donde los estudiantes toman la prueba basada en la web en oleadas. El software bloquea la estación de trabajo para que los estudiantes no puedan copiar los materiales de prueba, consultar referencias en línea o enviar correos electrónicos a otros (Swanson et al., 2010).

La evidencia de la validez del constructo también puede ser respaldada por la correlación de los puntajes de las pruebas de progreso con el GPA acumulativo de los

estudiantes (Al Alwan et al., 2011; Findyartini et al., 2014). Al Alwan, et al. (2011) encontraron que las correlaciones son más altas en los estudiantes de niveles superiores en el plan de estudio en comparación con los estudiantes en niveles menores que oscilaron entre 0,38 y 0,77. Además, en otro estudio de Boshuizen et al. (1997), la prueba de progreso y la prueba de razonamiento clínico revelaron el mismo patrón de puntajes crecientes a lo largo de los años y tenían una alta intercorrelación.

El rendimiento de la prueba de progreso también se ha correlacionado significativamente con la prueba de admisión a la Facultad de Medicina de EE. UU. (MCAT) (Kerfoot et al., 2011). Se han observado correlaciones significativas entre las pruebas de progreso y diferentes exámenes de licencia, como los Exámenes nacionales de licencia de Alemania (Nouns & Georg, 2010), el examen de licencia del Consejo médico de Canadá (Blake et al., 1996) y el Examen de licencia médica de EE. UU. (USMLE) Paso 1 (Johnson et al., 2014; Kerfoot et al., 2011), y Paso 2 (Kerfoot et al., 2011). No obstante, las pruebas de progreso pueden identificar a los estudiantes con bajo rendimiento, como lo han demostrado Kerfoot et al., en cuyo estudio, la prueba de progreso identificó correctamente a aquellos estudiantes de segundo año que puntuaron por debajo de la media en el Paso 1 con una sensibilidad del 75 %, una especificidad del 77 % y un valor predictivo positivo del 41 % (Kerfoot et al., 2011).

Asegurar la equidad en las pruebas de progreso respalda aún más la validez de la prueba. Los estudios no han mostrado diferencias significativas entre el rendimiento de estudiantes masculinos y femeninos (Findyartini et al., 2014; Tomic et al., 2005). También, se puede concluir que una prueba de progreso de elementos, diseñada correctamente, no discrimina sistemáticamente a los estudiantes de medicina con discapacidades específicas de aprendizaje (Ricketts et al., 2010).

Definitivamente, unas pruebas de progreso con validez predictiva significativa son base fundamental para una continua revisión curricular integral. Los puntajes de las pruebas de progreso son una fuente importante de información para los autores de los ítems, los profesores del programa, el cuerpo docente y el comité de revisión (Wrigley et al., 2012). La prueba de progreso puede usarse como una herramienta de diagnóstico para el currículo (Al Alwan et al., 2011; Findyartini et al., 2014) y mostrar cómo existen diferentes patrones de aprendizaje en diferentes áreas del currículo (Ricketts et al., 2010). En áreas donde todos los estudiantes se desempeñan mal, el plan de estudios se puede revisar (Aarts et al., 2010) y, posteriormente, el efecto de los cambios se puede monitorear a través de puntajes de pruebas de progreso futuro (Ricketts et al., 2010). Todo lo anterior comprende fuentes valiosas de evidencia de validez relacionada con las consecuencias (Downing, 2003). En el caso de las colaboraciones institucionales, las pruebas de progreso pueden dar lugar a comparaciones de programas y currículos (Muijtjens et al., 2007; Schuwirth et al., 2011). Sin embargo, estas comparaciones entre currículos pueden ser una amenaza para la validez si no se toman en consideración ciertas cuestiones. Muijtjens, et al. observaron que en una prueba de progreso colaborativa interinstitucional los estudiantes obtuvieron mejores resultados en los ítems producidos en sus propias escuelas (Muijtjens et al., 2007). Por lo tanto, los elementos de la prueba de progreso estaban sujetos al sesgo de origen. Para abordar estos problemas, todas las escuelas participantes deben contribuir con la misma cantidad de elementos de prueba (Muijtjens et al., 2007). Del mismo modo, se ha demostrado que compartir los materiales de la prueba es viable, pero se deben hacer esfuerzos para eliminar la introducción de cualquier traducción y sesgo cultural que pueda comprometer la validez y la imparcialidad de la prueba (Verhoeven et al., 2005).

Igualmente, Johnson et al. (2014) describieron un uso innovador del examen integral de ciencias básicas (CBSE, por sus siglas en inglés) de la junta nacional de examinadores

médicos (NBME, por sus siglas en inglés), como una prueba de progreso durante el plan de estudios médicos previo a las prácticas en una nueva facultad de medicina. El objetivo principal del estudio era proporcionar validación externa de las evaluaciones de opción múltiple desarrolladas internamente. Este nuevo programa de medicina posee características similares a las estudiadas en la presente investigación, es decir, un plan de estudios integrado de educación médica de pregrado; las ciencias básicas fundamentales se enseñan a partir de módulos del cuerpo humano que aprovechan las sinergias tradicionales entre disciplinas (la fisiología se enseña junto con la anatomía, por ejemplo); utiliza un enfoque basado en sistemas (módulos S), que se enfoca en el estudio del proceso de la enfermedad y culmina cuando los estudiantes completan el Paso 1 del USMLE. Además, en la última fase del plan de estudios se traslada el conocimiento y las habilidades a la práctica, representado por pasantías y rotaciones en los últimos años académicos. Los investigadores observaron correlaciones significativas entre casi todos los exámenes internos y los puntajes de CBSE a lo largo del tiempo, así como con los puntajes del USMLE Paso 1. La fuerza de las correlaciones de los exámenes internos con el Paso 1 de CBSE y USMLE aumentó ampliamente con el tiempo durante el plan de estudios, también el progreso de los estudiantes medido por el CBSE fue lineal a lo largo del tiempo.

Otro ejemplo más cercano al de la presente investigación, tanto geográficamente como por la similitud del enfoque curricular, es el presentado por Dobronski, (2007) en su tesis de grado; allí, analizó siete años de realización de pruebas de progreso en una universidad ecuatoriana. Los resultados del Quarterly Profile Examination (QPE), realizado periódicamente por los estudiantes investigados, fueron comparados con los logros en las pruebas NBEM y USMLE de una cohorte análoga de la facultad de medicina de una universidad norteamericana. El autor logra concluir que: 1.- Las pruebas de progreso miden la adquisición y la retención de las ciencias básicas y clínicas de cada estudiante por

separado, y por lo tanto cada uno tiene una buena idea de su crecimiento; 2.-Las pruebas de progreso constituyen excelentes predictores de cómo será el desempeño en el examen USMLE 1, escenario al que se enfrentarán en el futuro; 3.- El QPE, como prueba de progreso, es una medida de la calidad del currículo y de la instrucción recibida en la institución médica; además, es capaz de evaluar certeramente los objetivos del programa educacional.

Finalmente, con las pruebas de progreso en la Universidad del Norte de Barranquilla (exámenes comprensivos en las ingenierías y los exámenes de fin de ciclo básico-clínico en medicina), a pesar de que vienen realizándose hace casi dos décadas y que a través de ellas se realizan continuas revisiones curriculares, no se han realizado investigaciones que valoren la capacidad predictiva de estas pruebas sobre pruebas estandarizadas como las pruebas de Estado (Saber Pro). Además, en la revisión bibliográfica realizada no se hallaron publicaciones que muestren la coherencia de estos exámenes, tanto con las pruebas de admisión (pruebas de estado - Saber 11) como con el crecimiento académico que adquiere el ingeniero o el médico a través de dichos planes de estudio en la Universidad del Norte.

5. Objetivos

5.1. Objetivo general

Determinar la relación entre el rendimiento en las pruebas Saber (11 y Pro) de médicos y su rendimiento académico en los ciclos básico y clínico de un programa de medicina.

5.2. Objetivos específicos

1. Estimar la relación entre el rendimiento en las pruebas Saber 11 de médicos y su rendimiento en la prueba Saber Pro, general y específicas, de médicos de un programa de medicina.
2. Estimar la relación entre el rendimiento en las pruebas Saber 11 de médicos y su rendimiento académico en los ciclos básico y clínico (Examen de suficiencia de fin de ciclo básico y clínico) de un programa de medicina.
3. Estimar la relación entre el rendimiento en las pruebas Saber 11 de médicos y su rendimiento académico (promedio académico ponderado a final de carrera) de un programa de medicina.
4. Estimar la relación entre el rendimiento académico en los ciclos básicos y clínicos de médicos de un programa de medicina.
5. Estimar la relación entre el rendimiento en las pruebas Saber Pro de médicos y su rendimiento académico en los ciclos básico y clínico (Examen de suficiencia de fin de ciclo básico y clínico) de un programa de medicina.
6. Estimar la relación entre el rendimiento en las pruebas Saber Pro de médicos y su rendimiento académico (promedio académico ponderado a final de carrera) de un programa de medicina.

6. Metodología

6.1. Contexto de la investigación

El programa de medicina de la Universidad del Norte posee en su plan de estudio dos pruebas de suficiencia en sus ciclos básico-profesional y profesional. Las evaluaciones se aplican en momentos finales de los correspondientes ciclos de competencias básicas (preclínicas) y clínicas, respectivamente. Dichas pruebas son asignaturas sin créditos académicos en el plan de estudios; sin embargo, el no aprobarlas luego de tres intentos se considera un bloqueo para la consecución de avances de carrera, completar el ciclo clínico y graduarse, respectivamente. Desde el punto de vista curricular, se han planteado correlaciones entre el promedio académico, el resultado en estas pruebas de suficiencia de ciclos y las pruebas Saber 11, con el fin de obtener información que retroalimente la estandarización de dichas evaluaciones programáticas (Martínez et al., 2015). Sin embargo, los datos obtenidos son muy relativos y dependientes de múltiples variables no académicas, lo que es coherente con lo expresado en varios informes publicados por el Instituto de Estudios en educación (IESE) de la Universidad del Norte.

Hace 8 años se creó el comité de evaluación para cada programa de la división de salud de la Universidad del Norte, entre otras razones, debido a las exigencias de los procesos de acreditación y a la búsqueda de mediciones objetivas que evidenciaran tanto el valor agregado por el programa académico del egresado como las debilidades curriculares a corregir. El desarrollo de estos entes de supervisión interna ha estimulado la esperanza de contar con instrumentos estandarizados que nos aporten información predictiva del futuro del egresado. Por ejemplo, a partir de la revisión semestral de la creación de las pruebas de fin de ciclo y la correlación secuencial de sus resultados con los datos obtenidos de estos mismos estudiantes en pruebas realizadas antes del inicio de la carrera y a portas del momento culmen

de egreso, respectivamente (Saber 11 y Saber Pro), se podría hacer evidente la capacidad predictiva de estos instrumentos.

6.2. Muestra

Este estudio trabaja con datos secundarios, es decir, no se recolectaron datos en el marco de este trabajo, sino que simplemente se procedió a la integración de información académica y socioeconómica de una cohorte de estudiantes de pregrado de medicina de una universidad del Caribe colombiano, disponible como registro administrativo en dicha universidad. Para efectos de este estudio, se define como cohorte de estudiantes al conjunto de alumnos que tomaron las pruebas Saber Pro en el mismo año, pudiendo o no coincidir el año en que iniciaron su pregrado o tomaron las demás pruebas utilizadas en este trabajo. Se trabajó con dos cohortes de estudiantes, una integrada por los que tomaron la prueba Saber Pro en 2018 y otra por aquellos que la hicieron en 2019.

El tamaño de la muestra es de 347 estudiantes, de los cuales 150 (43%) tomaron Saber Pro en 2018 y los restantes en 2019. La muestra de este estudio está compuesta en un 55% por mujeres, de las cuales 10.4% provienen de colegios públicos y 23.3% de colegios bilingües; además, 11.5% pertenecían al programa “Ser pilo paga”. La distribución por estrato es la siguiente: 10.1% pertenecen al estrato 1; 16.3 % al 2; 18.3 % al estrato 3; 20.4 % al estrato 4; 18.6 % al estrato 5 y 16.3 % al estrato 6. Además, 77 % tomó la prueba Saber 11 antes del segundo semestre de 2014, (periodo en el que las pruebas Saber 11 fueron modificadas, según se detalla en la siguiente sección) y el porcentaje restante lo hizo en el segundo semestre de dicho año o en años siguientes.

6.3. Instrumentos

La información administrativa a la que se accedió para la muestra de este estudio corresponde a resultados en diferentes pruebas de rendimiento académico que los estudiantes tomaron como parte de su desarrollo académico.

6.3.1. Pruebas Saber 11

La prueba Saber 11 es aplicada por el ICFES (Instituto colombiano para la evaluación de la educación) de forma anual a los estudiantes que estén cursando el último año de bachillerato en todo el territorio nacional. Esta evaluación se administra dos veces al año (semestre 1 y 2). Hasta el primer semestre de 2014 la prueba estaba conformada por 8 subpruebas (lenguaje, matemáticas, biología, química, física, filosofía, ciencias sociales, inglés) con una escala calificativa que va de 0 a 100 (tanto en las subpruebas como en la escala global–promedio simple de rendimiento en las subpruebas) y 5 niveles de desempeño. Desde el segundo semestre de 2014, Saber 11 se compone de cinco módulos (lectura crítica, matemáticas, sociales y ciudadanas, ciencias naturales e inglés). La escala de puntajes va de 0 a 100 para cada módulo evaluado y de 0 a 500 para el total global (promedio ponderado de los puntajes modulares), con 4 niveles de desempeño (ICFES, 2020a). Esta prueba es obligatoria, dado que todos los estudiantes deben tomarla como parte de su calendario académico y es considerada una prueba de alta consecuencia. Es decir, los resultados de las pruebas Saber 11 inciden en las oportunidades de acceso a universidades del país, dado que dichos puntajes son considerados por las casas de estudio a la hora de la admisión de los estudiantes. En este estudio se trabajó con los resultados globales o totales y en dos grupos: aquellos que tomaron la prueba hasta el primer semestre de 2014 y aquellos que lo hicieron luego de este punto, puesto que las escalas son diferentes y no comparables directamente.

