UNIVERSIDAD DE COSTA RICA
SISTEMA DE ESTUDIOS DE POSGRADO

ESTABLECIMIENTO DE PUNTOS DE CORTE PARA PRUEBAS ESTANDARIZADAS REFERIDAS A CRITERIOS: EL CASO DE LA PRUEBA DE HABILIDADES CUANTITATIVAS

Tesis sometida a la consideración de la Comisión del Programa de Estudios de Posgrado en Educación para optar al grado y título de Maestría Académica en Educación con énfasis en Evaluación Educativa

ROBERTO GUZMÁN GÓMEZ

Ciudad Universitaria Rodrigo Facio, Costa Rica

2025

Dedicatoria

Con especial cariño a mi hija Abigail, mi mayor motivación. A mi familia, por su apoyo incondicional. Y a mis compañeras y compañeros de trabajo, por su apoyo y confianza en esta etapa.

Agradecimientos

Quisiera agradecer al Dr. Luis Rojas Torres, mi tutor durante este proyecto. Su orientación, su paciencia y sus valiosos comentarios fueron fundamentales para el desarrollo y la culminación de este proyecto. Su dedicación y conocimiento han sido una fuente constante de inspiración para mi persona.

Agradezco al equipo del Departamento de Docencia Universitaria por su continuo apoyo y el voto de confianza depositado en mí durante este proceso formativo.

Un agradecimiento especial al equipo de jueces y juezas, quienes con su generosa colaboración proporcionaron parte de la información central que permitió el desarrollo de este proyecto.

Extiendo mi gratitud al equipo de la Prueba de Habilidades Cuantitativas por su valioso apoyo y colaboración. Gracias por abrirme sus puertas.

Finalmente, a mi familia, mi más profundo agradecimiento por su amor incondicional, su paciencia infinita y su apoyo constante a lo largo de toda esta etapa. Su aliento fue mi mayor motivación para alcanzar esta meta.

"Esta tesis fue aceptada por la Comisión del Programa de Estudios de Posgrado en Educación de la Universidad de Costa Rica, como requisito parcial para optar al grado y título de Maestría Académica en Educación con énfasis en Evaluación Educativa."

Dra. Grettel Arias Orozco
Representante de la Decana
Sistema de Estudios de Posgrado

Dr. Luis Rojas Torres
Director de Tesis

Dra. Graciela Ordóñez Gutiérrez
Asesora

M.Sc. Mónica Mora Badilla
Asesora

Dra. Patricia Marín Sánchez
Directora del Programa de Posgrado en Educación

Roberto Guzmán Gómez
Candidato

Tabla de contenido

Resumen

Esta investigación aborda el establecimiento de puntos de corte (también conocido como cut score) en pruebas estandarizadas, específicamente en la Prueba de Habilidades Cuantitativas de la Universidad de Costa Rica, la cual se utiliza como parte del proceso de admisión para algunas carreras en esta universidad que se caracterizan por tener una carga significativa de cursos de matemática en sus planes de estudio.

Mediante la implementación de los métodos bookmark (basado en la identificación de un ítem clave) y Angoff (basado en la estimación de la probabilidad de respuesta correcta por expertos), el estudio determinó la puntuación mínima que una persona debe alcanzar para demostrar el nivel de habilidad requerido. Como resultado, se obtuvo un punto de corte de 67/100, el cual identifica a estudiantes con un perfil que demuestra habilidades fundamentales para el éxito académico, tales como relacionar, validar, clasificar y generalizar. Los análisis estadísticos realizados en esta investigación verificaron la existencia de una relación significativa entre el desempeño en la prueba y el éxito en el curso universitario Cálculo I. Este hallazgo valida que el punto de corte calculado no es arbitrario, sino que posee valor predictivo y es confiable como criterio de selección.

En el contexto educativo, este tipo de investigaciones son fundamentales para la toma de decisiones informadas y justas sobre el acceso, la permanencia y el acompañamiento académico del estudiantado, además de servir como insumo para el diseño de estrategias pedagógicas más personalizadas y basadas en evidencia.

Índice de tablas

1. Introducción

La investigación *Establecimiento de puntos de corte para pruebas estandarizadas referidas a criterios, el caso de la prueba de Habilidades Cuantitativas* tuvo como objetivo colaborar con el proceso de fijar puntos de corte para la Prueba de Habilidades Cuantitativas (PHC) que aplican algunas carreras de la Universidad de Costa Rica como requisito de ingreso para las personas aspirantes. La definición de las notas de corte se realizó por medio de la aplicación de los métodos Bookmark y Angoff, ambos con un historial significativo en su aplicación para proyectos como este.

El primero de esos métodos parte de la organización de los ítems de una prueba de menor a mayor dificultad, posteriormente un grupo de jueces analiza individualmente cuál de esos ítems marca una división entre las capacidades de las personas que aplican la prueba, se le conoce como el ítem bookmark por su función de “separador” en la prueba, luego, se comparan los juicios de cada persona incluida en el panel de jueces y se promedia el resultado individual para establecer el punto de corte respectivo.

Por su parte, con el método Angoff se considera a cada ítem de manera aislada y los jueces fijan la probabilidad de que 2 de cada 3 personas acierten ese ítem, una vez que se tienen los resultados para cada ítem de la prueba, se promedian los valores y así se fija el punto de corte por medio de ese método.

La relación de esta propuesta de investigación con el tema de estudio de la maestría se centra en el aporte de los resultados al análisis e implementación de dos métodos para establecer puntos de corte en pruebas educativas estandarizadas. El proceso que se realizó como parte de la investigación, así como sus resultados inciden

desde lo teórico práctico en el campo de la evaluación a nivel nacional, esto en tanto algunas de las pruebas que se realizan en el país carecen de la fijación objetiva de puntos de corte, así como de evidencia empírica que justifique o permita describir las habilidades de las personas que realizan dichas pruebas.

La investigación se realizó con las bases de datos de la población de personas estudiantes que desarrolló la prueba de Habilidades Cuantitativas en el año 2021, pues de acuerdo con los planes de estudio de las carreras, se supone que esa población ya habrá estudiado y aprobado los cursos de Cálculo que son de interés para esta investigación.

Desde el punto de vista social y educativo, los resultados de la investigación aportan en la caracterización de las personas que se postulan a ingresar a la Universidad de Costa Rica en las carreras que consideran los cursos de Cálculo en sus planes de estudio. A la vez, los resultados inciden en otros procesos para el establecimineto de puntos de corte en el campo educativo nacional e internacional, principalmente al evaluar las competencias de las personas que pretenden ingresar a la educación universitaria.

1.1. Antecedentes

Al plantear la propuesta de investigación se procedió a realizar una primera exploración general de la temática a nivel de la literatura y las bases de datos que se relacionaban con los principales términos a considerar en este proyecto, mismos que se utilizaron como guías para la búsqueda de las fuentes de información. Sin embargo, varios de esos referentes se desecharon debido a que no resultaban significativos para el trabajo que se proponía, por ejemplo por las temáticas abordadas o el enfoque de la investigación que se desarrolló. También otro criterio para descartarlas fue la fecha de

publicación respectiva, pues en algunos casos eran fuentes antiguas para las que se encontraban otras fuentes similares y más recientes.

A continuación, se presenta un recorrido por algunos de esos referentes, los mismos se han organizado según las temáticas que abordan.

Una primera categoría que se rescata de la primera búsqueda de información, es la que corresponde a los elementos cercanos al razonamiento matemático, algunas de las publicaciones encontradas, versan sobre el razonamiento lógico matemático, la habilidad matemática y otros elementos que se considera que inciden en la resolución de problemas o el enfrentar desafíos con el uso de contenidos matemáticos. Por ejemplo, se destaca el estudio de Calua (2016), estudio que buscó medir la potencia predictiva de algunas variables respecto al éxito académico de las personas estudiantes. De manera similar, Hernández (2017) en su tesis de maestría, desarrolló un sistema de evaluación para medir el razonamiento matemático, para lograr su objetivo se apoyó en pruebas adaptativas por medio de computadores.

Por su parte, en el caso de Uruguay, Rodríguez (2017), ofrece una sistematización del proceso de construcción y los alcances de una prueba que se preparó para diagnosticar las competencias en lectura y matemática previo al ingreso a la universidad.

Otra de las categorías significativas en la revisión de la literatura, es la que refiere al uso de los puntos de corte y de las pruebas estandarizadas en el área médica. Destaca el importante número de publicaciones que tratan esos temas, además de que en su mayoría son de habla inglesa, lo cual es congruente con la línea de desarrollo de la evaluación en América del Norte, la cual persigue la estandarización. Ejemplo de estos trabajos es el de Ruano et al. (2018) y el de Anderson (2020).

Se encontró otro grupo de publicaciones que hace referencia a la medición de habilidades varias, estas no se consideraron debido a que sus alcances diferían de los de pruebas como la PHC. En ese listado de pruebas se puede mencionar los estudios de Guevara (2017), Minervino y Dias (2017), Illesca y Alfaro (2017), el de Pérez y Valmaseda (2019) y el del Instituto Colombiano para la Evaluación de la Educación (2020).