6.3.2. Examen de suficiencia de Ciclo básico

El examen de suficiencia de ciclo básico es una asignatura sin créditos con dos posibles resultados (aprobada o reprobada) administrada a los estudiantes de quinto semestre de medicina. Esta evaluación se realiza 2 veces al año (semana 12 a 14 de cada semestre lectivo) y está a cargo del departamento de medicina de la universidad a través del Comité de evaluación. Esta prueba está conformada por 100 ítems que evalúan conocimiento de las

áreas de fundamentos de ciencias básicas médicas y sistemas corporales (nerviosos, músculo-esquelético, endocrinos, cardiovasculares, sanguíneos, reproductores, renales, inmunológicos, digestivos y respiratorios). La escala de calificaciones va de 1 a 5.

6.3.3. Examen de suficiencia de Ciclo clínico

El examen de suficiencia de ciclo clínico es una asignatura sin créditos con dos posibles resultados (aprobada o reprobada) administrada a los estudiantes de décimo semestre de medicina. Esta evaluación se realiza 2 veces al año (semana 10 a 12 de cada semestre lectivo) y está a cargo del departamento de medicina de la universidad a través del Comité de evaluación. Esta evaluación está conformada por 100 ítems que evalúan conocimientos de las áreas de salud pública (epidemiología, salud y adolescencia, salud ocupacional, gerencia en salud) y clínicas (quirúrgicas, ginecobstétricas, pediátricas, medicina interna y psiquiátricas). La escala de calificaciones va de 1 a 5.

6.3.4. Pruebas Saber Pro

La prueba Saber Pro también es administrada por el ICFES y está dirigida a estudiantes que han aprobado el 75% de los créditos de sus respectivos programas de formación universitaria profesional, independientemente de su condición étnica, discapacidad y estado de libertad. Este examen evalúa competencias genéricas y específicas en dos sesiones distintas. En la primera sesión se evalúan las genéricas y los estudiantes también completan un cuestionario de contexto. Esta sección de la prueba se compone de 5 módulos que evalúan competencias en lectura crítica, razonamiento cuantitativo, competencias ciudadanas, comunicación escrita e inglés. El puntaje global de las pruebas Saber Pro se estima a partir de estas competencias genéricas. Para la evaluación de competencias específicas (segunda sesión) hay módulos asociados a temáticas y contenidos específicos de acuerdo con el área de formación profesional de los estudiantes. En total, se disponen de 36 módulos para las diferentes áreas de formación de pregrado. La escala de puntajes va de 0 a

300 para cada módulo evaluado y para el total global, con 4 niveles de desempeño a excepción de inglés, que tiene 5 niveles de desempeño (ICFES, 2020b).

El Estado ha establecido como requisito de graduación para los estudiantes de pregrado el haber tomado los módulos de evaluación de competencias genéricas de las pruebas Saber Pro, mientras que los módulos de evaluación de las específicas son de carácter optativo (Ley 1324 de 2009), atendiendo que solo aquellos que se encuentran en el territorio nacional pueden tomar estos módulos. En este estudio se trabajó con los resultados por módulo, tanto de la prueba genérica como de la específica del área de salud, esta última compuesta por tres módulos:

(a) Atención primaria a la salud: evalúa la competencia que permite aplicar conceptos básicos de salud pública que determinan la priorización de las acciones a seguir de acuerdo con las condiciones de salud del individuo, la familia y la comunidad en el marco político y normativo nacional e internacional (ICFES, 2018)

(b) Fundamentación en diagnóstico y tratamiento: evalúa competencias de los estudiantes de programas profesionales de medicina para aplicar el conocimiento de ciencias básicas y clínicas en la elaboración de diagnósticos y proponer un plan de manejo para la recuperación de la salud humana, a partir de situaciones o casos clínicos de común ocurrencia en la práctica de médicos (ICFES, 2018).

(c) Promoción de la salud y prevención de la enfermedad: evalúa la competencia para aplicar conceptos básicos de promoción de la salud y prevención de la enfermedad que permitan la priorización de las acciones a seguir, de acuerdo con las condiciones de salud de las personas, las poblaciones y la normatividad vigente (ICFES, 2018).

6.3.5. Rendimiento promedio de la carrera

El rendimiento promedio de la carrera es, como su nombre lo indica, el promedio de las calificaciones obtenidas en todas las asignaturas matriculadas (la cantidad varía en función a los créditos asociados a las asignaturas que toma el estudiante) durante la vida académica del estudiante de pregrado, ponderadas por el número de créditos de cada asignatura matriculada. El mínimo de créditos requeridos para el grado es de 271. La escala de calificaciones del rendimiento promedio de la carrera va de 1 a 5.

6.4. Origen de los datos

Los datos utilizados en el presente estudio provienen de los registros administrativos de estudiantes matriculados en una universidad de la Costa caribe. En este caso, se integró toda la información disponible sobre los distintos resultados académicos de los estudiantes de las cohortes 2018 y 2019 utilizando como variable de anclaje el ID (código numérico) asignado al estudiante en el sistema de información de la universidad. La identidad de los estudiantes es desconocida para el investigador. Tal como se explicó, la cohorte de estudiantes está definida por aquellos estudiantes que tomaron la prueba Saber Pro en los años mencionados. A partir de esta nómina, se integraron los demás puntajes correspondientes a las mediciones descritas en el apartado anterior.

6.5. Plan de análisis de datos

La base de datos fue dividida en dos grupos en función al periodo en que tomaron las pruebas Saber 11, dado que esta prueba presentó varias particularidades: (i) cambió a partir del segundo semestre de 2014; (ii) los puntajes están en diferentes escalas; (iii) los puntajes no son comparables entre sí. En consecuencia, todos los análisis fueron conducidos dos veces en los casos en que los datos de esta prueba fueron utilizados. El primer conjunto de datos (estudiantes que tomaron la prueba Saber 11 antes del segundo semestre de 2014) consta de

267 observaciones y el segundo (estudiantes que tomaron la prueba Saber 11 a partir del segundo semestre de 2014) de 80.

El primer paso fue construir un índice considerado como proxy del nivel socioeconómico de los estudiantes a partir de la integración de las variables asociadas al tipo de colegio al que asistió el estudiante (público o privado, bilingüe o no), su estado de beca (si pertenece al programa Ser pilo paga o no) y el estrato al que pertenece (del 1 al 6), pues estas variables son todas indicativas del nivel socioeconómico y están –en consecuencia– altamente correlacionadas (Tabla 1). Para la construcción del índice se utilizó el análisis de componentes principales (PCA), que es una técnica de reducción de datos generalmente utilizada para casos como estos. El análisis paralelo (Horn, 1965) determinó que estas variables se pueden agrupar en un solo componente que explica el 84 % de la varianza total. El cálculo del índice fue realizado utilizando la regresión de Thurstone (1935), a partir de los pesos estimados con el PCA.

Tabla 1

Correlación entre variables Proxy del nivel socioeconómico

	SPP	Estrato	Privado	Bilingüe	Pesos*
SPP	1.00				-0.278
Estrato	-0.81	1.00			0.259
Privado	-0.70	0.69	1.00		0.268
Bilingüe	-0.91	0.70	0.91	1.00	0.286

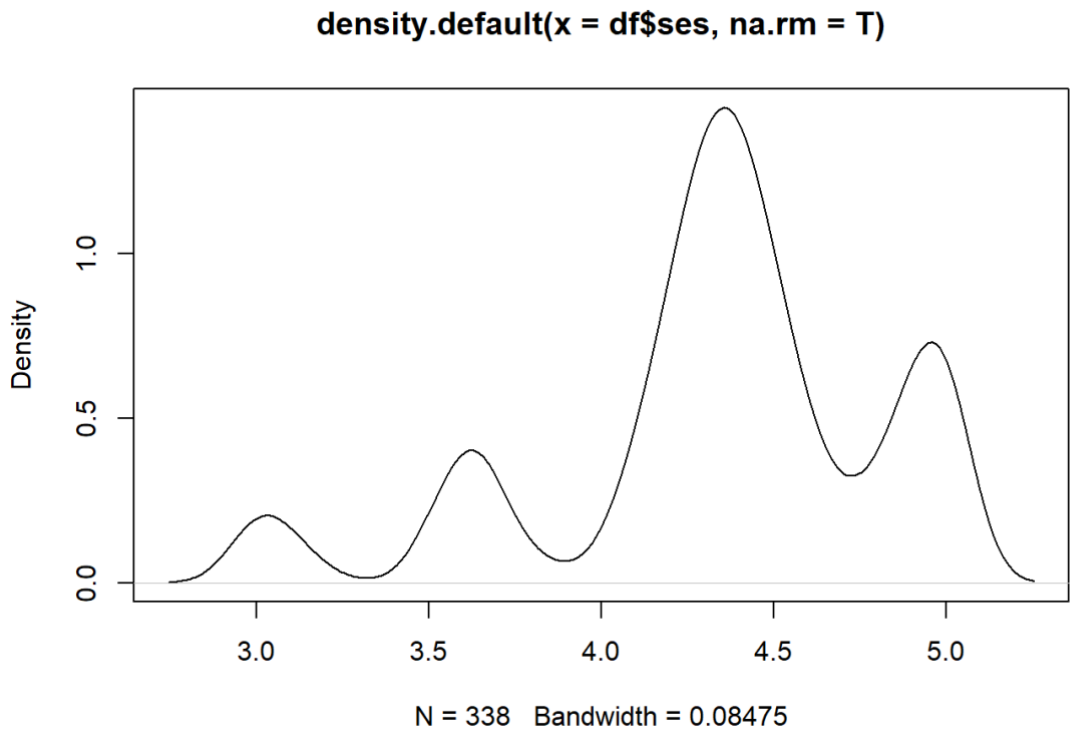
Nota: SPP: Ser Pilo Paga. Correlación tetracórica para las variables binarias entre sí (SPP, Privado, Bilingüe) y policórica para las categóricas ordinales (estrato) y las binarias.

* Pesos estimados a partir del PCA, utilizados para la estimación del índice proxy de nivel socioeconómico.

El índice estimado está sesgado hacia la izquierda (asimetría = 0.894), lo que quiere decir que la mayoría de los estudiantes se ubican en la parte superior de la distribución; mientras que la curtosis es 0.692 (Figura 1). El valor promedio del índice es 4.329 (DE=0.486), en un rango que va de 3 a 5.

Figura 1

Densidad de Kernel, índice de nivel socioeconómico estimado.



El segundo paso fue computar los descriptivos (promedio, mediana, desviación estándar, asimetría, kurtosis, etc.) de las variables asociadas al rendimiento en las diferentes mediciones descritas, el índice socioeconómico, y su cruce con variables categóricas (sexo, ser pilo paga, tipo de institución, estrato). Como tercer paso se generó la matriz de correlación de Pearson de todas las variables continuas (resultados en las 5 pruebas y el índice de nivel socioeconómico) a ser utilizadas en los análisis. Se evalúa la significancia al 0.05 y se utiliza como referencia para evaluar la fuerza de la relación. Se espera que todas las

correlaciones sean positivas, con excepción del nivel socioeconómico y las diferentes variables de rendimiento académico estudiadas en este trabajo.

De igual manera, como cuarto y principal paso se procedió a la estimación de las regresiones lineales definidas a partir de las restricciones que impone el tamaño de la muestra. En este sentido, se pretende medir la capacidad predictiva de las variables de interés controlado por características del estudiante, medidas a través de las variables de contexto (sexo y nivel socioeconómico). Por ejemplo, se pretende estimar en qué medida el rendimiento en el examen de suficiencia de ciclo básico predice el rendimiento en las pruebas Saber Pro específicas y si esta predicción difiere entre hombres y mujeres, para esto la ecuación de regresión sería:

$$Yi = \alpha + \delta Pi + jXji + i$$

Donde Y_i representa el puntaje en alguna de las variables de interés atendiendo su ubicación temporal (es decir, examen de suficiencia de fin de ciclo básico o clínico, promedio de fin de carrera, prueba Saber Pro general –y sus elementos conformantes– o Saber Pro específicas del área de medicina); α representa el rendimiento promedio de la variable de interés que asume posición de dependiente en la regresión (Y_i); P es la variable de interés que asume la posición de variable independiente o explicativa (es decir, pruebas que temporalmente se ubican antes que la estudiada como variable dependiente); δ representa el cambio en una unidad de la variable dependiente que se asocia al cambio en una unidad de la variable explicativa de interés; X_{ji} representa las características de los estudiantes (sexo, nivel socioeconómico); β_i es la variación en la predicción de la variable dependiente de interés que se asocia a la característica personal en cuestión; y ε_i es el error o diferencia entre el valor observado de *Y* el estimado por la regresión. Este error es la parte de *Y* que es predicha por otras variables que no están en el modelo, además del error de medición.

Se espera que los puntajes logrados en las distintas pruebas sean predictores positivos y significativos del rendimiento en las pruebas ubicadas temporalmente, con posterioridad a la considerada predictora en el análisis de regresión. Todas las pruebas de hipótesis sobre significancia estadística de los coeficientes estimados serán conducidas con un nivel de significancia del 0,05.

6.6. Valores perdidos

No existen valores perdidos en los datos utilizados, dado que provienen de registros administrativos.

7. Resultados

7.1. Descriptivos

7.1.1. Saber 11

Los descriptivos de la prueba Saber 11 se presentan por separado para cada escala (antes y después del segundo semestre de 2014), atendiendo que estas escalas no son directamente comparables, tal como se explicó en la sección de metodología. El promedio de los estudiantes que tomaron la prueba en la escala vieja (antes del 2º semestre de 2014) fue 60 puntos (DE=5,5) con puntajes que varían entre 50 y 81 puntos. La mediana coincide con el promedio para esta variable, mientras que la asimetría y kurtosis también son iguales (0,6) entre sí (Tabla 2; Figura 2a). Al comparar el rendimiento promedio de esta prueba entre estudiantes por sexo, tipo de institución de la que proviene (privada o pública; bilingüe o no) y estrato, se verifica que el rendimiento promedio sea prácticamente el mismo (Tabla 3) con diferencias entre grupos que como máximo llegan a 3 puntos.

El promedio para esta prueba en la nueva escala (después del segundo semestre de 2014) fue de 360 puntos (DE = 25) para los estudiantes que conforman este estudio, con un intervalo que va de 250 a 420 puntos. La mediana nuevamente es igual al promedio, mientras que se puede observar que la distribución está sesgada hacia la izquierda (-1,7) y con curtosis positiva (6,2) indicando una distribución leptocúrtica (Figura 2b). En la comparación de rendimiento promedio por sexo y tipo de institución, las diferencias se mantienen bajas. Sin embargo, se observa cierta variabilidad en el promedio cuando los estudiantes son agrupados por estrato. Por ejemplo, los estudiantes del estrato 2 son los que mejor promedio tuvieron en la escala nueva de Saber 11 (367 puntos), mientras que el estrato 3 tiene promedio de 358 puntos, y los demás estratos presentan promedios alrededor de los 361 puntos. Estas diferencias entre estratos responden a la política de admisión de la institución para estudiantes becarios, quienes generalmente provienen de estratos bajos.

7.1.2. Examen de ciclo básico

La nota del examen de fin de ciclo básico para el grupo que forma parte de este estudio varía entre 3,0 y 4,9; mientras que el promedio es de 3,5 (DE = 0,4) y la mediana de 3,4. Se observan valores para asimetría de 0,8 y curtosis de 0,5 para esta variable (Figura 2c). Cuando los estudiantes son agrupados en función al sexo, tipo de colegio del que provienen y estrato, los valores promedios de este indicador son muy similares, pues los promedios se ubican entre 3,4 a 3,6 en cualquiera de los casos.

7.1.3. Examen de ciclo clínico

La nota del examen de fin de ciclo clínico para el grupo que forma parte de este estudio varía entre 3,0 y 5,0; mientras que el promedio es de 3,6 (DE = 0,5) al igual que la mediana. Se observa valor de 0,9 para asimetría y valor cercano a cero de curtosis (0,3) para esta variable (Figura 2d). Cuando los estudiantes son agrupados en función al sexo, tipo de colegio del que provienen y estrato, los valores promedios de este indicador son prácticamente iguales, dado que los promedios son 3,6 o 3,7 en cualquiera de los casos.

Tabla 2

Descriptivos de evaluaciones de desempeño de estudiantes de pregrado

Evaluaciones	Mínimo	Cuartil 1	Mediana	Media	Cuartil 3	Máximo	DE	Asimetría	Curtosis
Saber 11 escala vieja	50	56	60	60	64	81	5.5	0.6	0.62
Saber 11 escala nueva	250	352	359.5	380.2	372.5	420	25.1	-1.69	6.16
Examen de ciclo básico	3	3.2	3.4	3.47	3.7	4.9	0.36	0.81	0.51
Examen de ciclo clínico	3	3.2	3.6	3.62	3.9	5	0.47	0.87	0.3
Rendimiento promedio de la carrera	3.36	3.67	3.81	3.84	3.99	4.47	0.22	0.41	-0.43
Saber Pro global	123	170	181	179.8	191.5	228	17.2	-0.48	0.46
Comunicación escrita	0	140.5	166	162.1	185	300	37.02	-0.98	5.84
Razonamiento cuantitativo	79	159	180	174.8	193	229	27.45	-0.89	0.94
Lectura crítica	86	171	189	186.7	203	300	24.78	-0.09	2.57
Competencias ciudadanas	0	165	181	178	194	300	26.87	-1.04	6.25
Inglés	0	182	198	197.3	210	300	27.85	0.06	10.52
Saber Pro específicas									
Atención primaria a la salud	0	170	165	191	192	232	38.12	-2.75	10.24
Fundamentación en diagnóstico y tratamiento	0	141.5	160	156	178	250	37.08	-2.12	7.55
Promoción de la salud y prevención de enfermedades	0	164	179	172	192	226	37.6	-3.02	11.29

Figura 2

Densidad de Kernel, evaluaciones de desempeño de estudiantes de pregrado

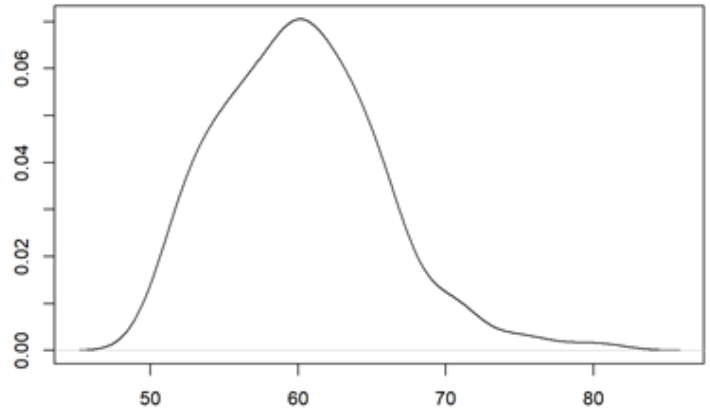


Figura 2a. Saber 11 vieja escala

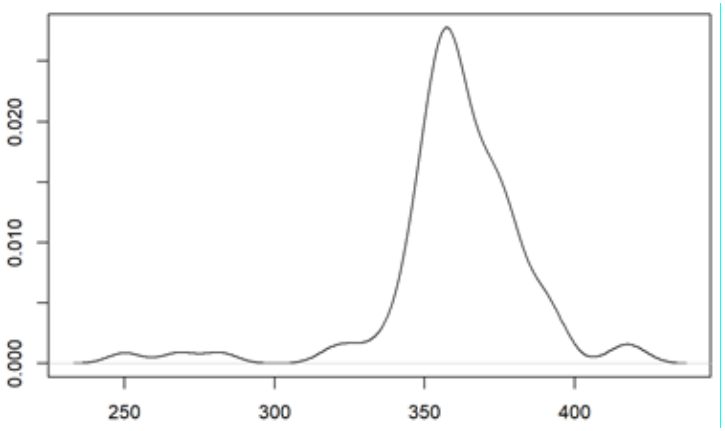


Figura 2b. Saber 11 nueva escala

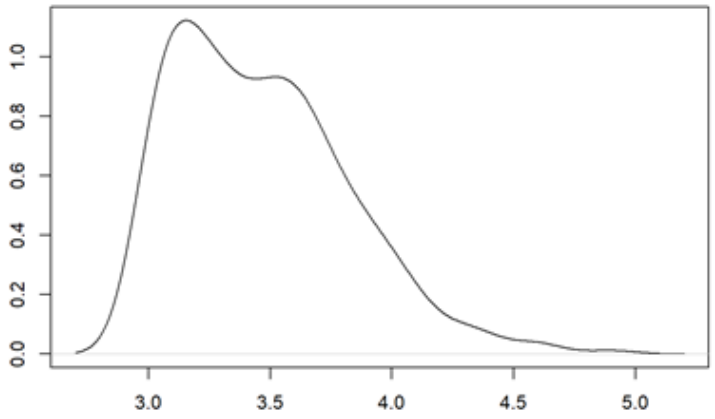


Figura 2c. Examen de ciclo básico

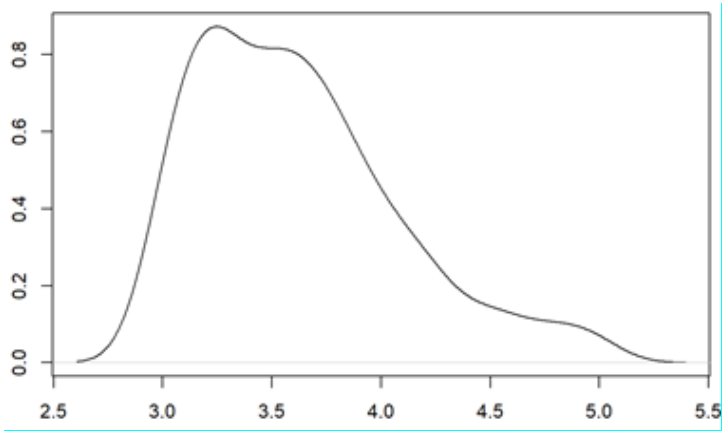


Figura 2d. Examen de ciclo clínico

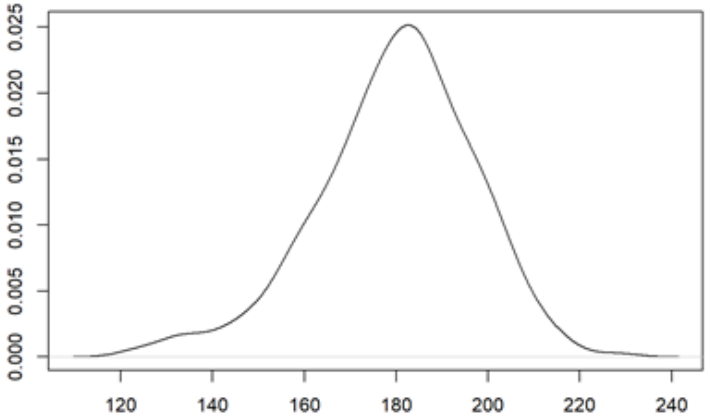


Figura 2e. Pruebas Saber Pro

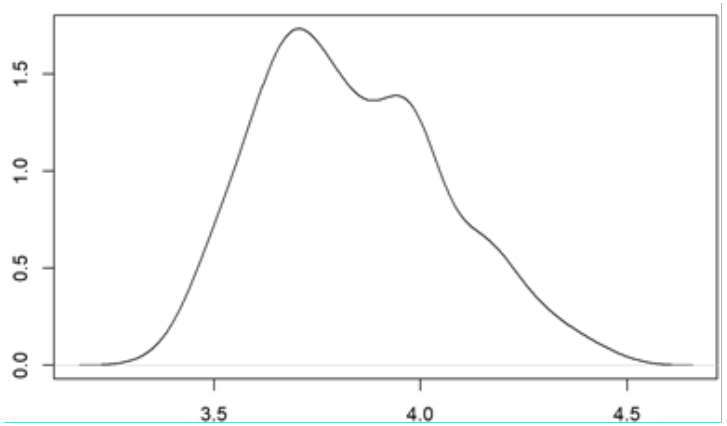


Figura 2f. Promedio de fin de carrera

7.1.4. Saber Pro

Los descriptivos de la prueba Saber Pro se presentan por separado para las pruebas denominadas genéricas (que sirven para la estimación del puntaje global), y para las pruebas específicas que deben realizar los estudiantes de los programas de salud (que se encuentran dentro del territorio nacional). El promedio del puntaje global fue 180 puntos (DE = 17) con valores que varían entre 123 y 228 puntos. La mediana (181) es prácticamente igual al promedio para esta variable, mientras que la asimetría y kurtosis son -0,5 y 0,5, respectivamente (Tabla 2; Figura 2e). Al comparar el rendimiento promedio de esta prueba entre estudiantes por sexo, tipo de institución de la que proviene (privado o público; bilingüe o no) y estrato, se verifica que el rendimiento promedio es prácticamente el mismo (Tabla 3), con diferencias entre grupos que como máximo llegan a 3 puntos.

El promedio para esta prueba en la nueva escala (después del segundo semestre de 2014) fue de 360 puntos (DE = 25) para los estudiantes que conforman este estudio, con un intervalo que va de 250 a 420 puntos. La mediana nuevamente es igual al promedio, mientras que se observa que la distribución está sesgada hacia la izquierda (-1,7) y con curtosis positiva (6,2) indicando una distribución leptocúrtica. En la comparación de rendimiento promedio por sexo y tipo de institución, las diferencias se mantienen bajas. Sin embargo, se observa cierta variabilidad en el promedio cuando los estudiantes son agrupados por estrato. Por ejemplo, los estudiantes del estrato 2 son los que mejor promedio tuvieron en la escala nueva de Saber 11 (367 puntos), mientras que el estrato 3 tiene promedio de 358 puntos, y los demás estratos presentan promedios alrededor de los 361 puntos. Estas diferencias entre estratos responden a la política de admisión de la institución para estudiantes becarios, que generalmente provienen de estratos bajos.

7.1.5. Rendimiento promedio de la carrera

El rendimiento promedio de carrera para el grupo que forma parte de este estudio varía entre 3,4 y 4,5; mientras que el promedio es de 3,8 (DE = 0,2) al igual que la mediana. Se observan valores bajos (cercanos a cero) de asimetría (0,4) y curtosis (-0,4) para esta variable (figura 2f). Cuando los estudiantes son agrupados en función al sexo, tipo de colegio del que provienen y estrato, los valores promedios de este indicador son prácticamente iguales, dado que los promedios son 3,8 o 3,9 en cualquiera de los casos.

Tabla 3

Descriptivos de evaluaciones de desempeño por características demográficas de los estudiantes

Variables	Sexo		Colegio		Colegio		Estrato					
	Hombre	Mujer	Público	Privado	Bilingüe	No bilingüe	1	2	3	4	5	6
Saber 11 escala vieja	60.8	59.6	59.9	60.2	59.4	62.1	61	58.1	59.9	61.2	59.6	61.1
Saber 11 escala nueva	360.2	360.2	362.2	359.6	360.3	358.7	361	366.8	358.4	351.2	360.9	381.8
Examen de ciclo básico	3.5	3.5	3.6	3.5	3.5	3.4	3.6	3.5	3.5	3.5	3.5	3.4
Examen de ciclo clínico	3.6	3.6	3.5	3.6	3.6	3.7	3.6	3.6	3.6	3.7	3.6	3.7
Rendimiento Promedio de la carrera	3.8	3.9	3.9	3.8	3.8	3.9	3.9	3.8	3.8	3.8	3.8	3.9
Saber Pro global	179.9	179.7	182.1	179.5	178	185.7	183.6	177.9	180.6	178.4	178.2	182
Comunicación escrita	158	165.4	161.6	162.1	160.8	166.1	158.2	156.9	172.2	157.3	161	163.6
Razonamiento cuantitativo	177.9	172.3	182.3	174	174.1	177.3	185.3	178.5	172.8	176.3	167.2	176.3
Lectura crítica	186.7	186.6	186.9	186.6	185.3	191.2	191	188.1	186.9	186.4	183.9	185.5
Competencias ciudadanas	179.3	177.2	189.2	176.8	177.8	179.2	190.7	179.8	174.1	174.8	182.2	173.5
Inglés	197.9	196.8	190.9	198.1	192	214.9	193.3	186.4	197.1	197	197.5	211.3
Saber Pro específicas												
Atención primaria a la salud	171.1	171	182.1	169.8	173.6	162.8	183.6	176	180.2	170.6	165.1	159.8
Fundamentación en diagnóstico y tratamiento	158.1	154.2	166.6	154.7	157.6	150.8	167.8	158.4	163.5	155.4	151.4	146.7
Promoción de la salud y prevención de enfermedades	169.6	173.9	179	171.2	174.5	163.7	176.2	174.2	177.5	174.5	170.6	160.1

7.2. Correlaciones

La asociación entre las mediciones de desempeño fue medida con el coeficiente de Pearson. En la mayoría de los casos (74% de los coeficientes), la correlación fue significativa ($p<0.05$), tal como se indica con los sombreados en la Tabla 4. Según Cohen (1992), la correlación es pequeña si el valor absoluto del coeficiente se ubica entre 0,10 y 0,29 ($|0,10|\leq r<|0,30|$); es mediana si está entre 0,30 y 0,49 ($0,30\leq r<0,50$); y es alta si es 0,50 o mayor ($r\geq 0,50$). Se entiende entonces que valores inferiores a 0,10 pueden ser considerados triviales ($r<|0,10|$). Sólo son interpretados los coeficientes estadísticamente significativos.