Considerando que la PHC se encuentra en la categoría de las pruebas estandarizadas, se descartó la categoría de publicaciones, que refiere a las limitaciones o críticas de las pruebas estandarizadas. Entre esos trabajos se encuentra el escrito de Popham (2014) quien además de criticar de manera constructiva las pruebas, brinda orientaciones para rectificar algunos de los vicios que existen en su construcción.

Otra vertiente que se encontró fue la de las publicaciones, que ofrecen distintas alternativas de las pruebas estandarizadas para la medición de habilidades, entre ellas destaca el trabajo de Gallardo et al. (2015), en esa publicación se señalan los alcances y las posibilidades del uso del portafolio como una evidencia del logro de habilidades y destacan que al ser un proceso más cualitativo, subsana la concentración de la atención en la medición de los saberes y habilidades en un solo momento, como sucede con las pruebas estandarizadas.

Los antecedentes que se han mencionado hasta este punto permiten establecer un marco general sobre lo abordado en otras investigaciones, sin embargo, a continuación se presenta un apartado que busca detallar otras investigaciones cuya especificidad les hace aún más relevantes y cercanas para esta investigación.

1.2. Antecedentes relacionados con la investigación

A continuación se presenta a la persona lectora una breve descripción de una serie de investigaciones que son cercanas a esta propuesta, en el apartado se considera una exposición de los antecedentes que se consideraron más relevantes, se ha dividido según las temáticas de interés, considerando desde los referentes más antiguos a los más recientes y organizados desde lo internacional, lo regional hasta lo nacional. Para recopilar la información, se realizó una revisión en bases de datos como EBSCOhost, JStor, ProQuest, ScienceDirect y ELibro, así como en repositorios como Kérwá y Kímuk, además de buscadores académicos en Internet, entre ellos Google Scholar y Scopus. Las fuentes obtenidas se filtraron por año de publicación (de 2014 en adelante) y por áreas de interés, por ejemplo consultando el área de ciencias sociales, ciencias básicas y matemática. Para la recolección de datos, se emplearon los siguientes términos clave: pruebas estandarizadas, puntos de corte, método bookmark, método Angoff, razonamiento matemático, valoración de constructos, pruebas criteriales, validez y confiabilidad y prueba de habilidades cuantitativas.

Seguidamente se muestran los resultados obtenidos tras las consultas realizadas en las bases de datos.

1.2.1. Uso de pruebas estandarizadas

Para iniciar este recorrido, se presenta el estudio realizado por Šifrar y Trenc (2014) en Eslovenia. En esta investigación, se aborda el tema de la calibración de una prueba estandarizada, específicamente con la prueba Matura que busca establecer el nivel de conocimiento de la lengua española como segunda lengua en el marco europeo. Para desarrollar el análisis recurren a una muestra de 189 y 196 postulantes a la prueba

en los niveles básico e intermedio respectivamente y en un trabajo conjunto con los jueces para desarrollar el método Angoff lograron evidenciar que se debe revisar la nota de aprobación en el nivel más bajo, mientras que el criterio de los jueces coincide con los puntajes respectivos para el nivel intermedio de dominio del español. Además de las variaciones que sugieren con su indagación, también destacan los alcances del método Angoff para establecer los puntos de corte en pruebas estandarizadas y señalan los principales aspectos que la hacen una de las opciones más populares para tal fin.

En Baréin, Al-Musawi (2016), aborda un tema similar al descrito anteriormente, pero en este caso en un entorno universitario. En su estudio, considera 348 estudiantes de la Universidad de Bahrain para medir los logros o el progreso de esa muestra de estudiantes por medio de un análisis de la prueba aplicada utilizando la Teoría de Respuesta al Ítem (TRI). Como parte de los resultados obtenidos, logra mostrar que de los 31 ítems de la prueba aplicada, 17 no se ajustan a los requerimientos de la prueba, mientras que el resto muestran altos índices de confiabilidad (entre 0,87 y 0,93). Como principal conclusión del estudio, se señala la recomendación de aplicar la prueba analizada como una medida confiable para el logro de las personas estudiantes en el tema de evaluación educativa.

En el caso de Uruguay, Rodríguez (2017), analiza el proceso de creación y desarrollo de dos pruebas para evaluar competencias en el área de matemática y lectura para el ingreso a los Centros Universitarios Regionales de la Universidad de la República. Como parte del proceso se crean y establecen los estándares de las evaluaciones en colaboración con expertos. Además, en el test se aplican análisis basados en la Teoría

de Respuesta al Ítem (TRI) para determinar los niveles de desempeño, un enfoque similar al utilizado hasta ahora en el caso de PHC.

En cuanto al cuestionario, se preseleccionaron 186 ítems de los que se emplearon 88 en la confección de dos cuadernillos, esto por medio del pilotaje y de los análisis respectivos según la TCT (Teoría Clásica de los Test) y la TRI. En test de matemática aplicaron 1380 personas y en la de lenguaje 1244. Las pruebas alcanzan altos niveles de confiabilidad, con valores superiores a 0,80 en el Alfa de Cronbach, y permiten establecer notas de aprobación sin recurrir a métodos como Angoff, Bookmark o sus variantes.

Como resultados señalan los bajos niveles de suficiencia que se encontraron en las evaluaciones de matemática, lo cual es congruente con los resultados arrojados por otras test nacionales e internacionales.

Bajo una línea similar de trabajo, Durán (2019) en Chile, analiza las evaluaciones de competencias que desarrolla el Departamento de Evaluación, Medición y Registro Educacional de la Universidad de Chile como parte del proceso de admisión de sus estudiantes. Aborda el proceso de construcción de nuevos cuestionarios y realiza análisis de los mismos y los acompaña con entrevistas a los postulantes. Como muestra contaron con 6392 estudiantes que se postularon para ingresar en el 2017 y 2500 en el año 2018. Destacan como resultados del estudio que aunque la prueba es una versión actualizada de las preexistentes, cuenta con algunos niveles de sesgo que perjudican a la población de más bajo nivel, aspectos a los que se debe prestar atención en este tipo de test estandarizados y que van más allá de lo que puede arrojar el análisis de los mismos por medio de la TRI o de la TCT.

Por su parte, Demarchi (2020), desde el contexto de Colombia, señala algunos puntos de encuentro con Durán del Fierro y parte de los aportes más significativos de su investigación es el énfasis que realiza en las evidencias de sesgo que se tiene por nivel socioeconómico y género en algunas pruebas estandarizadas que se aplican a nivel de la región americana, entre ellas las PISA. Apela a la necesidad de contar con suficientes evidencias de validez y confiabilidad en este tipo de pruebas y de estar siempre vigilante a estos aspectos. En su metodología recurre a la revisión documental de un amplio rango de pruebas y resultados que se han obtenido en la región luego de la aplicación de estas pruebas estandarizadas.

Finalmente, a nivel nacional se destaca la investigación de Smith (2014) en Costa Rica, quien en el marco del Instituto de Investigaciones Psicológicas de la Universidad de Costa Rica, realiza una reflexión y revisión de una serie de instrumentos de medición que recolecta luego de la sistematización de información tras una revisión bibliográfica. Más allá de los instrumentos que propone la autora, lo más relevante de su estudio es la diferenciación que realiza entre los alcances y recomendaciones para emplear cada instrumento y a la vez reseña algunas experiencias, entre ellas la prueba de aptitud académica de la UCR y la PHC. Señala además que las alternativas propuestas para evaluar no son las únicas, y que lo que más interesa es prestar atención a la adecuación de cada instrumento en cada caso y a la población en la que se desea implementar.

Desde la revisión de estos insumos, se rescata el abordaje de las pruebas estandarizadas que incluyen algunas experiencias en el establecimiento de los puntos de corte por medio de la TRI y de la TCT, además el señalamiento de algunas limitaciones

en su implementación y en los alcances que pueden tener sus resultados en las diversas poblaciones en las que se aplican.

Para el caso de otras instituciones que emplean puntos de corte para asignar la aprobación o reprobación de una prueba, destaca el caso del Consejo de Educación Vial (CEV), dependencia del Ministerio de Obras Públicas y Transportes (MOPT), así como de la Dirección de Gestión y Evaluación de la Calidad del Ministerio de Educación Pública de Costa Rica. En cada caso emplean un determinado valor para establecer si se aprueba o no una evaluación.

En el caso del CEV, la señora Jackeline Ruíz (directora), en conjunto con los señores Gary Jiménez (encargado de la prueba práctica de manejo) y Alejandro Vargas (encargado de la prueba teórica de manejo), indican en una entrevista realizada el día 5 de mayo de 2023 por el tesiario a cargo de esta investigación, que las pruebas que cada uno aplica poseen una nota de aprobación que se asignó más por un criterio legal (artículos 217 y 221 de la ley de tránsito) que por uno basado en otro tipo de objetivos o evidencias. En el caso de la prueba teórica de manejo, con el decreto 138 del MOPT, se realizan esfuerzos para que su aprobación sea con una nota equiparada con la nota mínima del ciclo diversificado de la educación pública, es decir, pasó de 80 a 70.