De los 89 coeficientes, 18 resultaron triviales; es decir, inferior a 0,10. Por su parte, la mayoría de las correlaciones (38 coeficientes) está entre |0,10| y |0,20|; mientras que un número relativamente elevado (19 coeficientes) se ubica entre |0,21| y |0,29|. Es decir, 57 de los 89 coeficientes presentan una correlación valorada como pequeña. Complementariamente, 17 coeficientes de correlación están entre |0,30| y |0,49|, mostrando así una relación de tamaño mediano. Finalmente, varios coeficientes (12 en total) reportan valores considerados altos ($r\geq|0,50|$) según los parámetros de Cohen (1992).

Tabla 4

Correlación de Pearson

Variables	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15
1 **Saber Pro global**	1														
2 Comunicación escrita	0.60	1													
3 Razonamiento cuantitativo	0.56	0.15	1												
4 Lectura crítica	0.60	0.11	0.21	1											
5 Competencias ciudadanas	0.61	0.11	0.17	0.37	1										
6 Inglés	0.61	0.16	0.20	0.24	0.26	1									
Saber Pro Específicas															
7 Atención primaria a la salud	0.28	0.14	0.21	0.17	0.22	0.10	1								
8 Fundamentación en diagnóstico y tratamiento	0.26	0.12	0.16	0.17	0.18	0.15	0.75	1							
9 Promoción de la salud y prevención de la enfermedad	0.22	0.18	0.10	0.15	0.13	0.11	0.77	0.71	1						
10 **Examen de ciclo básico**	0.32	0.21	0.23	0.08	0.25	0.19	0.21	0.38	0.18	1					
11 **Examen de ciclo clínico**	0.08	0.08	0.12	0.02	0.04	0.01	0.03	0.06	0.00	0.16	1				
12 **Rendimiento promedio de la carrera**	0.47	0.20	0.40	0.20	0.33	0.29	0.14	0.34	0.09	0.64	0.15	1			
13 **Nivel socioeconómico**	-0.02	0.02	-0.14	-0.04	-0.14	0.23	0.19	-0.16	-0.11	-0.11	0.10	-0.11	1		
14 **Saber 11 escala vieja**	0.55	0.17	0.45	0.36	0.36	0.41	0.19	0.25	0.16	0.39	0.06	0.45	0.15	1	
15 **Saber 11 escala nueva**	0.55	-0.02	0.46	0.49	0.42	0.25	0.28	0.22	0.08	0.01	-0.07	0.43	0.14	na	1

Las correlaciones más bajas se dieron con el rendimiento en el examen de ciclo clínico, pues los valores son inferiores a 0,10 en 9 de las 12 posibles asociaciones con esta variable ($r<|0,10|$); es decir prácticamente todas las correlaciones con el examen de ciclo clínico fueron triviales según los parámetros de Cohen (1992). Además, la correlación entre examen de ciclo clínico y la escala nueva de Saber 11 es negativa, lo que contradice lo esperado.

La vieja escala de Saber 11 reporta una correlación de tamaño mediano ($r = 0,39$) con las pruebas de ciclo básico, el cual tiene una asociación mediana ($r = 0,33$) con las pruebas Saber Pro. De igual forma, se observa una correlación mediana ($r = 0,38$) entre el rendimiento en el examen de ciclo básico y el del módulo específico de Saber Pro para medicina «fundamentación en diagnóstico y tratamiento». El rendimiento promedio de fin de carrera es el que presenta mayor número de correlaciones de tamaño mediano ($r\geq|0,30|$) con las pruebas Saber 11 (nueva y vieja escala) y Saber Pro (global, razonamiento cuantitativo, competencias ciudadanas, inglés y fundamentación en diagnóstico y tratamiento).

De manera esperada, las mayores correlaciones (altas o $r\geq|0,50|$ en todos los casos) se dieron entre el puntaje global de Saber Pro y los módulos que la componen (competencias genéricas). Además, Saber 11 (nueva y vieja escala) mostraron correlaciones altas ($r\geq|0,50|$) con Saber Pro (global); y medianas ($r\geq|0,30|$) con algunos de sus componentes (razonamiento cuantitativo, lectura crítica y competencias ciudadanas). Los módulos específicos de Saber Pro también reportan correlaciones altas entre sí. Finalmente, se observa una correlación alta ($r = 0,64$) entre el rendimiento en la prueba de ciclo básico y el rendimiento promedio de fin de carrera; en contraposición al tamaño pequeño de la relación que esta última variable tiene con el rendimiento en la prueba de ciclo clínico ($r = 0,14$).

7.3. Regresiones

Cabe recordar que el análisis de regresión se condujo para cada conjunto de datos por separado (Tablas 5 y 6): con el conjunto de estudiantes que tomaron la prueba Saber 11 antes del segundo semestre de 2014 (n = 267) y con los estudiantes que tomaron dicha prueba posterior al primer semestre de 2014 (n = 80), conforme lo señalado en la sección del plan de análisis de datos. Cada coeficiente de regresión indica el cambio (positivo o negativo) en la variable dependiente ante una variación (positiva o negativa) de la variable predictora, cuando dicho coeficiente es estadísticamente significativo ($p<0,05$), y manteniendo todo lo demás constante. Si el coeficiente es negativo, se tiene una relación inversa entre variable predictora y variable dependiente; y si es positivo, una relación directa.

Los resultados del modelo muestran que no existe diferencia significativa entre hombres y mujeres en las evaluaciones estudiadas en este trabajo (Saber 11, examen de ciclo básico, examen de ciclo clínico, Saber Pro, y promedio de fin de carrera) en el primer grupo de datos, excepto cuando en el modelo está presente la variable razonamiento cuantitativo en el modelo de predicción del promedio de fin de carrera. En otras palabras, el sexo del estudiante es predictor (Tabla 5, CE = 0,038; p = 0,04) del rendimiento promedio de fin de carrera, solo cuando razonamiento cuantitativo es parte del modelo.

En el segundo conjunto de datos, el sexo es variable predictora significativa del rendimiento promedio de fin de carrera cuando en el modelo están presentes algunas de las siguientes dimensiones de Saber Pro: comunicación escrita (Tabla 5, CE = 0,072; p = 0,04), razonamiento cuantitativo (Tabla 5, CE = 0,071; p = 0,04), inglés (Tabla 5, CE = 0,073; p = 0,04), atención primaria a la salud (Tabla 5, CE = 0,072; p = 0,04), promoción de la salud y prevención de enfermedades (Tabla 5, CE = 0,072; p = 0,04), fundamentación en diagnóstico y tratamiento (Tabla 5, CE = 0,08; p = 0,01). En todos los casos, ser mujer predice una

mejora del rendimiento de fin de carrera de alrededor de 0,07 unidades cuando en el modelo están incluidas ciertas dimensiones de Saber Pro.

Se encontró diferencia significativa entre estudiantes de distintos niveles socioeconómicos en los resultados de la prueba Saber 11 (Tabla 6, CE = 2,450; p = 0,01), ciclo básico (Tabla 6, CE = -0,143; p = 0,02), promedio de fin de carrera (Tabla 5) y Saber Pro: global (Tabla 5, CE = 7,000; p =0,01), inglés (Tabla 5, CE = 31,800; p = 0,00), atención primaria a la salud (Tabla 5, CE = -18,400; p = 0,01) y fundamentación en diagnóstico y tratamiento (Tabla 5, CE = -17,500; p = 0,02), para el primer grupo de datos. Nótese los valores negativos para algunos de los coeficientes, lo que indica que a medida que los estudiantes pertenecen a niveles socioeconómicos más elevados su rendimiento tiende a disminuir en dichas dimensiones (examen de ciclo básico, atención primaria a la salud y fundamentación en diagnóstico y tratamiento); algunas de estas relaciones ya se mostraron negativas en la correlación bivariada. En el segundo grupo de datos (estudiantes con puntajes Saber 11 en la nueva escala), el nivel socioeconómico deja de ser estadísticamente significativo en todos los casos (p>0,05). Esto se debe al ingreso de estudiantes “Ser pilo paga” al programa, quienes provienen de estratos bajos, pero que en general presentan rendimiento similar a los de estratos más elevados.

En la relación intertemporal de las pruebas analizadas, se tiene que el rendimiento en las pruebas Saber 11 predicen positiva y significativamente los resultados en la prueba de ciclo básico (Tabla 6, CE = 0,028; p = 0,00, primer grupo de datos), el rendimiento promedio de carrera (ambos grupos de datos), y el rendimiento en las pruebas Saber Pro (ambos grupos de datos), tanto en el puntaje global como en casi todas sus dimensiones (excepción: comunicación escrita -ambos grupos de datos- y promoción de la salud y prevención de enfermedades -segundo grupo de datos); mientras que no predice estadísticamente el rendimiento en las pruebas de ciclo clínico (ambos grupos de datos, Tabla 6).

Por su parte, los resultados de la prueba de ciclo básico en el primer conjunto de datos predicen significativamente las pruebas de ciclo clínico (Tabla 6, CE = 0,238; p = 0,005), promedio de fin de carrera, y Saber Pro: puntaje global (Tabla 5, CE = 8,1; p = 0,002), comunicación escrita (Tabla 6, CE=18,7; p = 0,01), competencia ciudadana (Tabla 6, CE = 10,6; p = 0,003), y promoción de la salud y prevención de enfermedades (Tabla 6, CE = 29,2; p = 0,00). En el segundo conjunto de datos, el rendimiento en el examen de ciclo básico es predictor significativo del promedio de fin de carrera (Tabla 5), de inglés (Tabla 5, CE = 33,4; p = 0,01), promoción de la salud y prevención de enfermedades (Tabla 5, CE = 18,205; p = 0,04) y fundamentación en diagnóstico y tratamiento (Tabla 5, CE = 40,645; p =0,00).

Finalmente, la prueba de ciclo clínico predice estadísticamente el promedio de fin de carrera únicamente en el primer grupo de datos, dejando de ser predictor significativo en el segundo conjunto de datos (Tabla 5). Esta prueba tampoco predice de manera significativa (p>0,05) ningún puntaje en las pruebas Saber Pro ni en sus dimensiones generales o específicas en ninguno de los dos grupos de datos analizados, excepto en el modelo donde la variable dependiente es el rendimiento en razonamiento cuantitativo, en el primer grupo de datos, en donde el examen de ciclo clínico es predictor positivo y significativo (Tabla 5, CE = 6,60; p = 0,04).

Tabla 5

Regresiones para promedio de fin de carrera, Saber Pro general y específico

Variables	PS = Saber Pro (M1)			PS = Comunicación escrita (M2)		
	CE	**SE**	**p-valor**	**CE**	**SE**	**p-valor**
Datos previos al segundo semestre de 2014 (n=267)						
VD: Promedio de fin de carrera	R2 = 0,556			R2 =0.544		
PS	0,00	0,00	**0,00**	0,00	0,00	0,15
Examen de ciclo clínico	0,04	0,02	**0,03**	0,04	0,02	**0,03**
Examen de ciclo básico	0,32	0,03	**0,00**	0,33	0,03	**0,00**
Saber 11	0,01	0,00	**0,01**	0,01	0,00	**0,00**
Nivel socioeconómico	0,09	0,03	**0,00**	0,11	0,03	**0,00**
Sexo (mujer=1)	0,03	0,02	0,10	0,03	0,02	0,10
VD: PS	R2 =0,350			R2 = 0.078		
Examen de ciclo clínico	1,60	1,90	0,41	4,70	5,00	0,35
Examen de ciclo básico	8,10	2,60	**0,00**	18,70	7,00	**0,01**
Saber 11	1,49	0,18	**0,00**	0,76	0,47	0,11
Nivel socioeconómico	7,00	2,60	**0,01**	2,10	7,00	0,76
Sexo (mujer=1)	1,70	1,80	0,35	8,70	4,80	0,07
Datos posteriores al primer semestre de 2014 (n=80)						
VD: Promedio de fin de carrera	R2 = 0,555			R2 =0.557		
PS	0,00	0,00	0,98	0,00	0,00	0,62
Examen de ciclo clínico	-0,05	0,04	0,21	-0,05	0,04	0,20
Examen de ciclo básico	0,42	0,06	**0,00**	0,42	0,06	**0,00**
Saber 11	0,00	0,00	**0,00**	0,00	0,00	**0,00**
Nivel socioeconómico	0,00	0,03	0,91	0,00	0,03	0,90
Sexo (mujer=1)	0,07	0,03	0,05	0,07	0,03	**0,04**
VD: PS	R2 =0.328			R2 =0.015		
Examen de ciclo clínico	-1,17	3,29	0,72	-5,19	8,55	0,55
Examen de ciclo básico	5,98	4,54	0,19	9,15	11,81	0,44
Saber 11	0,32	0,06	**0,00**	-0,03	0,15	0,86
Nivel socioeconómico	-0,15	2,35	0,95	1,43	6,12	0,82
Sexo (mujer=1)	-1,55	2,81	0,58	2,59	7,30	0,72

cont.

Variables	PS = Razonamiento cuantitativo (M3)			PS = Lectura crítica (M4)		
	CE	**SE**	**p-valor**	**CE**	**SE**	**p-valor**
Datos previos al segundo semestre de 2014 (n=267)						
VD: Promedio de fin de carrera	R2 =0.550			R2 = 0.543		
PS	0,00	0,00	**0,02**	0,00	0,00	0,19
Examen de ciclo clínico	0,04	0,02	0,05	0,04	0,02	**0,02**
Examen de ciclo básico	0,33	0,03	**0,00**	0,34	0,03	**0,00**
Saber 11	0,01	0,00	**0,00**	0,01	0,00	**0,00**
Nivel socioeconómico	0,11	0,03	**0,00**	0,11	0,03	**0,00**
Sexo (mujer=1)	0,04	0,02	**0,04**	0,03	0,02	0,08
VD: PS	R2 = 0.227			R2 = 0.134		
Examen de ciclo clínico	6,60	3,20	**0,04**	-0,10	3,00	0,97
Examen de ciclo básico	5,40	4,50	0,22	1,30	4,10	0,75
Saber 11	1,99	0,30	**0,00**	1,52	0,28	**0,00**
Nivel socioeconómico	1,80	4,50	0,69	2,50	4,10	0,55
Sexo (mujer=1)	-4,70	3,00	0,12	2,50	2,80	0,37
Datos posteriores al primer semestre de 2014 (n=80)						
VD: Promedio de fin de carrera	R2 =0.569			R2 =0.568		
PS	0,00	0,00	0,14	0,00	0,00	0,15
Examen de ciclo clínico	-0,06	0,04	0,17	-0,05	0,04	0,21
Examen de ciclo básico	0,41	0,06	**0,00**	0,39	0,06	**0,00**
Saber 11	0,00	0,00	**0,00**	0,00	0,00	**0,00**
Nivel socioeconómico	0,01	0,03	0,76	0,00	0,03	0,95
Sexo (mujer=1)	0,07	0,03	**0,04**	0,07	0,03	0,06
VD: PS	R2 =0.237			R2 =0.318		
Examen de ciclo clínico	3,20	5,51	0,56	0,44	6,31	0,94
Examen de ciclo básico	5,57	7,61	0,47	-24,01	8,71	**0,01**
Saber 11	0,41	0,10	**0,00**	0,55	0,11	**0,00**
Nivel socioeconómico	-4,56	3,94	0,25	-1,22	4,52	0,79
Sexo (mujer=1)	0,20	4,70	0,97	-4,71	5,39	0,38

cont.

Variables	PS = Competencia ciudadana (M5)			PS = Inglés (M6)		
	CE	**SE**	**p-valor**	**CE**	**SE**	**p-valor**
Datos previos al segundo semestre de 2014 (n=267)						
VD: Promedio de fin de carrera	R2 = 0.547			R2 = 0.542		
PS	0,00	0,00	0,05	0,00	0,00	0,29
Examen de ciclo clínico	0,04	0,02	**0,02**	0,05	0,02	**0,02**
Examen de ciclo básico	0,33	0,03	**0,00**	0,33	0,03	**0,00**
Saber 11	0,01	0,00	**0,00**	0,01	0,00	**0,00**
Nivel socioeconómico	0,11	0,03	**0,00**	0,09	0,03	**0,00**
Sexo (mujer=1)	0,03	0,02	0,06	0,03	0,02	0,08
VD: PS	R2 =0.150			R2 =0.340		
Examen de ciclo clínico	0,00	3,50	1,00	-3,60	2,80	0,20
Examen de ciclo básico	10,60	4,80	**0,03**	4,80	3,90	0,22
Saber 11	1,61	0,33	**0,00**	1,57	0,27	**0,00**
Nivel socioeconómico	-3,10	4,90	0,53	31,80	3,90	**0,00**
Sexo (mujer=1)	-0,50	3,30	0,87	2,50	2,70	0,35
Datos posteriores al primer semestre de 2014 (n=80)						
VD: Promedio de fin de carrera	R2 =0.556			R2 =0.558		
PS	0,00	0,00	0,77	0,00	0,00	0,49
Examen de ciclo clínico	-0,05	0,04	0,22	-0,05	0,04	0,24
Examen de ciclo básico	0,42	0,06	**0,00**	0,40	0,06	**0,00**
Saber 11	0,00	0,00	**0,00**	0,00	0,00	**0,00**
Nivel socioeconómico	0,00	0,03	0,95	0,00	0,03	1,00
Sexo (mujer=1)	0,07	0,03	0,05	0,07	0,03	**0,04**
VD: PS	R2 =0.211			R2 =0.169		
Examen de ciclo clínico	3,92	4,91	0,43	-8,60	8,82	0,33
Examen de ciclo básico	4,45	6,78	0,51	33,40	12,18	**0,01**
Saber 11	0,33	0,09	**0,00**	0,37	0,15	**0,02**
Nivel socioeconómico	-4,68	3,51	0,19	8,76	6,31	0,17
Sexo (mujer=1)	-0,76	4,19	0,86	-6,39	7,53	0,40

cont.

Variables	PS = Atención primaria a la salud (M7)			PS = Promoción de la salud y prevención de enfermedades (M8)		
	CE	**SE**	**p-valor**	**CE**	**SE**	**p-valor**
Datos previos al segundo semestre de 2014 (n=267)						
VD: Promedio de fin de carrera	R2 = 0.548			R2 = 0.546		
PS	0,00	0,00	**0,03**	0,00	0,00	0,79
Examen de ciclo clínico	0,05	0,02	**0,02**	0,04	0,02	**0,02**
Examen de ciclo básico	0,34	0,03	**0,00**	0,33	0,03	**0,00**
Saber 11	0,01	0,00	**0,00**	0,01	0,00	**0,00**
Nivel socioeconómico	0,10	0,03	**0,00**	0,11	0,03	**0,00**
Sexo (mujer=1)	0,04	0,02	0,06	0,03	0,02	0,07
VD: PS	R2 =0.081			R2 =0.068		
Examen de ciclo clínico	2,00	5,10	0,70	3,20	4,80	0,50
Examen de ciclo básico	13,50	7,10	0,06	29,20	6,60	**0,00**
Saber 11	1,21	0,48	**0,01**	1,05	0,45	**0,02**
Nivel socioeconómico	-18,40	7,10	**0,01**	-9,30	6,60	0,16
Sexo (mujer=1)	3,10	4,90	0,53	-2,90	4,50	0,52
Datos posteriores al primer semestre de 2014 (n=80)						
VD: Promedio de fin de carrera	R2 =0.562			R2 =0.563		
PS	0,00	0,00	0,29	0,00	0,00	0,26
Examen de ciclo clínico	-0,05	0,04	0,22	-0,05	0,04	0,27
Examen de ciclo básico	0,40	0,06	**0,00**	0,40	0,06	**0,00**
Saber 11	0,00	0,00	**0,00**	0,00	0,00	**0,00**
Nivel socioeconómico	0,00	0,03	0,92	0,00	0,03	0,99
Sexo (mujer=1)	0,07	0,03	**0,04**	0,07	0,03	**0,04**
VD: PS	R2 =0.117			R2 =0.073		
Examen de ciclo clínico	-1,76	5,74	0,76	-7,41	6,40	0,25
Examen de ciclo básico	13,80	7,92	0,09	18,20	8,84	**0,04**
Saber 11	0,25	0,10	**0,01**	0,08	0,11	0,47
Nivel socioeconómico	0,59	4,11	0,89	3,61	4,58	0,43
Sexo (mujer=1)	-0,92	4,90	0,85	-1,34	5,47	0,81

cont.

Variables	PS = Fundamentación en diagnóstico y tratamiento (M9)		
	CE	**SE**	**p-valor**
Datos previos al segundo semestre de 2014 (n=267)			
VD: Promedio de fin de carrera	R2 = 0.540		
PS	0,00	0,00	0,06
Examen de ciclo clínico	0,04	0,02	**0,02**
Examen de ciclo básico	0,34	0,03	**0,00**
Saber 11	0,01	0,00	**0,00**
Nivel socioeconómico	0,10	0,03	**0,00**
Sexo (mujer=1)	0,04	0,02	0,05
VD: PS	R2 =0.148		
Examen de ciclo clínico	-0,30	5,10	0,95
Examen de ciclo básico	9,90	7,10	0,17
Saber 11	1,13	0,48	**0,02**
Nivel socioeconómico	-17,50	7,20	**0,02**
Sexo (mujer=1)	7,40	4,90	0,13
Datos posteriores al primer semestre de 2014 (n=80)			
VD: Promedio de fin de carrera	R2 =0.630		
PS	0,00	0,00	**0,00**
Examen de ciclo clínico	-0,03	0,04	0,43
Examen de ciclo básico	0,29	0,06	**0,00**
Saber 11	0,00	0,00	**0,00**
Nivel socioeconómico	0,00	0,03	0,96
Sexo (mujer=1)	0,08	0,03	**0,01**
VD: PS	R2 =0.330		
Examen de ciclo clínico	-7,25	5,49	0,19
Examen de ciclo básico	40,64	7,58	**0,00**
Saber 11	0,20	0,10	**0,04**
Nivel socioeconómico	0,60	3,93	0,88
Sexo (mujer=1)	-2,97	4,69	0,53

Tabla 6

Regresiones para examen de ciclo clínico, básico y Saber 11

Variables	Datos previos al segundo semestre de 2014 (n=267)			Datos posteriores al primer semestre de 2014 (n=80)		
	CE	**SE**	**p-valor**	**CE**	**SE**	**p-valor**
VD: Examen de ciclo clínico	R2 =0,047			R2 =0,139		
Examen de ciclo básico	0,24	0,09	**0,01**	0,14	0,16	0,40
Saber 11	0,00	0,01	0,79	0,00	0,00	0,75
Nivel socioeconómico	0,14	0,09	0,11	0,13	0,08	0,10
Sexo (mujer=1)	0,08	0,06	0,19	-0,26	0,09	**0,01**
VD: Examen de ciclo básico	R2 =0,172			R2 =0,035		
Saber 11	0,03	0,00	**0,00**	0,00	0,00	0,91
Nivel socioeconómico	-0,14	0,06	**0,02**	-0,08	0,06	0,16
Sexo (mujer=1)	0,05	0,04	0,20	-0,06	0,07	0,37
VD: Saber 11	R2 =0,032			R2 =0,022		
Nivel socioeconómico	2,45	0,98	**0,01**	-5,71	4,54	0,21
Sexo (mujer=1)	-1,04	0,68	0,12	1,50	5,35	0,78

8. Discusión de Resultados

El valor predictivo de las pruebas de progreso sobre los exámenes estandarizados realizados por el Estado en los médicos egresados de una institución universitaria, es un tema de actualidad. Su importancia radica en la capacidad potencial de supervisión y seguimiento a múltiples factores en los que las instituciones de educación superior podrían enfocar sus modificaciones para lograr, de mejor manera, los RAE en el médico recién egresado (Pinilla & Parra, 2009) y con ello, parte de los criterios requeridos por los sistemas de acreditación tanto nacionales como internacionales. El Observatorio de Educación del Caribe colombiano de la Universidad del Norte, por ejemplo, asegura que a través de estas correlaciones se puede evidenciar el valor agregado que brindan las instituciones, en el recorrido del estudiante desde el primer semestre hasta el futuro profesional (Valencia et al., 2020).

Las pruebas de progreso intrainstitucionales que durante la formación demuestran los avances parciales o totales del estudiante a través de su plan de estudios no han sido contundentemente significativas, a pesar de las expectativas lógicas que nacen de la obtención de sus datos (Plessas 2015). Para Dobronski (2007), analizar la información de estas pruebas y luego compararla con pruebas estandarizadas externas resulta invaluable para el control de factores metodológicos en la formación médica.

A excepción de la investigación de Dobronski, (2007), realizada para su tesis de grado para optar a su licenciatura de medicina, no se encontraron referentes bibliográficos actualizados sobre estudios en egresados de programas de medicina de los últimos 10 años, con criterios equiparables al nuestro. Por lo tanto, se organiza la discusión de los resultados en apartados que analizan el poder predictivo, de forma intertemporal, del rendimiento de los estudiantes en las pruebas Saber 11, examen de ciclo básico, examen de ciclo clínico, el GPA de carrera y la prueba Saber Pro con sus diferentes dimensiones. En este sentido, se describirán las correlaciones significantes desde las regresiones analizadas.

8.1. Poder predictivo de la prueba Saber 11

El rendimiento en las pruebas Saber 11 predice positiva y significativamente los resultados en la prueba de ciclo básico (CE = 0,028; p = 0,00), el rendimiento promedio de carrera (CE = 0,01; p = 0,00), y el rendimiento en las pruebas Saber Pro (CE = 0,032; p = 0,00,), tanto en el puntaje global como en casi todas sus dimensiones (excepción: comunicación escrita -ambos grupos de datos- y promoción de la salud y prevención de enfermedades -segundo grupo de datos).

8.1.1. Entre resultados de Saber 11 y Saber Pro

Al igual que en nuestra investigación, los estudiantes que se destacaron académicamente en el colegio continúan destacándose en la IES. Se resalta que los dos grupos de datos mostraron correlaciones altas ($r \geq |0,50|$) con el puntaje global de Saber Pro. Bahamón y Reyes (2014) encontraron en un grupo de 68 estudiantes de psicología, diferencias estadísticamente significativas en áreas homologables de estas pruebas externas como lenguaje, escritura y lectura crítica. Sin embargo, es difícil hablar de valor predictor en esa relación, ya que se analizaron de manera individual los puntajes de las competencias evaluadas por medio de una Prueba T Student. Mientras que, Castro et al. (2018) encontraron una relación positiva entre los puntajes de la prueba Saber 11 y Saber Pro, esto es: a mayor puntaje en la prueba Saber 11, mayor puntaje en Saber Pro (ß=9.698; $p<0.05$). Esta observación fue realizada a partir de una regresión exploratoria en 1806 estudiantes que tomaron la prueba Saber 11 para los años 2005 y 2006, y luego la prueba Saber Pro en los años 2009-2010 en el departamento de Antioquia (Colombia). De la misma manera, Melo et al. (2014), en un análisis de eficiencia de ciertos factores de la educación colombiana, hallaron una correlación positiva entre los resultados de las pruebas Saber 11 con el resultado promedio del grupo de estudiantes que presentaron la prueba Saber Pro en el segundo

semestre de 2011, de 0,88. Inclusive, por grupos de referencia, dicha correlación supera el 0,9 para las carreras de Medicina, Derecho y Ciencias económicas y administrativas.

Estos desenlaces se encuentran en sintonía con otros trabajos que utilizan la base de datos Saber (Ramírez, 2014). Es explicable entonces, que las pruebas Saber 11 sean definitivamente la valoración más usada para la admisión en las instituciones de educación superior en Colombia, haciendo parte de las evaluaciones estandarizadas proporcionadas por el Estado colombiano para la supervisión longitudinal de la educación. A pesar de que, en algunos casos como el de Acautt et al. (2021), donde se contrastaron los resultados totales y por área de estas mismas variables (Saber 11 y Saber Pro) en 14 estudiantes de enfermería, no se encontraron diferencias significativas, tanto en la comparación de promedios en general como en la del rendimiento en dimensiones específicas. Este último estudio, metodológicamente no describe la realización de algún análisis estadístico específico de predicción; seguramente, por el pequeño número tomado como muestra. Por lo tanto, dicha investigación no reviste el poder metodológico para emitir conclusiones relevantes.