Para el caso de las pruebas realizadas por el MEP y sus valores mínimos (70 para el ciclo diversificado y 65 para los ciclos I, II y III), se intentó coordinar encuentros con los encargados de dicha dirección, o al menos obtener sus opiniones por medio de contactos vía correo electrónico, pero a la fecha no se ha obtenido respuesta de los mismos.

1.2.2. Experiencias en el uso de métodos para establecer puntos de corte

Otro eje temático que se ha considerado para esta sección es el empleo de métodos para el establecimiento de puntos de corte y para su desarrollo se empezará con la investigación realizada por los autores Clauser et al. (2017) en Estados Unidos, quienes en el contexto de las evaluaciones para las ciencias médicas y sus estándares de aprobación, buscan recolectar evidencias de validez para las pruebas que se aplican en el ámbito médico para los futuros profesionales en el área. En el panel de expertos contaron con 30 jueces divididos en grupos de 10 miembros agrupados para asegurar un balance de las pruebas en temas de la especialidad médica, género y región geográfica. Concluyen que los puntos de corte que se establecen por cada uno de los jueces varían, y que por ello es importante acompañar el establecimiento de los cortes con otras técnicas que permitan validar los niveles que se consideran inicialmente y así asegurar una mayor validez en los criterios obtenidos.

Para el caso europeo, específicamente en Alemania, Kampa et al. (2019a) abordan el caso del método Angoff y establecen puntos de corte para una prueba de ciencias que se administró a 3641 estudiantes de secundaria. Contaron con un panel de nueve expertos para fijar esos puntajes. A los puntajes que asignaron lo jueces se les contrastó con otros análisis estadísticos como ANOVA, t-test y análisis de regresiones con el fin de contar con evidencias de validez del proceso, de las consecuencias y de la prueba a nivel interno y externo. Como conclusión del estudio, se identifican puntos en común con investigaciones previas, particularmente en que la aplicación de métodos como el de Angoff, por sí sola, no es suficiente. Es necesario que se complemente con otros análisis que refuercen la validez y confiabilidad de los resultados. Si bien el criterio de los jueces

constituye un buen punto de partida, siempre puede ser enriquecido con más datos. Además, se destaca que las interpretaciones de estos resultados deben realizarse con cautela, debido a posibles fallos o incertidumbres inherentes al proceso.

Más recientemente, en Estados Unidos, en la investigación de Baldwin et al. (2020), se agrega más información sobre las recomendaciones para aplicar los métodos de puntos de corte. En este caso su labor se centró en analizar específicamente la consistencia interna entre los valores que fijaron los jueces considerados para establecer puntos de corte en una prueba. Consideraron 75 ítems de selección única, lo cuales fueron analizados desde el punto de vista de los puntajes que asignaron los jueces cuando los analizaron para establecer puntos de corte por medio del método Bookmark. El estudio concluye aspectos similares a los ya indicados en las dos investigaciones precedentes, es decir, Clauser et al. (2017) y Kampa et al. (2019a) y a la vez indica que los puntajes o valores que se asignen a un ítem por parte de un juez, depende en gran medida de la información previa de la que disponga, por ello estos valores son variables y se enfatiza en lo indispensable que es la cantidad y calidad de la información previa de la que dispongan los jueces.

En una línea similar, en España, se desarrolló un proyecto de fin de maestría en el que se logró fijar un punto de corte con la implementación del método Bookmark para establecer niveles de desempeño en lo que refiere a algunas competencias digitales (Villalobos, 2018). De su estudio, la autora señala que se fijó un precedente a nivel de metodología para el futuro empleo de dicho método para fijar niveles de competencias digitales.

En el área de Ciencias de la Salud, se señala el caso de Ecuador, en el mismo se implementa el método Angoff para fijar el punto de corte para la habilitación de profesionales en el área (Ruano et al., 2018). Este caso no es aislado en la región, pues también se reconoce el caso de México, donde los métodos de punto de corte fueron una herramienta para dotar de mayor validez y propósito a las pruebas de certificación de médicos, específicamente en el caso de otorrinolaringólogos.

De lo expuesto hasta este punto, destacan diversos estudios realizados en distintos momentos y lugares que señalan las limitaciones de los métodos más populares para establecer puntos de corte. Sin embargo, esto no implica que deban ser descartados. Uno de los aspectos más importantes que se rescata de estas publicaciones es la necesidad de contar con otros recursos para comparar los puntajes obtenidos por los jueces, esto prestando atención a un interés de ofrecer resultados confiables y válidos. Entre las limitaciones que se señalan para la implementación de los métodos, se indica la falta de capacitación o de conocimiento de los aspectos relevantes por parte de las personas que conforman el jurado, por ello señalan que este proceso se debe potenciar a fin de obtener mejores resultados en los juicios que emiten para cada método. También, como aspecto favorable de la implementación de los métodos, se destaca que se fija un valor de aprobación objetivo que cuenta con un sentido e interpretaciones claras a partir de la experiencia del equipo de jueces.

Estos son elementos fundamentales para considerarlos al momento de generar propuestas de investigación en temas cercanos a los abordados en estas publicaciones.

1.2.3. Antecedentes sobre el razonamiento matemático

Otra de las aristas que se ha considerado para este estudio es el constructo del razonamiento matemático y afines, esto en tanto esa habilidad es la que se busca medir en la PHC, es por ello que ahora se realizará un recorrido por algunas investigaciones sobre el tema.

En primer lugar se tiene a Reguant y Martínez (2014), quienes desde España presentan algunas de las limitaciones y alcances del proceso de operacionalización de algunas variables en el contexto educativo, esto desde una revisión de la literatura. Destacan que algunos de estos constructos son multidimensionales y por ello su estudio es retador, además de que cuentan con una gran cantidad de variables e indicadores. Proponen como ejemplo el caso del éxito escolar, el cual va más allá de la mera obtención de notas altas.

Karsenty (2014) en Holanda, recurre a la revisión documental para llevar a cabo una deconstrucción del concepto de razonamiento matemático en la que considera una arista cognitiva y otra práctica para su análisis. Indica que una de las particularidades de ese constructo es que su estudio debe ser multidisciplinar y destaca que en los estudios previos que consultó, las personas investigadoras se centraron únicamente en una de las aristas y de ahí viene su aporte, en esa multidimensionalidad para dicho constructo.

En el mismo país, Tall (2014), señala una vez más que las habilidades de razonamiento matemático son un constructo multidimensional y multicomponente y que su medición puede tener variaciones en tanto no se puede aislar a la persona de sus conocimientos o vivencias previas. Para llegar a dichas conclusiones, recurre a la revisión de literatura desde los clásicos a lo más reciente.

En una línea similar y desde Argentina, Orlando (2014) en sus tesis doctoral plantea una revisión de insumos bibliográficos, documentales y de los resultados obtenidos en algunas pruebas como PISA y el Test de Razonamiento Matemático para analizar la incidencia de diversas variables en este constructo. En su caso, concluye que el razonamiento matemático se complementa con la solución de problemas y el rendimiento académico, del cual ya se ha hablado en esta propuesta por parte de otros autores. Destaca además que los resultados que se obtienen en la prueba que desarrolló, resultan ser un factor de predicción para el éxito académico. Es llamativo de esta investigación que como parte de los aspectos que considera para medir el impacto de variables externas en el razonamiento matemático, considera incluso el nivel académico alcanzado por los padres de familia.

En España, Cuesta et al. (2015), dan continuidad al tema de la medición del razonamiento matemático, pero en este caso lo complementan con el razonamiento verbal y el uso de las tecnologías. El proceso de investigación inició con una fase diagnóstica de las necesidades de los educandos para elaborar un módulo de formación sobre esos temas con el apoyo de recursos tecnológicos. Concluyen que su “trabajo ha supuesto un pequeño pero contundente paso hacia delante en el dominio de estas tecnologías como instrumento enriquecedor de las labores pedagógicas” (Cuesta et al., 2015, p. 49). Por ello se entiende que los recursos como las TIC pueden ayudar a enriquecer los procesos formativos, incluso en el caso del desarrollo de habilidades como el razonamiento matemático. Una de las características de este estudio fue que se realizó bajo una óptica cualitativa y consideró el seguimiento de una única persona estudiante como sujeto de investigación.

En el caso de Canadá, Lyons y Ansari (2015) señalan una vez más lo complejo que es la medición del razonamiento matemático en todos los niveles escolares, esto debido a los múltiples factores que inciden en él. Como parte de las conclusiones de su estudio, señalan que, aunque se quisiera, la medición de ese constructo no depende únicamente de factores académicos, sino que abarca otros aspectos, lo que contribuye a su complejidad. Para su estudio consideran una muestra de 1500 niños pertenecientes al sistema educativo de Canadá, específicamente en la escuela elemental y se les aplicó una prueba que se procedió a analizar y de ahí las conclusiones del estudio. El proceso se centran en aspectos de conteo y aritmética.