8.1.2. Entre resultados de Saber 11 y pruebas de progreso (exámenes de fin de ciclo) y GPA (promedio final de carrera):

En la Universidad del Norte de Barranquilla, desde hace más de 15 años (con excepción de los dos semestres que por la pandemia causada por el virus Covid-19 no se tuvo en cuenta), esta prueba es el único y principal insumo de clasificación de admisibilidad a los programas de pregrado, incluyendo la carrera de medicina. Estas razones hacían prever un hipotético poder predictivo de la prueba Saber 11 con relación a las pruebas de progreso y ponderaciones intra-facultad de medicina (examen de ciclo básico; examen de ciclo clínico y promedio final de carrera). El coeficiente de regresión que resultó positivo significativamente para el examen de final de ciclo básico y el promedio final de carrera contrastó con el nulo poder de predicción estadística sobre el examen de fin de ciclo clínico.

Similares fueron los resultados expuestos por Kerfoot et al. (2011), donde la prueba de admisión a la Facultad de Medicina de EE. UU. (MCAT) se ha correlacionado significativamente con el rendimiento de la prueba de progreso. Adam et al. (2015) mostraron evidencias análogas, donde la variable de selección para admisión de la carrera “puntuación total de la Prueba de Aptitud Clínica del Reino Unido (UKCAT)” fue un predictor significativo de los componentes de los exámenes escritos del año 4, clínicos del año 5 y predictor único del examen escrito del quinto año, tomados estos tres últimos como pruebas de progreso, lo que los presenta como coherentes con nuestros resultados. En el mismo estudio de Adam et al. (2015), otra variable de selección denominada "puntuación académica HYMS", basada en el rendimiento académico previo (educación secundaria), fue predictor significativo de las dos primeras pruebas de progreso anteriormente mencionadas. El puntaje HYMS es, teóricamente, una variable menos confiable por sus diferentes evaluadores (docente de cada nivel escolar y de cada escuela) y diversos contextos de obtención (procedencia, modelo escolar y social).

Asimismo, Dabaliz et al. (2017) utilizando un análisis de regresión lineal multivariado como el nuestro, encontraron variables de preingreso como los test de inglés, IELTS (p=0,04, B=0,08) y el TOEFL (p=0,017, B=0,01), que predijeron significativamente el rendimiento ponderado de los años preclínicos (1, 2 y 3 / básicos), de la carrera de medicina en una universidad de Arabia Saudita.

Caso aparte presenta el hecho que la prueba Saber 11, en los dos grupos de datos, no predice estadísticamente el rendimiento en la prueba de ciclo clínico. Hipotéticamente, esta última debería comportarse como una prueba de progreso y evidenciarse una correlación positiva significativa con los cambios en el rendimiento de la prueba Saber 11, como lo mostrado por Adam et al. (2015) en referencia a la prueba de selección para admisión de la carrera de Medicina en universidades del Reino Unido (UKCAT) específicamente, frente al

examen clínico del quinto año. Sin embargo, el puntaje HYMS basado en el rendimiento académico previo, no predijo significativamente la misma prueba clínica del quinto año de la carrera. El análisis de la diferencia de poder predictivo entre las dos pruebas de admisión en Reino Unido, sobre la prueba clínica de quinto año en el estudio de Adam et al. (2015), puede ser de ayuda para explicar la ausencia de poder predictor de la prueba Saber 11, frente al examen de fin de ciclo clínico de nuestra investigación.

Seguidamente, se puede considerar lo siguiente: 1. la prueba de aptitud clínica de Reino Unido (UKCAT) es una prueba de admisión a los programas de medicina de Reino Unido, e intenta evidenciar aptitudes deseables en el estudiante de nuevo ingreso a los programas de salud ante situaciones clínicas. 2. Tanto el puntaje HYMS en el estudio de Adam et al. (2015), en Reino Unido, como la prueba Saber 11 en Colombia, valoran el proceso de enseñanza aprendizaje de la educación secundaria previa. 3. el examen clínico de quinto año y el examen de fin de ciclo clínico son realizados en momentos análogos del plan de carrera de ambos estudios; un punto temporal de la carrera donde el estudiante ha desarrollado, en gran porcentaje, su capacidad de análisis de casos clínicos y sus pacientes; la valoración de la mencionada capacidad aplicada es el fundamento de estas pruebas clínicas.

La analogía entre los dos estudios, basada en las diferencias y similitudes mencionadas en las tres consideraciones, tiende a aclarar la ausencia del valor predictivo entre pruebas que, aunque son secuenciales en la evolución académica y cognitiva del estudiante de medicina, en este caso, se basan en las diferencias específicas en las capacidades a evaluar. Sin embargo, deberemos volver al tema del examen de fin de ciclo clínico cuando discutamos su propio valor predictivo hacia otras pruebas.

8.1.3. Saber 11 y GPA

Generalmente, los promedios de final de carrera (GPA) son resultados ponderados del transcurso del estudiante a través del plan de estudios, donde se entrega valor relativo

(CRÉDITOS, ULAS) a cada componente (asignatura). Siendo el GPA la evidencia numérica del comportamiento académico de un estudiante, es relevante que la prueba Saber 11, en nuestra investigación, predice significativamente el rendimiento del médico a través de su carrera. A conclusiones similares llegaron Tomatis et al. (2016), quienes por un análisis de regresión lineal demostraron que hay asociación positiva entre la nota de ingreso que resultó ser predictor del promedio general de la carrera, donde por cada punto que aumenta la nota del ingreso, aumenta 0,38 el promedio de la carrera ($p<0,0001$). En nuestro caso y el de Tomatis et al. (2016), se tomaron cortes que iniciaron y finalizaron sin interrupción ni pérdida el plan de estudios completo.

Los resultados fueron diferentes para Dabaliz et al. (2017) que, en un análisis de regresión lineal multivariado, encontraron pruebas como los puntajes de la escuela secundaria y puntajes de pruebas estandarizadas, como las de la Prueba Nacional de Logro y la Prueba de Aptitud General, no mostraron valor predictivo del GPA en un programa de medicina de una universidad de Arabia Saudita. También para Wilkinson et al. (2011), el test de admisión a la licenciatura de medicina en Australia (UMAT) tuvo una validez predictiva limitada para el rendimiento académico. Si bien la correlación era positiva en estos últimos trabajos, esta fue parcial y especialmente visible (no significativa) en los primeros años de carrera.

8.2. Poder predictivo del examen de fin de ciclo básico

Como se esperaba, los resultados de la prueba de ciclo básico en el primer conjunto de datos predicen significativamente las pruebas de ciclo clínico, el promedio de fin de carrera y el puntaje global de Saber Pro (además sus dimensiones de comunicación escrita, competencia ciudadana y promoción de la salud y prevención de enfermedades). Esta prueba de progreso de los RAE del ciclo básico, en el plan de estudios de medicina estudiado, posee la valoración de los conceptos tanto básicos biomédicos (morfología-bioquímica-fisiología-microbiología-patología y farmacología) como los fundamentales de la salud pública

(atención primaria- familia sociedad y salud- bioestadística). Dichos conceptos se entregan desde una visión sistémica integrada, basada en adaptaciones metodológicas, tanto de la secuencia funcional de las áreas disciplinares en cada asignatura como de situaciones clínicas transversales que facilitan su entendimiento en contexto. Lo anterior pudiese explicar que de acuerdo a lo adquirido en este trayecto del programa se pueda predecir positiva y significativamente las pruebas intertemporalmente posteriores a la prueba de ciclo básico (ciclo clínico, GPA y Saber Pro global), inclusive la dimensión de Saber Pro relacionada con atención primaria.

De forma similar a nuestro estudio, Johnson et al. (2014) encontraron correlaciones significativas entre evaluaciones internas de escogencia múltiple y única respuesta de las evaluaciones de un nuevo programa de medicina (curricularmente parecido por su enfoque basado en sistemas), con el CBSE y el USMLE Paso 1, lo que demuestra la validez predictiva de sus evaluaciones de progreso. De igual manera, Al Alwan et al. (2011) y Findyartini et al. (2014) encontraron fuerte correlación de los puntajes de las pruebas de progreso con el GPA acumulativo de los estudiantes. Al Alwan, et al. (2011) encontraron que las correlaciones eran más altas en los estudiantes de niveles superiores en el plan de estudio. También, Boshuizen et al. (1997) revelaron el mismo patrón de puntajes crecientes a lo largo de los años entre la prueba de progreso y la prueba de razonamiento clínico, teniendo una alta correlación intertemporal.

El valor predictivo del examen de fin de ciclo básico sobre la prueba Saber Pro, tomado como examen estatal de educación superior, se comportó de igual forma que en investigaciones internacionales, en las que se han descrito correlaciones significativas entre las pruebas de progreso y diferentes exámenes de licencia, como los exámenes nacionales de licencia de Alemania (Nouns & Georg, 2010), el examen de licencia del Consejo médico de

Canadá (Blake et al., 1996) y el examen de licencia médica de EE. UU. (USMLE) Paso 1 (Johnson et al., 2014; Kerfoot et al., 2011) y Paso 2 (Kerfoot et al., 2011).

En el segundo conjunto de datos, si bien las condiciones generales académicas son similares, el 50% de dicha submuestra son estudiantes de la primera promoción del programa Ser pilo paga; esto, sumado a la debilidad metodológica declarada en relación al tamaño de la población del segundo grupo de datos, hace que las pocas diferencias cognitivas de entrada y progreso no permitan evidenciar algunas distinciones de nivel. Por ello, el rendimiento en el examen de ciclo básico es predictor significativo del promedio de fin de carrera y de dimensiones de la prueba Saber Pro como la de inglés (CE = 33,4; p = 0,01), promoción de la salud y prevención de enfermedades (CE = 18,205; p = 0,04) y fundamentación en diagnóstico y tratamiento (CE = 40,645; p =0,00); dejando de ser significativas las diferencias ante la prueba del ciclo clínico y el puntaje global de Saber Pro.

Si analizamos estos resultados de ambos grupos de datos desde la perspectiva de predicción del examen de fin de ciclo básico, hacia las competencias propias de un médico, es relevante el valor predictivo sobre la dimensión específica de Saber Pro para médicos. Allí, por cada punto de aumento en la prueba de ciclo básico, el rendimiento en esa dimensión específica de Saber Pro (fundamentación en diagnóstico y tratamiento) fue de 10 y 40 puntos para el primer y segundo grupo de datos, respectivamente. En esta misma línea, algunos investigadores norteamericanos intentaron correlacionar evaluaciones internas de facultades de medicina, similares a las que representan procesos básicos y/o clínicos con el aprendizaje esperado al final de carrera (Greatrix et al., 2021; Wang, et al., 2021) y/o su desempeño práctico, al final de la formación habilitante (Taber et al., 2020; Tamblyn et al., 2002), la mayoría con resultados estadísticamente poco significativos o no tan evidentes como el nuestro.

8.3. Poder predictivo del examen de fin de ciclo clínico

La prueba de ciclo clínico en la presente investigación, considerando su relación de intertemporalidad, debería mostrar una correlación significativa con el promedio de fin de carrera, el puntaje de la prueba Saber Pro global y las dimensiones análogas a las valoradas en el fin de ciclo clínico (por ej. fundamentación en diagnóstico y tratamiento). Sin embargo, únicamente predice de manera estadísticamente significativa el promedio de fin de carrera, en el primer grupo de datos más no en el segundo conjunto de datos. Además, en ninguno de los dos grupos de datos analizados predice el puntaje en las pruebas Saber Pro global o sus dimensiones específicas, de manera significativa ($p>0,05$). La calidad de la variable criterio, y/o el nivel de solapamiento que existe entre la variable predictora y el criterio desde la teoría podría también estar afectando el poder predictivo hipotético; factores que no están contemplados en el análisis.

Recordemos que, para Thorndike (2005), la utilidad de una prueba como predictora depende de qué tan informativa sea la prueba (1) y qué tan bien se relacione con la variable criterio (2). En ese sentido (1), un 80 % de la prueba de ciclo clínico está fundamentada en informar la capacidad de análisis del examinado, específicamente, ante situaciones clínicas de las áreas de desarrollo de la práctica médica (Ginecología-Obstetricia, Pediatría, Cirugía, Medicina Interna, Urgencias, Psiquiatría); el restante 20% está diseñado para la valoración de las capacidades de decisión ante situaciones de Salud Ocupacional, Gerencia en Servicios de Salud, Epidemiología e investigación. Estas aclaraciones nos sitúan ante una prueba que, si bien es escrita, evalúa constructos prácticos específicos que seguramente el egresado médico debió obtener y desarrollar a través de sus rotaciones y pasantías ante el paciente real y/o simulado. De allí que en la segunda característica, la prueba de ciclo clínico como variable predictora solo tendría utilidad con una de las variables criterio mencionadas al inicio de este apartado; específicamente, debería estar relacionada con la dimensión de fundamentación en

diagnóstico y tratamiento de la prueba Saber, por recurrencia (estudiantes que se examinan para Saber Pro y la prueba clínica a final del quinto año de la carrera) e intertemporalidad (estudiantes que se examinan para la prueba de ciclo clínico a final del quinto año de carrera y en el transcurso del sexto año de la carrera).

Podríamos inferir que, de las fuentes de validez que APA (2014) menciona como estándares para pruebas, hay dos directamente afectadas: "(ii) relaciones con otras variables, dado que es altamente probable que existan otras pruebas que persiguen el mismo objetivo (o el objetivo exactamente opuesto) que el de la prueba en cuestión, la relación con variables externas a la prueba también provee evidencias de validez"; y "(v) consecuencias de las pruebas, que refiere a la interpretación de los puntajes en relación al propósito que se buscaba (por ejemplo, determinar la competencia de las personas en una dimensión específica) así como sus consecuencias indirectas (por ejemplo, ranquear universidades en función a puntajes que no buscaban eso)". Esta última está evidenciada si recordamos dos características administrativas, tanto de la prueba hipotéticamente predictora (prueba de ciclo clínico) como de la variable criterio (prueba Saber Pro).

La prueba de ciclo clínico (variable hipotéticamente predictora) al final, solo requiere ser aprobada; no hay incentivo especial, más allá de premiar el día de la graduación al mejor puntaje (diploma especial). La penalidad posible solo se produce si se reprueba por tercera vez (el estudiante aparece como no activo hasta aprobar). Recordemos también que, en la valoración de la prueba Saber Pro (variable criterio), hay un conflicto de intereses desarrollados entre instituciones, que busca mejorar los rankings (criterio de acreditación) y estudiantes que perciben la prueba Saber Pro exclusivamente como requisito de grado (Ley 1324, 2009), sin puntajes mínimos exigidos; solo algunas convocatorias para residencias médicas (último lustro) tienen en cuenta el puntaje obtenido en la prueba Saber Pro en un porcentaje mínimo.