Rahmi y Surya (2017) realizan un estudio en Indonesia y para ello consideraron una muestra de 40 estudiantes del nivel VIII-3 de la secundaria Sablina Tembung Junior. Su metodología se basó en un enfoque cualitativo y, entre sus principales conclusiones sobre la medición del razonamiento matemático, destacan que, aunque la habilidad de razonamiento mostrada por los estudiantes es baja, su capacidad para resolver problemas matemáticos es mayor. Esto es coherente con lo que se ha señalado en otras investigaciones y a la vez sienta un precedente para el planteamiento de este tipo de pruebas por la complejidad que conlleva este tipo de constructos.

Sukirwan y Herman (2018), señalan que en el caso de Indonesia la tendencia es similar a lo ya expuesto y una vez más se indica la complejidad en el estudio e implicaciones de las habilidades de razonamiento matemático. En su caso buscaban revisar la calidad de la formación recibida para el desarrollo de esa habilidad, se centraron en dos perspectivas, la primera de ella fue el razonamiento imitativo y el segundo fue el creativo. Concluyen que algunas tareas ligadas con el razonamiento matemático se

realizan más por imitación que por creatividad propia de cada estudiante. Como muestra consideraron 33 estudiantes de IX año de la secundaria de la ciudad de Tangerang a los cuales se les aplicó un test de razonamiento matemático.

En el mismo país, Lestari y Jailani (2018), desarrollan un estudio con 122 estudiantes divididos en 4 salones de clase en una escuela secundaria. En este caso, la investigación consistió en exponer a las personas estudiantes a estrategias de clase que incluyeran actividades colaborativas, además de estrategias metacognitivas. Luego de esto, se les aplicó un test para medir los alcances de ese experimento y concluyen que aquellas actividades metacognitivas y colaborativas influyeron de manera positiva en el desarrollo de las habilidades por parte de las personas estudiantes.

En otra investigación más, Saxton et al. (2019), en Estados Unidos, recurren al análisis de bases de datos con resultados de distintas pruebas a fines al constructo de razonamiento matemático con el fin de analizar la tendencia de esos resultados y los factores que intervienen en la misma. Concluyen que el razonamiento matemático es complejo en su medición y también en su interpretación, esto debido a los procesos neuronales que conlleva y que, además, implica un manejo de símbolos, la comprensión de un lenguaje y la ejecución de procesos, también señalan que incluso el sentido de la visión influye en la medición de ese constructo.

En el caso de Cuba en 2016, Carballo y Guelmes (2016) realizan un estudio que se basa en la revisión de la literatura preexistente sobre el tema e indican una serie de recomendaciones para abordar distintas variables de investigación en el contexto educativo. Concluyen que tanto la reflexión como la revisión de la teoría y la metodología son parte importante de las orientaciones para saber cómo acercarse o medir

determinada variable en el campo educativo y que finalmente ese aspecto será una de las fuentes de validez y confiabilidad de esos estudios.

En México, Hernández (2017) realiza una propuesta de test adaptativo por medio de computadoras para medir el nivel de razonamiento matemático con el que cuenta una persona. Para ello realizó un acercamiento al constructo y en sí, a algunas de las variables que se acercan al mismo para poder desarrollar su propuesta. Señala que la implementación de un Test Adaptativo Informatizado permite que se realicen las mediciones sobre el constructo de una manera fiable y más económica que si fueran impresos en tinta y papel, además de que esta modalidad ofrece la posibilidad de administrar los ítems por medio de una base de datos y así generar distintas combinaciones aleatorias para varios tomadores del test al mismo momento.

Más recientemente, Cristancho (2019) en Colombia aborda el constructo del razonamiento matemático desde una perspectiva epistemológica. Para ello, recurre a la revisión de la literatura que consideró adecuada para el análisis del tema. Señala que, como se ha mencionado, un constructo como el razonamiento matemático puede intentarse medir, pero siempre habrá variables involucradas que no pueden ignorarse. Más bien, la consulta con expertos y el contar con información de diversas fuentes ayuda a conocer esos aspectos y, en un caso como el de esta propuesta de investigación, realizar los ajustes que sean necesarios en beneficio de toda la población.

De esta sección se resalta la complejidad de la medición del constructo de razonamiento matemático y de la interferencia que tiene de parte de varias dimensiones.

1.2.4. La Prueba de Habilidades Cuantitativas

Para culminar esta sección de antecedentes, se hará alusión a los resultados de algunos trabajos que han profundizado en la prueba específica que ocupa esta investigación, es decir, la PHC.

Al respecto Rojas (2014) en el contexto de la Universidad de Costa Rica consideró la cohorte de 2010 de las carreras de Física, Meteorología, Farmacia, Matemática y Ciencias Actuariales para analizar por medio de los resultados obtenidos en los cursos de iniciales de matemática de esas carreras y en la PHC si era posible predecir el rendimiento académico que tendrían los postulantes en sus carreras. Concluye que de estos elementos, el que mostró mayor potencia predictiva del éxito en las carreras fue la PHC. Es decir, a mayor nota obtenida en la prueba, mayor rendimiento académico o aprobación de cursos tendría cada persona estudiante.

En cuanto a la construcción de la prueba, se consideró analizar cuáles son los temas que se abordan en al área de matemática a nivel de la educación secundaria de Costa Rica, esto en tanto el abordar temas nuevos inducía a una varianza en la medición del constructo que no era relevante para el objetivo de la prueba. Por ello, se definió mantener las mismas áreas temáticas de la educación primaria y secundaria, es decir, geometría, aritmética, álgebra y análisis de datos, esto según Rojas y Ordóñez (2019).

A partir del esquema general de la prueba, se optó por conformarla con 40 ítems de selección única, pues esto aseguraba un equilibrio entre el número de ítems, el cansancio al realizar la prueba y su duración. El total de reactivos se dividió de manera equitativa en las 4 áreas ya mencionadas y por ello resultaron 10 ítems por temática (Rojas et al., 2019).

En la publicación de Rojas y Ordóñez (2019), se aborda con detalle el proceso de construcción y validación de la prueba y se indica cómo se han obtenido los cortes de aprobación vigentes a ese momento, sin embargo, estos no responden a otros métodos como el Angoff o bookmark, sino que se plantearon como una recomendación desde la Teoría Clásica de los Test (TCT) y de la Teoría de Respuesta al Ítem (TRI). Sientan también las bases para construir otras pruebas estandarizadas similares a la PHC con aplicaciones para otras áreas o constructos.

Por otra parte, Víquez et al. (2021) realizan una revisión de la literatura al igual que el caso anterior, y la realimentan con la experiencia que han acumulado en la PHC y proponen un folleto de introducción y preparación a la prueba para quienes deseen conocer más sobre ella. En el texto se brindan algunos ejemplos de ítems que se pueden encontrar en la prueba y de algunas recomendaciones para su realización.

En otras publicaciones como la tesis doctoral de Ordóñez (2023) y en el artículo de Ordóñez y Rojas (2024) los investigadores buscan evidencias sobre las habilidades de razonamiento cuantitativo que muestran un grupo de estudiantes en la PHC y las comparan con las requeridas para los cursos de Química General I y de Introducción a la Química. Para ello, recurren a la aplicación de protocolos de pensamiento en voz alta con el objetivo de conocer las estrategias que las personas examinadas aplican para resolver distintos ítems de la PHC. Además, cotejan los resultados obtenidos luego de observar a los estudiantes con las estrategias de solución a problemas que empleaban los docentes en los cursos en los que se enfocó la investigación.

Como parte de las conclusiones de los estudios, se confirma que las habilidades de razonamiento cuantitativo que son requeridas para los cursos de Química,

corresponden a las habilidades de relacionar, clasificar, ejemplificar, así como a las de realizar inferencias de veracidad o falsedad para una proposición y a las habilidades para cuantificar o calcular. Además, se evidencia una relación de predicción entre la nota obtenida en la PHC y el resultado obtenido en los cursos de Química.

Además, en Ordóñez y Rojas (2024) se evidencia que en los procesos de resolución de problemas del área de Química que emplean los profesores en sus lecciones, recurren a los componentes del razonamiento cuantitativo, aunque en muchas situaciones pasa desapercibido.

En una línea cercana de investigación, Mora y Rojas (2023a) realizan un análisis de los procesos de respuesta a una serie de problemas de razonamiento cuantitativo que resuelve un grupo de 20 estudiantes que obtuvieron notas altas en la PHC. Para ello, también recurren al pensamiento en voz alta para registrar los procedimientos empleados por los sujetos de estudio. Concluyen que los procesos que desarrollaron los estudiantes coinciden con los constructos de razonamiento cuantitativo para los que se construyeron los reactivos analizados.

En cuanto a los aportes de los antecedentes a la propuesta de investigación, no se ha encontrado referencia a un proceso de indagación que dé lugar a la construcción de puntos de corte para la PHC, considerando sus implicaciones más allá del requisito de ingreso para las carreras que lo consideran de ese modo y cómo estas podrían incidir en el desarrollo académico y profesional de los examinados.

También, se concluye de la revisión realizada que las pruebas estandarizadas son apropiadas para la medición de los constructos como el razonamiento matemático, pero que esta tarea es compleja por las dimensiones que posee ese constructo lo que se

denota en los estudios citados y es un insumo a considerar para la propuesta investigativa. Además, se ha abordado el tema de los puntos de corte desde distintos puntos de vista, los cuales incluyen la TRI, la TCT, los métodos Angoff y Bookmark, sus variantes y combinaciones. En cuanto a los tamaños de muestra, es importante resaltar que es variable y depende del objetivo y escenario de estudio, al igual que el elegir un estilo cualitativo o cuantitativo. Por ello, una combinación de ambos fue la opción que se consideró en el escenario propuesto y esto se debería complementar con una elección de una muestra amplia para consolidar resultados más confiables y válidos.

1.3. Delimitación de la problemática a investigar

Para este apartado de la delimitación de la problemática, se iniciará con la justificación de la investigación, a continuación la pregunta problema que guía la investigación y finalmente se considerarán los objetivos del estudio y algunos aspectos que permiten delimitar el trabajo por desarrollar.

1.3.1. Justificación

La literatura sobre el diagnóstico del nivel de preparación de estudiantes universitarios en Costa Rica resalta el esfuerzo de la Universidad de Costa Rica en el desarrollo de pruebas estandarizadas para aspirantes a carreras con alta carga matemática, como Física, Química, Farmacia, Estadística, Matemática, Meteorología e ingenierías, esto se constata en publicaciones como Rojas (2014) y Rojas y Ordoñez (2019). Estas disciplinas, donde el éxito en asignaturas como Cálculo I es fundamental, presentan desafíos significativos para estudiantes con habilidades cuantitativas insuficientes, lo que se traduce en tasas de reprobación y repitencia preocupantes (Rojas-Torres, 2014). Una limitación importante en el uso actual de la Prueba de Habilidades

Cuantitativas (PHC), a pesar de su valor diagnóstico, es la ausencia de puntos de corte validados que permitan una interpretación clara y uniforme del nivel de habilidad de los aspirantes. Esta falta de criterios estandarizados puede generar incertidumbre en los estudiantes sobre su preparación real y dificulta la implementación de estrategias de apoyo académico precisas, afectando potencialmente su trayectoria y el rendimiento institucional.

También, en Rojas (2014) y Rojas y Ordoñez (2019), se destaca que el interés por parte de la Universidad se ha concretado por medio de la creación y aplicación del examen de Diagnóstico en Matemática o el DiMa y desde el año 2010 se reforzó con la Prueba de Habilidades Cuantitativas, la cual por medio de la evaluación del constructo de razonamiento matemático, busca seleccionar a las personas estudiantes que poseen mejores habilidades para encarar los primeros cursos de esa área en sus carreras y así predecir de manera eficaz el éxito que tendrán en dichas asignaturas (Rojas-Torres, 2014). Sin embargo, la ausencia de puntos de corte validados limita la capacidad de interpretar los puntajes de la PHC en términos de niveles de habilidad específicos. Esto introduce un riesgo de subjetividad en la toma de decisiones y dificulta la comunicación clara del nivel de preparación de los estudiantes, comprometiendo la efectividad predictiva que se busca con la prueba.

Pese al tiempo durante el cual se ha implementado esta prueba, la misma no posee una categorización de las habilidades de cada persona que la realiza referida a un punto de corte, si no que se hace con apoyo de la Teoría de Respuesta al Ítem (TRI) y de la Teoría Clásica de los Test (TCT), los encargados de la PHC realizan una recomendación de nota mínima a las unidades académicas, pero tal y como se visualiza en la resolución

VD-11757-2021 de la Vicerrectoría de Docencia de la Universidad de Costa Rica, esta nota es variable para cada cohorte y las unidades académicas son quienes definen el valor final a considerar. Esta variabilidad en los criterios introduce un riesgo de inconsistencia e inequidad en la evaluación diagnóstica

Con esta investigación, lo que se pretende es poder complementar las notas mínimas ya establecidas con notas de puntos de corte que propicien que esta prueba estandarizada pueda brindar información para la Universidad y las unidades académicas de las herramientas o habilidades con las que cuenta una persona postulante a sus carreras y que pueden filtrar cuáles de esas destrezas son indispensables y cuáles podrían estar en otros niveles de relevancia, lo que permitirá la toma de decisiones más informada y la posible implementación de estrategias de apoyo académico más focalizadas.

Específicamente, lo que se plantea es colaborar con jueces expertos en el tema y con personas clave de cada unidad académica interesada en la PHC para establecer notas mínimas de aprobación y reactivos del banco de ítems de la prueba en las que se valoren determinadas habilidades y que sean significativos en la categorización de las personas estudiantes según el constructo de razonamiento matemático implícito en la PHC. Este proceso que se basa en el juicio de expertos permitirá definir niveles claros para la categorización de los estudiantes basados en el desempeño y que proporcionarán niveles de referencia para interpretar sus resultados de manera más precisa.

Los estudios que se han realizado sobre la prueba se centraron en su proceso de elaboración e implementación, así como sus niveles de dificultad, entre ellos Rojas y Ordoñez (2019), sus evidencias de confianza y validez, por ejemplo Rojas (2013), Rojas

(2014) y Rojas y Ordoñez (2019). También, algunos otros escritos abordan su potencia predictiva, como Rojas (2013) y algunas prácticas previas para resolver dicha prueba, entre ellas Herrera (2016) y Víquez et al. (2021). Sin embargo, esta investigación se distingue al centrarse específicamente en el establecimiento de puntos de corte basados en el análisis detallado de las habilidades de razonamiento matemático necesarias para el éxito en Cálculo I. Este enfoque busca llenar un vacío importante al proporcionar criterios específicos para la interpretación de los resultados de la PHC en términos de niveles de habilidad.

Lo anterior quiere decir que si se consideran los estudios que se han realizado hasta el momento y se contrastan con esta propuesta, la relevancia de esta se centra en la posibilidad de establecer niveles de habilidad en una prueba sobre razonamiento matemático en nuestro país, además de implicar el ejercicio de adaptar los métodos bookmark y Angoff para establecer los puntos de corte por primera vez para una prueba de ese tipo en Costa Rica.

La pregunta central que guiará esta investigación será ¿cuál es el punto de corte de la PHC asociado al nivel de razonamiento apropiado para cursar la asignatura de Cálculo I? Ante esta cuestión, también surgen las siguientes preguntas ¿cuáles niveles de desempeño o habilidad se establecen para la PHC?, ¿qué habilidades son las que requieren las personas estudiantes para afrontar el curso de Cálculo I?, ¿cómo se establecen los puntos de corte en pruebas estandarizadas?, ¿qué niveles de desempeño surgen y qué significan una vez planteados los puntos de corte con los informantes?

La información obtenida beneficiará a la Universidad de Costa Rica, a sus estudiantes, a las unidades académicas y al país. En particular, permitirá contar con un

mayor conocimiento sobre las habilidades matemáticas de los aspirantes a la UCR que aplican a carreras con el requisito de la PHC y que, además, deben cursar MA1001 o asignaturas similares. Esta información permitirá a la institución tomar decisiones de admisión más informadas, a las unidades académicas diseñar estrategias pedagógicas más efectivas, y a los estudiantes comprender mejor sus fortalezas y áreas de mejora.

El principal aporte de esta propuesta es definir el significado de cada rango de calificaciones en la PHC en términos de las habilidades de razonamiento matemático de los estudiantes. Esto como complemento a los hallazgos de otros estudios sobre la incidencia de la PHC en el desempeño académico, como los presentados en los textos de Rojas (2013), Rojas (2014), Rojas y Ordoñez (2019), Ordóñez (2023) y Ordóñez y Rojas (2024).

Como ya se mencionó, el tema que se propuso para la investigación fue el *establecimiento de puntos de corte para pruebas estandarizadas referidas a criterios, el caso de la prueba de Habilidades Cuantitativas* y a raíz de eso, el objeto de estudio que se ha considerado corresponde a los *puntos de corte para pruebas estandarizadas*.

Respecto a la relación del área temática planteada con el énfasis de la maestría, cabe señalar que esta se evidencia en tanto la propuesta de investigación busca un acercamiento al análisis y establecimiento de los puntos de corte para una prueba estandarizada referida a criterios, tópico el cual forma parte de los objetos de estudio de la maestría a la que pertenece esta propuesta. Además, esta prueba se emplea en el campo educativo, específicamente como requisito de ingreso de los postulantes a varias carreras de la Universidad de Costa Rica, las cuales han definido desde su unidad académica la necesidad de realizar esta prueba para determinar el nivel de razonamiento

matemático con el que cuentan los postulantes a sus carreras y así predecir el posible éxito que tendrán los mismos en los primeros cursos de matemática de sus carreras.

Sobre las implicaciones sociales y éticas que conlleva el desarrollo de este estudio, se destaca el posible impacto para las personas estudiantes que tendrá el indagar sobre el nivel que poseen las personas que se postulan a ingresar a la UCR en cuanto a sus habilidades para el razonamiento matemático, tema que podría ser abordado no solamente por la Universidad, sino también por las autoridades del MEP en busca de la mejora de los procesos formativos. Por otro lado, a nivel ético, expondría algunos elementos que si bien no forman parte de los elementos explícitos del proceso de admisión a una carrera, podrían considerarse como complementarios y a la vez orientadores para que las personas estudiantes de nuevo ingreso se preparen mejor al inicio de su carrera universitaria y que a la vez saquen provecho de las opciones de nivelación que ofrece la Universidad y no como una imposición, sino como posibilidad para mejorar y enfrentar con éxito el inicio de su carrera universitaria.