Finalmente, y sin evidencias estadísticas que lo confirmen, queda la percepción de la necesidad de revisar la prueba de ciclo clínico, desde su diseño, revisión y ejecución. Los hechos, de no haber sido predecible de manera estadísticamente significativa por la prueba Saber 11 y el no poseer valor predictivo estadísticamente significativo, ante las pruebas que hipotéticamente lo podía hacer nos obliga a revisar la prueba como tal (lo que no era objetivo directo de la presente investigación).

9. Conclusiones

1. El rendimiento en la prueba de ciclo básico predice significativamente el rendimiento en pruebas subsiguientes: clínico, promedio de fin de carrera y dimensiones de Saber Pro. En el primer conjunto de datos predice las pruebas de ciclo clínico (CE = 0,238; p = 0,005), promedio de fin de carrera y Saber Pro: puntaje global (CE = 8,1; p = 0,002), comunicación escrita (CE = 18,7; p = 0,01), competencia ciudadana (CE = 10,6; p = 0,003), y promoción de la salud y prevención de enfermedades (CE = 29,2; p = 0,00). En el segundo conjunto de datos, el rendimiento en el examen de ciclo básico es predictor significativo del promedio de fin de carrera, de inglés (CE = 33,4; p = 0,01), promoción de la salud y prevención de enfermedades (C = 18,205; p = 0,04) y fundamentación en diagnóstico y tratamiento (CE = 40,645; p = 0,00).
2. En general, todas las correlaciones con el examen de ciclo clínico fueron triviales según los parámetros de Cohen (1992). Las correlaciones más bajas se dieron con su rendimiento, pues los valores son inferiores a 0,10 en 9 de las 12 posibles asociaciones con esta variable ($r<|0,10|$).
3. La prueba de ciclo clínico no tiene poder predictivo de evaluaciones subsiguientes. Esta prueba solo predice el promedio de fin de carrera en el primer grupo de datos, dejando de ser predictor significativo en el segundo conjunto de datos; tampoco predice ningún puntaje en las pruebas Saber Pro, en sus dimensiones generales o específicas, en ninguno de los dos grupos de datos analizados.

10. Recomendaciones

La educación médica ha evolucionado desde un enfoque en el proceso de enseñanza a un enfoque en los resultados y la demostración de competencias. Predecir para modelar los progresos ha demostrado ser útil en muchos campos y la educación médica no es la excepción. Las pruebas de progreso evalúan el crecimiento del alumno a lo largo del tiempo mediante la administración de exámenes de contenido y dificultad similares en el plan de estudios. Sin embargo, los problemas metodológicos pueden limitar la generalización de las pruebas de progreso a contextos de mayor escala y su capacidad para predecir el desempeño futuro en los exámenes.

El estudio estadístico de los resultados en las mencionadas pruebas de progreso realizadas por los estudiantes de una universidad del caribe colombiano, insertadas en su plan de estudios, han resultado de gran valor para sentar las bases para la supervisión de los procesos dentro del marco de su diseño filosófico y de ejecución. Los datos que emergen continuamente del departamento de medicina y de la división de ciencias de la salud de una institución con la estructura mencionada son muy valiosos en cantidad y calidad, porque podrían alimentar la creación de especialidades en educación médica o un énfasis para educación en salud de la Maestría en Educación en la institución.

10.1 Consideraciones metodológicas para investigaciones futuras:

1. La investigación sugiere seguir **alimentando la base de datos** (por ej. más cohortes); también, más **datos que evidencien las dimensiones académicas** de las que se componen las pruebas.

2. **Conocer y analizar la percepción de los egresados, docentes y administrativos (incluyendo el comité curricular y de evaluación)** sería de invaluable importancia en el análisis global del ejercicio de estas pruebas de fin de ciclo básico y clínico.

10.2 Consideraciones conceptuales para el programa:

1. Si bien la mayoría de las expectativas e hipótesis formuladas previamente se confirmaron, las premisas no probadas y de difícil explicación (por ausencia de insumos externos a la investigación), colocan al nivel de indispensable una **revisión de los procesos de planeación, recolección de preguntas, revisión y puesta en marcha de ambas pruebas**. La verificación interna debe dirigirse a una homogeneidad en la política administrativa de las mismas; obviamente, manteniendo la coherencia entre lo pertinente / valorable del aprendizaje, a través de ellas y los diferentes momentos de su aplicación.

2. La ubicación temporal de ambas pruebas y su proceso punitivo, al ser reprobadas, conlleva a momentos puntuales al final de cada ciclo (básico y clínico) que no son retroalimentados en busca de recuperación de aprendizajes. Esta mejoría podría lograrse a través de situaciones de evaluación (formativas), estructuralmente similares, que retroalimenten previamente y en múltiples ocasiones la progresión hacia la ejecución de cada prueba de fin de ciclo. Concretamente, se hace referencia a **la implementación de evaluaciones formativas integrales por semestre cursado**, que incluso puedan contener interrogantes (adrede) de niveles superiores (que se puedan inferir o no) para monitorizar (auto y heteroevaluación) el progreso previo del estudiante.

11. Bibliografía

Aarts, R., Steidel, K., Manuel, B., & Driessen, E. (2010). Progress testing in resource-poor countries: A case from Mozambique, *Medical Teacher*, 32(6), 461-463, doi: 10.3109/0142159X.2010.486059

Acautt, S., Carreño, D. y Rey, L. (2021). Análisis de Desempeño de Competencias en Pruebas Saber Pro y Saber 11 de los Estudiantes de Enfermería de 10 Semestre de la Universidad de Santander Campus Bucaramanga durante el año 2020- B. (Trabajo de grado) https://repositorio.udes.edu.co/handle/001/5963.

Adam, J., Bore, M., Childs, R., Dunn, J., Mckendree, J., Munro, D., & Powis, D. (2015). Predictors of professional behavior and academic outcomes in a UK medical school: A longitudinal cohort study. *Medical teacher*, *37*(9), 868-880. doi:10.3109/0142159X.2015.1009023

Al Alwan, I., Al-Moamary, M., Al-Attas, N., Al Kushi, A., Al Banyan, E., Zamakhshary, M., Al Kadri, H., Tamim, H., Magzoub, M., Hajeer, A., & Schmidt, H. (2011). The progress test as a diagnostic tool for a new PBL curriculum. *Educ Health*, 24(3), 493-531. https://educationforhealth.net//text.asp?2011/24/3/493/101426

Albanese, M., & Case, S. M. (2016). Progress testing: critical analysis and suggested practices. Advances in health sciences education: theory and practice, 21(1), 221–234. https://doi.org/10.1007/s10459-015-9587-z

Ali, K., Coombes, L., Kay, E., Tredwin, C., Jones, G., Ricketts, C., & Bennett, J. (2015). Progress testing in undergraduate dental education: The Peninsula experience and future opportunities. *European Journal of Dental Education.* 20(3). doi.org/10.1111/eje.12149

American Educational Research Association; American Psychological Association; National Council on Measurement in Education. (2014). *Standards for educational and*

pysochological testing. Washington D. C.: American Educational Research Association.

Anders, P. L., Stellrecht, E. M., Davis, E. L., & McCall, W. D., Jr (2019). A Systematic Review of Critical Thinking Instruments for Use in Dental Education. *Journal of Dental Education*, 83(4), 381–397. https://doi.org/10.21815/JDE.019.043

Agencia Nacional de Evaluación de la Calidad y Acreditación (ANECA) de España. (2013). Guía de apoyo para la redacción, puesta en práctica y evaluación de los resultados del aprendizaje.

Aparicio, J. y Valencia, J. (2020). ¿Qué es el "valor agregado" en la educación universitaria? *Revista Intellecta*: https://www.uninorte.edu.co/web/intellecta/que-es-el-valor-agregado-en-la-educacion-universitaria.

Asociación Colombiana de Facultades de Medicina. (2017). *ASCOFAME*. https://ascofame.org.co/web/consenso_monteria/#1509566110678-c74119ac-432a

Ayer, W., & Boulet, J. (2001). Establishing the validity of test score inferences: performance of 4th-year U.S. medical students on the ECFMG Clinical Skills Assessment. *Teaching and learning in medicine, 13*(4), 214-220. doi:10.1207/S15328015TLM1304_01

Bahamón, M. y Reyes, L. (2014). Caracterización de la capacidad intelectual, factores sociodemográficos y académicos de estudiantes con alto y bajo desempeño en los exámenes Saber Pro - año 2012. *Avances en Psicología Latinoamericana.* 32(3), 459-476.

Beason, A., Zuberi, R., Klamen, D., Hallam, J., Yousuf, N., Neumeister, E.. & Ward, J. (2019). The journeys of three ASPIRE winning medical schools toward excellence in student assessment. *Med Teach*, 457-46.

Boshuizen, H. P., Van der Vleuten, C. P., Schmidt, H. G., & Machiels-Bongaerts, M. (1997). Measuring knowledge and clinical reasoning skills in a problem-based curriculum. *Medical Education*, 31(2), 115-121

Blake, J. M., Norman, G. R., Keane, D. R., Mueller, C. B., Cunnington, J., & Didyk, N. (1996). Introducing Progress Testing in McMaster University´s Problem-Based Medical Curriculum: Psycometric properties and effect on Learning. *Academic Medicine*, 71(9), 1002-1007.

Bustamante, N. (21 de junio de 2013). Ministerio de Educación Nacional. https://www.mineducacion.gov.co/normatividad/1753/articles-324964_archivo_pdf_conceptos_Saber_Pro.pdf

Castro, M., Ruiz, J., y Guzmán, F. (2018). Cruce de las pruebas nacionales Saber 11 y Saber Pro en Antioquia, Colombia: una aproximación desde la regresión geográficamente ponderada (GWR). *Revista Colombiana de Educación*, (74), 63-79.

Castro, M. y Ruiz, J. (2019). La educación secundaria y superior en Colombia, vista desde las pruebas Saber. *Praxis & Saber*. 10(24), 341-366. doi.org/10.19053/22160159

Changiz, T., Yamani, N., Tofighi, S., Zoubin, F., & Eghbali, B. (2019). Curriculum Management/monitoring in undergraduate medical education: a systematized review. *BCM Med Educ*, 60.

Coates, H. (2008). Establishing the criterion validity of the Graduate Medical School Admissions Test (GAMSAT). *Medical Education, 42*(10), 999-1006.

Cohen, J. (1992). A power primer. *Psychological Bulletin*, 112(1), 155–159. https://doi.org/10.1037/0033-2909.112.1.155

Consejo Colombiano de Acreditación y Recertificación Médica, de Especialistas y Profesiones Afines (CAMEC). 2021. *Recertificación*. *http://camec.co/wp/recertificacion/*

Consejo Nacional de Acreditación – CNA. (2013). Lineamientos para la acreditación de programas de pregrado. Bogotá, Colombia: Consejo Nacional de Acreditación. https://www.mineducacion.gov.co/1621/articles-342684_recurso_1.pdf

Curtis, S., & Smith, D. (2020). A comparison of undergraduate outcomes for students from gateway courses and standard entry medicine courses. *BCM medical education*, 1-14.

Dabaliz, A., Kaadan, S., Dabbagh, M., Barakat, A., Shareef, M., Al-Tannir, M., & Mohamed, A. (2017). Predictive validity of pre-admission assessments on medical student performance. *International journal of medical education, 8*, 408-413. doi:10.5116/ijme.5a10.04e1

Delgado, M. (2011). ¿Será posible la formación ética y profesional de médicos y especialistas en el sistema de salud actual? *Revista Colombiana de Anestesiología*. 39:15-9.

Dixon, D. (2012). Prediction of Osteopathic Medical School Performance on the Basis of MCAT Score, GPA, Sex, Undergraduate Major, and Undergraduate Institution *Journal of Osteopathic Medicine*, 175-181.

Dobronski, L. (2007). Pruebas de progreso en escuelas de Medicina. La experiencia de siete años en la Universidad San Francisco de Quito con el Quarterly Profile Examination (Tesis de maestría. Colegio de Ciencias de la Salud de la Universidad San Francisco de Quito).

Domingue, B. (2012). Measuring effects of Colombian postsecondary institutions on Student learning. Documento presentado en el Seminario Internacional de Investigación sobre Calidad de la Educación, organizado por el ICFES.

Downing, S. (2003). Validity: on the meaningful interpretation of assessment data. *Medical Education.* 37(9). 830-841. https://doi.org/10.1046/j.1365-2923.2003.01594.x

Edel, R. (2003). El rendimiento académico: concepto, investigación y desarrollo. *Reice. Revista Iberoamericana sobre Calidad, Eficacia y Cambio en Educación, 1*(2). https://revistas.uam.es/reice/article/view/5354

Edwards, D., Friedman, T., & Pearce, J. (2013). Same admissions tools, different outcomes: a critical perspective on predictive validity in three undergraduate medical schools. *BCM Medical Education, 13*.

Epstein, R. M. (2007). Assessment in medical education. *The New England journal of medicine*, 356(4), 387–396.

Ferrer, J. y Arregui, P. (2003). Las pruebas internacionales de aprendizaje en América Latina y su impacto en la calidad de la educación: Criterios para guiar futuras aplicaciones. Programa de Promoción de la Reforma Educativa en América Latina y el Caribe. Grupo de Trabajo sobre Estándares y Evaluación. http://www.grade.org.pe/upload/publicaciones/archivo/download/pubs/Arreguiyferrer espaolpruebasinternacionales.pdf

Ferris, H., & O'Flynn, D. (2015). Assessment in medical education; what are we trying to achieve? *International Journal of Higher Education*, 139-144.

Findyartini, A., Werdhani, R. A., Iryani, D., Rini, E. A., Kusumawati, R., Poncorini, E., & Primaningtyas, W. (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(4), 366–373. doi:10.3109/0142159X

Frank, J., Taber, S., & Van Zanten, M. (2020). El papel de la acreditación en la educación de las profesiones de la salud del silgo XXI. *BMC Med Educ*.

Freeman, A., Van Der Vleuten, C., Nouns, Z., & Ricketts, C. (2010). Progress testing internationally. *Medical teacher*, 32(6), 451–455. https://doi.org/10.3109/0142159X.2010.485231

Freeman, A. & Ricketts, C. (2010) Choosing and designing knowledge assessments: Experience at a new medical school, *Medical Teacher*, 32(7), 578-581, doi: 10.3109/01421591003614858

Gabalan-Coell, J., y Vasquez-Rizo, F., (2016). Saber 11 y rendimiento universitario: un análisis del progreso en el plan de estudios. *Ciencia, Docencia y Tecnología*, 27(53), 135-161.

Geisinger, K., Bracken, B., Carlson, J., Hansen, J., Kunsel, N., Reise, S., & Rodríguez, M. (2013). APA Handbook of Testing and Assessment in Psychology (1): *American Psychological Asociation.*

Gil, F., Rodríguez, V., Sepúlveda, L., Rondón. M. y Gómez-Restrepo, C. (2013). Impacto de las facultades de medicina y de los estudiantes sobre los resultados en la prueba nacional de calidad de la educación superior (Saber Pro). *Revista Colombiana de Anestesiología*; 41:196-204.

Goldhaber, D., & Özek, U. (2019). How Much Should We Rely on Student Test Achievment as a Measure of Success. *Educational Researcher*, 479-483.

Greatrix, R., Nicholson, S., & Anderson, S. (2021). Does the UKCAT predict performance in medical and dental school? A systematic review. *BJM Open*, 1-13.

Griffin, B., Yeomans, N. D., & Wilson, I. G. (2013). Students coached for an admission test perform less well throughout a medical course. *Internal medicine journal*, *43*(8), 927–932. https://doi.org/10.1111/imj.12171

Horn, J. L. (1965). A rationale and test for the number of factors in factor analysis. *Psychometrik.* 87(3), 179–185 https://doi.org/10.1007/BF02289447

ICFES. (2018). Promoción de la salud y prevención de la enfermedad. Colombia.

ICFES. (2020a). Informe Nacional de Resultados del Examen Saber 11 -2019. https://www.icfes.gov.co/documents/20143/1711757/Informe%20nacional%20de%20resultados%20Saber%2011-2019

ICFES. (2020b). Informe Nacional de Resultados del Examen PRO- 2019. https://www.icfes.gov.co/documents/20143/1711776/Informe%20nacional%20de%20resultados%20Saber%20Pro%202016-2019

Jiraporncharoen, W., Angkurawaranon, C., Chockjamsai, M., Deesomchok, A., & Euathrongchit, J. (2015). Learning styles and academic achievement among undergraduate medical students in Thailand. *Journal of educational evaluation for health professions, 12*(38). doi:doi.org/10.3352/jeehp.2015.12.38

Johnson, T. R., Khalil, M. K., Peppler, R. D., Davey, D. D., & Kibble, J. D. (2014). Use of the NBME Comprehensive Basic Science Examination as a progress test in the preclerkship curriculum of a new medical school. *Advances in physiology education*, *38*(4), 315–320. https://doi.org/10.1152/advan.00047.2014

Jornet, J. (2017). Evaluación estandarizada. *Revista Iberoamericana de Evaluación Educativa*, 5-8.

Julian, E. (2005). Validez de la prueba de admisión a la facultad de medicina para predecir el rendimiento de la escuela de medicina. *Academic Medicine, 80*(10), 910-917.

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

Kane, M. (2006). Validation. *Educational measurement, 4*(2), 17-64.

Keltner, C., Haedinger, L., Carney, P., & Bonura, E. (2021). Preclinical assessment performance as a predictor of USMLE step 1 scores or passing status. *Medical Science Educator*, 1453-1462.

Kennet-Cohen, T., Turvall, E., Saar, Y., & Oren, C. (2016). La validez predictiva de un proceso de selección de dos pasos para las escuelas de medicina. *Revista de Educación Biomédica*. doi:10.1155/2016/8910471

Kerfoot, B. Price M., Shaffer, K., McMahon, Graham T., Baker, H., Kirdar, J., Kanter, S., Corbett, E., Berkow, R., Krupat, E., & Armstrong, E. (2011). Online "Spaced Education Progress-Testing" of Students to Confront Two Upcoming Challenges to Medical Schools. *Academic Medicine*: 86(3), 300-306, doi: 10.1097/ACM.0b013e3182087bef

Krupat, E., Pelletier, S., & Dienstag, J. (2017). Academic Performance on First-Year Medical School Exams: How Well Does It Predict Later Performance on Knowledge-Based and Clinical Assessments? *Teaching and learning in medicine, 29*(2), 181-187. doi:10.1080/10401334.2016.1259109

Ley 1324 de 2009. (13 de julio de 2009). Ministerio de Educación superior. Diario oficial No. 47.409.

Lievens, F., & Sackett, P. (2006). Video-based versus written situational judgment tests: a comparison in terms of predictive validity. *Journal of Applied Psychology, 91*(5), 1181-1188.

MacKenzie, R., Cleland, J., Ayansina, D., & Nicholson, S. (2016). Does the UKCAT predict performance on exit from medical school? A national cohort study. *BMJ open*, 1-10.

Martínez, E., Gómez, M. y Visbal, L. (2015). Resultados de la Prueba de Ciclo Básico en el Programa de Medicina de la Universidad del Norte: análisis retrospectivo y retos futuros. *Revista Salud Uninorte, 31*(1), 53-58.

McManus, I., Woolf, K., Dacre, J., Paice, E., & Dewberry, C. (2013). The Academic Backbone: Longitudinal Continuities in Educational Achievement from Secondary

School and Medical School to MRCP (UK) and the specialist register in UK medical students and doctors. *BMC Med*, 1-27.

McManus, I., Dewberry, C., Nicholson, S., Dowell, J., Woolf, K., & Potts, H. (2013). Construct-level predictive validity of educational attainment and intellectual aptitude test in medical student selection: meta-regression of six UK longitudinal studies. *BMC Med, 11*. doi:10.1186/1741-7015-11-243

Melo, L., Ramos, J. y Hernández, P. (2014). La Educación Superior en Colombia: Situación Actual y Análisis de Eficiencia. Borradores de Economía, número 808. https://www.banrep.gov.co/sites/default/files/publicaciones/archivos/be_808.pdf.

Muijtjens, A.M.M., Schuwirth, L.W.T., Cohen-Schotanus, J. & Van Der Vleuten, C.P.M. (2007), Origin bias of test items compromises the validity and fairness of curriculum comparisons. *Medical Education*, 41(12) 1217-1223. https://doi.org/10.1111/j.1365-2923.2007.02934.x

Nimkuntod, P., & Tongdee, P. (2016). Preclinical Medical Students Achievement to Learning Outcomes in Special Tracts of Rural Doctors. *Journal of the Medical Association of Thailand* 105-110.

Nouns, Z., & Georg, W. (2010). Progress testing in German speaking countries, *Medical Teacher*, 32:6, 467 470, doi: 10.3109/0142159X.2010.485656

Organización de las Naciones Unidas para la Educación, la Ciencia y la Cultura (UNESCO). 2002. Informe sobre la educación superior en Colombia 2002, p. 3-149.

Patterson, J., Thaler, T., Hoffmann, M., Hughes, S., Oels, A., Chu, E., Mert, A., Huitema, D., Burch, S. & Jordan, A. (2018). Political feasibility of 1.5°C societal transformations: the role of social justice, Current Opinion in Environmental Sustainability, *BMC Medicine*. 31, 1-9. doi.org/10.1016/j.cosust.2017.11.002.

Pinilla, A., & Parra, M. (2009). Estrategias de evaluación para fortalecer el aprendizaje. En L. López, & M. Sáenz, *Reflexiones sobre educación universitaria* (pp. 233-250). Editorial Universidad Nacional de Colombia.

Pontificia Universidad Javeriana. (s. f.). Plan de estudios de la carrera de Medicina. https://www.javeriana.edu.co/documents/12362/12084568/Plan+de+estudios+Carrera+de+Medicina+Septiembre+2021/1e24fe9f-8812-4acc-b952-04c5a3c8daf9

Plessas, A., (2015). Validity of Progress Testing in Healthcare Education. International Journal of Humanities Social Sciences and Education, 2(8), 22-33.

Rademakers, J., Ten Cate, T. J., & Bär, P. R. (2005). Progress testing with short answer questions. *Medical teacher*, 27(7), 578–582. https://doi.org/10.1080/01421590500062749

Ramírez, C. (2014). Factores asociados al desempeño académico según nivel de formación pregrado y género de los estudiantes de educación superior Colombia. *Revista Colombiana de Educación*, (66), 203-224. http://www.scielo.org.co/scielo.php?script=sci_arttext&pid=S0120-39162014000100009&lng=en&tlng=es.

Ricketts, C., Freeman, A., Pagliuca, G., Coombes, L., & Archer, J. (2010). Difficult decisions for progress testing: how much and how often? Medical teacher, 32(6), 513–515. https://doi.org/10.3109/0142159X.2010.485651

Rodríguez, A. y Ruíz, M. (2011*). Indicadores de rendimiento de estudiantes universitarios: calificaciones VS créditos*. *Revista de Educación*, 467-492.

Ruíz, G., Ruíz, J. y Ruíz, E. (2010). Indicador global de rendimiento. *Revista Iberoamericana de Eduación*, 1-11.

Sánchez, J. (2014). La formación médica en Colombia. *Revista Educación y Desarrollo Social*, 8 (2), 168-183.

Sánchez-Bello, N., Galván-Villamarín, J. y Eslava-Schmalbach, J. (2016). Producción científica en las facultades de Medicina en Colombia en el periodo 2001-2015. *Revista de la Facultad de Medicina.* 64. 645. 10.15446.

Sánchez-Mendiola, M., Moreno-Salinas, J., Bautista-Godínez, T. y Martínez-González, A. (2019). *Gaceta médica de México*, 90-100. doi:10.24875/GMM.18004801

Schreurs, S., Cleutjens, K., Cleland, J., & Oude, M. (2020). Outcomes based selection into medical school: predicting excellence in multiple competencies during the clinical years. *Journal of the Association of American Medical Colleges*, 1411-1420.

Shadish, W., Cook, T., & Campbell, D. (2002). Experimental and quasiexperimental designs for generalized causal inference. *Houghton Mifflin Company*. https://www.alnap.org/system/files/content/resource/files/main/147.pdf

Schuwirth, L., Colliver, J., Gruppen, L., Kreiter, C., Mennin, S., Onishi, H., Pangaro, L., Ringsted, Ch., Swanson, D., Van Der Vleuten, C., Wagner-Menghin, M., (2011). Research in assessment: Consensus statement and recommendations from the Ottawa 2010 (Conference), *Medical Teacher*, 33:3, 224-233.

Sireci, S. & Sukin, T. (2013). Test Validity. En *APA Handbook of Testing and Assessment in Psychology.* Washington: American Psychological Association.

Swanson, D. B., Holtzman, K. Z., & Butler, A. (2010). Collaboration across the pond: the multi-school progress testing project. *Medical Teacher*, 32(6), 480-485. doi: 10.3109/0142159x.2010.485655

Taber, S., Akdemir, N., Gorman, L., Van Zanten, M., & Frank, J. (2020). A fit for purpose framework for medical education accreditation system design. *BMC medical education*. doi:10.1186/s12909-020-02122-4

Tamblyn, R., Abrahamowicz, M., & Dauphinee, W. (2002). Asociación entre los puntajes del examen de licenciatura y la práctica en atención primaria. *JAMA, 288*(23), 3019-3026. doi:10.1001/jama.288.23.3019

Times Higher Education (2022, 12 de marzo). Times World University Rankings 2019. https://www.timeshighereducation.com/world-university-rankings/2019/world-ranking

Thorndike, R., Cunningham, G., Thorndike, R. L., & Hagen, E. (1991*).* Measurement and evaluation in psychology and educatio*n.* Macmillan Publishing Co, Inc.

Tomatis, M., Burrone, M. y Romero, D. (2016). Validez predictiva del examen de ingreso a la carrera de medicina de la Facultad de Ciencias Médicas (UNC). *Revista de educación*, 7(9), 358-367.

Torre, D., Dong, T., Schreiber-Gregory, D., Durning, S., Pangano, L., Pock, A., & Hemmer, P. (2020). Exploring the predictors of post-clerkship USMLE step 1 scores. *Teaching and Learning in Medicine*, 330-336.

Thurstone, L.L. (1935). The Vectors of Mind. *Psychological Review, 41*(1), 1–32. https://doi.org/10.1037/h0075959

Universidad de los Andes. (s. f.). Plan de estudios del programa de Medicina. https://medicina.uniandes.edu.co/es/programas/pregrado/plan-de-estudios

Valencia, J., Aparicio, J. y Villegas-Mendoza, A. (2020). ¿Qué tanto valor agregado aporta a sus estudiantes las universidades colombianas con los mejores resultados en las pruebas SABER PRO 2016-2018? Serie Documentos No. 41. https://www.researchgate.net/publication/353559638_Que_tanto_valor_agregado_aportan_a_sus_estudiantes_las_universidades_colombianas_con_los_mejores_resultados_en_las_pruebas_SABER_PRO_2016-2018

Van der Vleuten, C. P., Schuwirth, L. W., Driessen, E. W., Dijkstra, J., Tigelaar, D., Baartman, L. K., & Van Tartwijk, J. (2012). A model for programmatic assessment fit for purpose. *Medical teacher*, 34(3), 205–214. https://doi.org/10.3109/0142159X.2012.652239

Van der Vleuten, C. P. (1996). The assessment of professional competence: developments, research and practical implications. *Advances in health Sciences Education*, 1 (1), 41-67.

Verhoeven, B. H., Snellen-Balendong, H. A., & Hay, I. T. (2005). The versatility of progress testing assessed in an international context: a start for benchmarking global standardization? *Medical Teacher*, 27(6), 514-520. doi: 10.1080/01421590500136238

Vergel, J. (29 de mayo de 2019). Medicina Rosarista, un momento histórico. *NOVA et VETERA.* https://www.urosario.edu.co/Periodico-NovaEtVetera/Nuestra-U/Medicina-rosarista-un-momento-historico/

Wang, L., Laird-Fick, H., Parker, C., & Solomon, D. (2021). Using Markov chain model to evaluate medical student's trajectory on progress test and predict USMLE step 1 scores - a retrospective cohort study in one medical school. *BMC Med Educ*.

Wiley, J. (2018). The Association for the Study of Medical Education. *Medical Education, 52*, 641-653.

Wilkinson, D., Zhang, J., & Parker, M., (2011). Validez predictiva de la Prueba de Admisión de Licenciatura en Medicina y Ciencias de la Salud para el rendimiento académico de estudiantes de medicina. *Revista médica de Australia,* 194 (7), 341-344

Wrigley, W., Van der Vleuten, C. P., Freeman, A., & Muijtjens, A. (2012). A systemic framework for the progress test: strengths, constraints and issues: AMEE Guide No. 71. *Medical teacher*, *34*(9), 683–697. https://doi.org/10.3109/0142159X.2012.704437

Wur, T. (s. f.). *The World University Rankings*.

https://www.timeshighereducation.com/world-university-rankings/2021