Por otra parte, considerando lo expuesto en los antecedentes de esta investigación, no es posible constatar que la PHC cuente con evidencias sobre la construcción de puntos de corte para el ingreso a la carrera y que además considere los alcances de la misma más allá del cumplimiento de los requisitos fijados por las carreras que la solicitan. Lo anterior responde a que las investigaciones referenciadas previamente sobre la prueba se han centrado en la valoración del éxito académico de las personas estudiantes en función de los resultados que obtienen en ella, es decir, su potencia predictiva para enfrentar el desarrollo de una carrera universitaria en distintas áreas.

1.3.2. Pregunta problema

La pregunta que orientó la investigación fue ¿cuál es el punto de corte de la PHC asociado al nivel de razonamiento cuantitativo apropiado para cursar Cálculo I?

1.3.3. Objetivos

Los objetivos de la investigación son:

Objetivo general: Analizar los puntos de corte apropiados para la PHC que indiquen el nivel de razonamiento cuantitativo adecuado para cursar Cálculo I

Objetivos específicos:

- Determinar los puntos de corte para la PHC a partir de la aplicación de los métodos Angoff y Bookmark.
- Describir los niveles de competencia de las personas estudiantes en cuanto a razonamiento cuantitativo según los valores obtenidos para los puntos de corte.
- Obtener evidencias de validez del uso de los puntos de corte establecidos.

Considerando los objetivos de esta investigación y su alcance, en la siguiente sección se exponen los fundamentos teóricos que permiten explicar los resultados obtenidos en el trabajo de campo.

2. Marco teórico

Este apartado se ha organizado en 4 secciones, en la primera se aborda la temática de la evaluación con pruebas educativas, en el segundo se diferencias los tipos de pruebas escritas, en el tercero se aborda el proceso de establecimiento de puntos de corte y en el cuarto se detallan los aspectos teóricos y las características de la PHC.

2.1. Evaluación con pruebas educativas

La evaluación de los aprendizajes ha sido un tema de interés en el área de educación y muestra de ello es la variedad de publicaciones que se pueden encontrar. Algunas de estas investigaciones como por ejemplo la de González et al. (2020), indican que la evaluación ha trascendido de una óptica tradicional en la que se le consideraba como un mecanismo de control y clasificación de los estudiantes como buenos y malos a uno contemporáneo en el que la evaluación es esencial para verificar los aprendizajes y tomar decisiones sobre el proceso formativo.

En el caso específico de la evaluación educativa, autores como Arribas (2017), destacan el papel que cumple en el monitoreo del proceso de formación, así como las implicaciones que puede tener en la vida de las personas estudiantes y en la valoración para la mejora continua de los procesos de formación de los que participan. Señala que los alcances de la evaluación pueden ir desde la verificación de la modificación de conductas, el cumplimiento de objetivos hasta el dar fe de si se cuenta o no con cierta competencia.

Por otro lado, tanto a nivel nacional como internacional, la evaluación educativa ha recibido críticas en cuanto a los mecanismo para la rendición de cuentas y hasta

transfondos políticos bajo un modelo neoliberal, lo cual puede resultar una limitante respecto a la labor que puede cumplir y el cómo se percibe esta junto con sus implicaciones (Parcerisa et al., 2022).

Desde un punto de vista más integral del proceso de evaluación, se le puede ver como lo sugiere Hernández et al. (2018, p. 150) es decir, “una herramienta de aprendizaje” . Para ello se supone que las experiencias de evaluación estén relacionadas con el proceso de formación en el que se integra a las personas estudiantes, a la vez que los resultados impactan en dicho proceso con miras a la mejora.

Al respecto, Arribas (2017, p. 383) indica que la evaluación de los aprendizajes es “el proceso de recogida, análisis e interpretación de resultados con el fin de valorarlos y que conlleva una toma de decisiones”. Esto refiere a que la evaluación funciona como método de recolección de información para la toma de decisiones.

De las definiciones que se encuentran en la literatura para la evaluación, por ejemplo Arribas (2017) o Jordán et al. (2018), se remarca la convergencia del concepto en tres elementos fundamentales que les sirve de características: el análisis de resultados, la valoración y finalmente la toma de decisiones en función de los hallazgos de los dos elementos previos.

Para efectos de esta investigación, y siguiendo la línea de Arribas (2017), se comprenderá la evaluación de los aprendizajes como aquel proceso por medio del cual se emite un juicio de valor a partir de las evidencias recolectadas en un determinado contexto y con una clara intención de considerar los resultados para la toma de decisiones sobre los procesos formativos. Además, se destaca que corresponde a un proceso, el cual es sistémico, intencionado, objetivo, continuo y enmarcado en un proceso de

formación. Lo anterior en tanto que con esta se busca recolectar y analizar información para mejorar el mismo proceso educativo en sus distintos niveles.

Basado en lo anterior, la PHC coincide con esta definición pues busca recolectar información a partir de la medición que se realiza de las habilidades en un campo específico por parte de las personas que finalizaron un proceso educativo como la educación secundaria y se perfilan para ingresar a la Universidad de Costa Rica, sin embargo, el aporte de esta propuesta de investigación es la oportunidad de enriquecer la interpretación de la información obtenida y la posible toma de decisiones a partir de ella.

2.1.1 Tipologías de la evaluación

La evaluación se puede organizar e incluso jerarquizar bajo distintos parámetros, entre ellos se tiene a los momentos, los agentes, las intenciones o sus funciones. En este punto, es de interés enumerar algunas de estas clasificaciones, las cuales son resumidas por García (2020) y se retoman en este documento. En primer lugar, se reconoce la diferencia entre la evaluación diagnóstica, procesual y final, nótese que esta diferenciación responde al momento del proceso formativo en el que se aplican, ya sea antes, durante o al final. Por otra parte, García (2020) así como Sánchez y Martínez (2022) señalan que se tiene el trinomio de la evaluación que se conforma por aquella que sirve de diagnóstico y que se suele realizar al inicio de un curso, de un tema o de una unidad didáctica. Por otra parte, la evaluación sumativa que se refleja en la nota de aprobación de un proceso y que tradicionalmente se aplica al cierre de un curso o de un proceso de formación y finalmente la evaluación formativa que consiste en una vía para recolectar información para la mejora del proceso, se aplica de manera paralela al

desarrollo de las actividades de mediación y no representa un valor explícito en las notas de aprobación para el curso.

Se destaca en el caso de la evaluación formativa que el rol de la persona estudiante debe ser activo, esto en tanto se busca darle voz sobre las modificaciones o ajustes que se pueden hacer en el proceso (F. J. García et al., 2019).

García (2020), indica que otra diferenciación corresponde al paradigma seleccionado para el desarrollo del proceso evaluativo y plantea dos posibilidades: la ruta cuantitativa y la cualitativa. También se debe diferenciar entre una evaluación normativa, una criterial y una personalizada, estas se distinguen por los criterios de comparación para los juicios. El primero considera el desempeño de un individuo respecto a sus pares, el segundo filtra al individuo a partir de una serie de estímulos o reactivos que dan cuenta de las competencias que posee y el tercero sigue el proceso de cada individuo de manera aislada y se centra en el proceso que sigue esa persona y su evolución. También se considera la selección según los agentes que realizan la evaluación. Puede tratarse de una persona valorando su propio proceso, otra persona evaluando el desarrollo de un individuo o un equipo de personas evaluándose entre sí y a sí mismos. Con base en estas características, se identifican tres tipos de evaluación: autoevaluación, heteroevaluación y coevaluación.

García (2020), continua la clasificación señalando que el ambiente formativo también incide en la clasificación de la evaluación y a raíz de ello se distingue el tipo presencial, a distancia y la mixta, cada una de ellas responde al lugar en el cual se gestiona el proceso evaluativo.

Considerando la clasificación de la evaluación abordada hasta este punto, se puede afirmar que la PHC posee características similares a una prueba diagnóstica. Esto se debe a su momento de aplicación y a su objetivo de determinar las habilidades de razonamiento cuantitativo de los examinados. Además, se caracteriza por ser cuantitativa y criterial, y según el agente que la aplica que es un instituto que es parte de la UCR, corresponde a una heteroevaluación que se realiza de manera presencial.

2.1.2. Evaluación referida a criterio

En apartados previos de esta sección se hizo mención de un listado de las posibles tipologías de evaluación, entre ellas se realizó una diferenciación de acuerdo con los puntos de referencia que se toman para comparar a los individuos que están sujetos a la evaluación. Particularmente, se considera la evaluación referida a criterio porque permite contrastar las capacidades de una persona respecto a una serie de indicadores que se deben cumplir para garantizar que se posea o no una habilidad.

Una prueba con referencia a criterio es aquella que “se emplea para determinar la posición de un individuo con respecto a un dominio de la conducta perfectamente definido” (J. Popham, 1983, p. 134). Es decir, no importa la posición o habilidades de esta persona en cuanto a su grupo compañeros o pares, lo que interesa es su resultado individual en comparación con los criterios que se hayan definido previamente (Villalobos et al., 2021). Por ejemplo, en Costa Rica, la aprobación de la prueba práctica para obtener el permiso de conducir requiere una nota superior a 80. Esto implica que no se compara el desempeño de los examinados entre sí, sino que la evaluación se basa en estándares previamente establecidos.

En esta investigación, la temática resulta pertinente porque la PHC es una prueba que refiere a criterios pues en primer lugar no resulta indispensable asignar una nota de aprobación, ni tampoco se pretende definir si una persona postulante se encuentra en determinado percentil para cumplir un requisito para acceder a una carrera, lo que interesa es conocer cuántos de los ítems se logró resolver de manera correcta y en función de ese resultado, determinar el nivel de competencia en el constructo del razonamiento cuantitativo.

2.2. Tipos de pruebas escritas

Con respecto a pruebas educativas, Villaroel y Bruna (2019) destacan que estas se deben aplicar por medio de textos escritos en un cuadernillo o formulario, pero más allá de eso, deben medir las habilidades, o en su defecto los aprendizajes, de manera contextualizada y para ello se sirven de estímulos o ítems que debe ser realistas, situados y que respondan a situaciones significativas y retadoras. Además, indican que para su solución se debe requerir la aplicación de habilidades cognitivas complejas.

En el caso de la PHC, su cuadernillo consta de 40 ítems de selección única y tiene como objetivo que los estudiantes apliquen sus habilidades de razonamiento cuantitativo para resolver enunciados en contextos reales. Por ello, puede incluirse dentro de esta clasificación.

Por otro lado, a continuación se realiza una diferenciación entre las pruebas de aula y las pruebas estandarizadas, las cuales difieren por los alcances y las consecuencias que poseen.

2.2.1. Pruebas de aula

Las pruebas de aula para Mahias y Polloni (2019) son aquellas que aplica una persona docente como parte de un proceso formativo, pero principalmente sumativo. Se caracterizan por responder al contexto específico del aula y consideran los rasgos particulares de la audiencia a la que va dirigida, misma que suele ser reducida. La preparación de dichas pruebas es responsabilidad de la persona encargada del proceso formativo. En cuanto a las consecuencias de la prueba, se limitan a mejorar el proceso que se desarrolla en el aula pero su alcance es reducido, tal y como lo indican Sitotaw y Tadele (2018) y Carvalho (2018).

Una de las desventajas que poseen las evaluaciones de aula es que las personas docentes “desarrollan experiencias experimentales, tienen en cuenta sus vivencias y expectativas” (F. J. García et al., 2019, p. 5). Además, Giménez et al. (2021) subrayan que las pruebas de aula se realizan con el empleo de pruebas escritas con enunciados formales y alejados de las vivencias e intereses del estudiantado. A lo anterior se suma que los ítems se preparan con poco conocimiento del área de evaluación, con poca claridad de los objetivos y habilidades que se desean evaluar (Gonzalez et al., 2017). Por ello, las características específicas y el proceso de construcción, aplicación y análisis de la PHC cobran más valor, pues la misma no se ve afectada por las limitaciones de las pruebas de aula.

Autores como Escorcia et al. (2014) señalan que la evaluación a nivel de aula se centra más en las valoración de rasgos normativos de la disciplina que en la aplicación de los conocimientos y habilidades en el contexto de la vida cotidiana.

Una de las limitaciones de las evaluaciones de aula, expuesta por Benítez y López (2018), se vincula con el juicio que emite el evaluador, pues usualmente es el mismo docente quien las realiza. Los autores lo resumen en una frase: “lo que habría que razonar es si el juicio es justo o está cargado de subjetividades por parte del docente” (p. 69).

2.2.2. Pruebas estandarizadas

Las pruebas estandarizadas se han configurado como una herramienta que permite realizar evaluaciones a grandes grupos de personas, aunque las percepción del tamaño del grupos puede variar en función del área disciplinar y de los objetivos de la evaluación. Con este tipo de pruebas se asume que la población tiene cierto grado de uniformidad y puede actuar como filtro para la entrada a un nuevo nivel escolar, además, en un marco de globalidad, permiten equiparar ciertas características deseables en países con grandes diferencias a nivel cultural e incluso académico, lo cual permite igualar algunas condiciones para competir en el mercado global (Parcerisa et al., 2022).

Algunos ejemplos cercanos de estas pruebas son las PISA, las pruebas de aptitud académica para ingresar a centros educativos o la misma PHC. Otros ejemplos de estas pruebas pueden ser las que se implementan para demostrar el dominio de alguna lengua extranjera, por ejemplo el TOEFL o el TOEIC en el caso de la lengua inglesa o también, algunas pruebas que se utilizan en el sector salud, por ejemplo, en el caso de Turquía, se aborda un ejemplo de implementación de una prueba para valorar la reserva cognitiva y para ello ajustan un test, que fue construido previamente, a las necesidades de la población que se consideró para la investigación que pretendía validar dicha prueba en ese contexto (Ozakbas et al., 2021).

Por otro lado, Demarchi (2020) expone que una prueba estandarizada se puede comprender como aquella que se aplica a una amplia población de estudiantes, considerando que entre ellos existe una serie de características homogéneas que los hacen ser sujetos de una evaluación similar.

Entre las ventajas de la construcción de pruebas estandarizadas se encuentra la posibilidad de disminuir “la subjetividad de las valoraciones, fijar parámetros de desempeño necesarios para el establecimiento de metas educativas” (M. Hernández et al., 2018, p. 150). A esto se suma que las evaluaciones estandarizadas se aplican por parte de un equipo de especialistas en el área, en evaluación o en psicometría, con ello se busca asegurar un conocimiento óptimo de “el estatus pedagógico y epistemológico de la disciplina en cuestión” (D. D. Pérez et al., 2020).

Además, tal y como lo señalan Gutiérrez y Acuña (2020), dos de las características que se asocian con las pruebas estandarizadas y que se señalan en algunas investigaciones como una de sus ventajas, son las evidencias de validez y confiabilidad, las cuales se pueden analizar debido a la perduración de los instrumentos por sus múltiples iteraciones, además de la existencia de especialistas en el campo y que asesoran el proceso de construcción y aplicación de estas pruebas.

En cuanto a sus alcances, Martínez et al. (2017), destacan que estas pruebas no se limitan a determinar si una persona estudiante puede reproducir lo que aprendió, sino que en el escenario de estas pruebas deben ser capaces de extrapolar lo aprendido y aplicarlo a circunstancias desconocidas dentro y fuera del aula.

Dentro de los aportes de las pruebas estandarizadas, también se destaca que son los instrumentos de medición más empleados en Psicología, Educación, Ciencias de la

Salud y Ciencias Sociales, y que además cuentan con un amplio desarrollo a nivel técnico y metodológico que permiten su mejoramiento para medir los rasgos observables o latentes, en la población focal específica y con un grado de precisión previamente establecido y controlado por procedimientos logísticos y administrativos e igualmente objetivos (Tristán y Pedraza, 2017).

Los rasgos característicos de una prueba estandarizada coinciden con las características de la PHC. Uno de los supuestos de aplicación de esta prueba es la homogeneidad de los examinados, ya que todos provienen de la educación diversificada y han cubierto un currículo mínimo común en todo el sistema educativo costarricense. Además, la PHC es desarrollada por un equipo de especialistas en el área, lo que garantiza su especificidad, objetividad y confiabilidad en el análisis de los resultados.

Finalmente, es oportuno destacar la noción de las pruebas estandarizadas de altas consecuencias, categoría a la cual pertenece la PHC por sus implicaciones actuales en la vida de las personas, estas pruebas son confeccionadas para obtener información para mejorar el proceso educativo y sus resultados conlleven un impacto en la vida de estudiantes, docentes y escuelas" (Torres y Contreras, 2022).

2.3. Establecimiento de puntos de corte

El objetivo principal que se persigue con esta propuesta investigativa es el lograr definir de manera consistente y objetiva los niveles de desempeño de las personas estudiantes que realizan la PHC previo al ingreso a las carreras de la UCR a las que se postulan, estos niveles de habilidad corresponden a lo que la literatura señala como puntos de corte, es decir, son unas notas determinadas que resumen la habilidad con la que cuenta una persona en relación con determinado constructo.

Con el fin de establecer los resultados deseables en las pruebas, se han realizado varios esfuerzos que recurren a la aplicación de la TRI y la TCT para definir los valores de aprobación. Otros esfuerzos se han dirigido a determinar una nota de corte para la aprobación del examen, entre los cuales se han utilizado la psicometría, la conformación de jurados, el establecimiento de metas políticas e incluso la deseabilidad social de un rasgo o conducta. Sin embargo, se debe señalar que en general, estos enfoques han recibido críticas al implementarlos en distintas pruebas debido a su grado de subjetividad (Kampa et al., 2019b).

Retomando los puntos de corte, se destaca que están referidos a una serie de criterios, que usualmente son fijados por jueces y permiten valorar el alcance de los objetivos de un plan o de un proceso de formación y facilitan la clasificación de las personas estudiantes en niveles según sus logros o habilidades (Sondergeld et al., 2020).

Lewis y Cook indican que esa definición de las habilidades de las personas, corresponden más bien a niveles de desempeño y la implementación de los métodos se ocupa de la tarea de identificar cuáles son los ítems que marcan la diferencia entre un nivel y el otro (2020).

En el texto de Shin y Lidster (2017), se hace la diferencia entre dos tipos de métodos para fijar los puntos de corte, el primero, refiere a aquellos que se centran en la prueba y en ellos, un grupo de expertos analiza los ítems de la prueba para decidir los niveles de rendimiento de una persona mínimamente competente en el área. En el segundo, la atención se pone en el examinado y acá, los expertos identifican a algunas personas estudiantes que pueden ser ejemplos de los estándares de rendimiento y en función de los resultados de esos estudiantes se fijan los puntos de corte.

Ahora bien, como una parte del proceso de evaluación, la fijación de los puntos de corte para algunos tipos de pruebas ha recibido la atención de varios expertos en el tema y es por ello que se cuenta con varias opciones para realizar esa tarea. Incluso, algunas de las investigaciones que se han analizado para este documento y que refieren a la fijación de puntos de corte, por ejemplo Clauser et al. (2017) o Durán del Fierro (2019), aconsejan utilizar al menos dos de los métodos para dar mayor fiabilidad de la tarea que se ha asumido. Por ejemplo, en la literatura reciente que se ha analizado, se cuenta con un caso en el que se fundamenta la necesidad de contar con la orientación apropiada para que los jueces implementen los métodos de punto de corte, en este caso, se partió de las recomendaciones y la propia metodología de los métodos Angoff y el Angoff extendido, esto a fin de garantizar la validez de los puntos de corte establecidos para una prueba desde el punto de vista de quienes integraron el jurado, los resultados obtenidos se corroboraron con pruebas de ANOVA para verificar la pertinencia de los resultados (Kampa et al., 2019b). Otro caso similar se origina en Corea, pero en esa ocasión se recurre a la fundamentación teórica de los métodos Angoff, el Ebel modificado y el método Hofstee (Park et al., 2020).

Previo a la mención de los métodos de punto de corte que se considerarán en esta investigación, es importante hacer mención del modelo de Rasch (1960), en tanto este define algunos de los postulados que dan pie al uso de los métodos para fijar puntos de corte.

El modelo de Rasch constituye una de las opciones disponibles en el marco de la Teoría de Respuesta al Ítem que surge para solventar algunas de las limitaciones de la Teoría Clásica de los Tests, misma que dejaba de lado algunos aspectos que podían

intervenir al momento de determinar la habilidad de los individuos ante determinado constructo. El modelo asume que es posible medir una habilidad latente por medio de las respuestas que un individuo brinde a un ítem específico de una prueba, por ello es uno de los modelos de Teoría de respuesta al ítem de un parámetro más usados para medir habilidades no observables y se empleó por primera vez para medir la inteligencia de los soldados daneses, pero de ahí se ha extendido al campo educativo e incluso al económico.

Este modelo también considera que esa habilidad latente se mide en la prueba de manera unidimensional y que se mantiene invariante y también requiere que los ítems sean independientes los unos de los otros. Además, sugiere que hay un comportamiento de escala lineal en el logro de las habilidades que se busca medir, es decir, no hay saltos importantes entre un nivel de habilidad y el siguiente superior o inferior.

Adicionalmente, el modelo de Rasch considera que al momento de determinar la probabilidad de acertar un ítem por parte de un individuo, se debe considerar dos aspectos: por un lado el parámetro de la dificultad del individuo y por otro, el parámetro de la dificultad del ítem (Cizek y Bunch, 2007).

Desde el punto de vista matemático, el modelo se expresa por medio de la fórmula:

$$p(x = 1|\theta_i, \beta_j) = \frac{e^{\theta_i - \beta_j}}{1 + e^{\theta_i - \beta_j}}$$

Se considera a $p(x = 1)$ como la probabilidad de acertar un ítem que posee una dificultad representada por β_j y una habilidad del individuo denotada por θ_i. En otras palabras, de expresa la probabilidad de acierto como la división de la potencia del número

e con exponente igual a la diferencia entre la habilidad del individuo y la dificultad del ítem, dividido por uno más la misma potencia.

Bajo la óptica de este modelo, se considera que en caso de existir alguna diferencia en la nota obtenida en determinada prueba, la misma se vincula con una variación en la habilidad de los sujetos evaluados. Es decir, si ante la aplicación de una prueba que mide determinado constructo el sujeto A obtiene una nota inferior a la del sujeto B, esto implica que la habilidad del sujeto A en el marco de ese constructo, es menor que la del sujeto B.

Algunas características del modelo es que asume que en caso de que el sujeto haya acertado un determinado ítem del test con una dificultad X, entonces debería haber acertado todos aquellos reactivos que tengan un nivel de dificultad menor a X. Esto cobra relevancia si se considera un test en el cual todos los ítems se ordenan de menor a mayor dificultad, pues si se respondió de manera correcta un determinado ítem, todos los anteriores también deberían estar correctos.

Tomando en cuenta lo expuesto hasta este punto, a continuación se abordará el detalle de dos de los métodos existentes que resultan relevantes para esta propuesta de investigación.

2.3.1. El método bookmark

El método bookmark consiste en la selección de los ítems que conforman el banco de la prueba para confeccionar un cuadernillo, ordenados de menor a mayor dificultad a partir del nivel que señala el análisis psicométrico de cada uno a partir de la TRI. Posteriormente, un equipo de jueces revisa esa prueba e identifica cuál es el ítem que

marca un nivel diferenciado de habilidad entre los examinados, con esto se logra definir una brecha entre un nivel de habilidad y otro. Luego, las decisiones de los jueces se someten a discusión por el panel de expertos con el objetivo de consensuar la ubicación de ese ítem marcador de los niveles de habilidad (Jalalizadeh et al., 2019).

Una de las características que se señala para este método desde textos como el de Mitzel et al. (2001) y el de Feseker et al. (2021), es que permite simplificar la tarea de los jueces en tanto resulta más sencillo identificar una pregunta que sirva de filtro que el intentar definir el nivel de complejidad real de cada ítem aislado.

El método bookmark o el método del marcador parte de la teoría de respuesta al ítem y considera la revisión del compendio de los ítems que conforman una prueba para la que se desea fijar un punto de corte. Es por ello que se inicia construyendo un folleto con los ítems de la prueba ordenados de menor a mayor dificultad y luego los jueces insertan un marcador en algún punto de dicho folleto.

Este método también tiene la ventaja de que se puede aplicar a otros tipos de ítem que no sean solamente dicotómicos o de selección múltiple, por lo que es oportuno para procesos de fijación de puntos de corte más complejos que los que atañen a otros métodos.

También, se señala como ventaja que el método considera la prueba como una unidad y esto es coherente con la realidad que enfrentará la persona al momento de desarrollar la prueba.

Otro acierto de este método es que facilita la tarea de los jueces, incluso si se debiera considerar más de dos niveles de habilidad al fijar los puntos de corte. Por ejemplo, si fuera necesario considerar los niveles básico, competente y avanzado,

bastará con que cada juez introduzca un marcador en dos puntos distintos del folleto que se le entrega y se entenderá que la posición del primero indicará un menor nivel de dificultad que el segundo marcador que se utilice.

La estimación de la nota de corte bajo el método bookmark, no considera la revisión detallada de cada ítem para fijar su dificultad, sino que considera la prueba como un todo, además, reduce la posibilidad de error y reduce el tiempo requerido para establecer el punto de corte.

La implementación tradicional del método bookmark implica la construcción de un cuadernillo de la prueba en el que los ítems se ordenan según su nivel de dificultad. Posteriormente, cada juez analiza este documento para indicar cuál ítem representa el nivel mínimo de habilidad requerido para aprobar el test, considerando el perfil deseable del estudiante. Este perfil deseable, previamente definido, sirve como referencia para que los jueces puedan realizar una evaluación objetiva y coherente. Estos aspectos se detallan a continuación.

2.3.1.1. El folletillo de la prueba

El cuadernillo o folletillo de la prueba es uno de los elementos característicos y fundamentales en todo proceso que conlleve la aplicación del método bookmark, el mismo se construye considerando los ítems de la prueba que se analiza y puede contener reactivos en distinto formato, es decir, no se limita a reactivos dicotómicos o de selección múltiple, sin embargo, la ubicación de cada ítem debe responder a la dificultad asociada a cada uno de ellos, esto bajo la valoración de la dificultad del ítem según el modelo de Rasch, el cual asume que la probabilidad de que un examinado responda correctamente un ítem depende de la dificultad del ítem y de la habilidad del individuo.

En caso de incluir ítems que conlleven un proceso al que se asignará un puntaje distinto de 1, debe aparecer en el cuadernillo tantas veces como puntos se le pueda asignar al ser respondido de manera correcta. Esto quiere decir que si hay un ítem por cuyo desarrollo correcto se asignarán 4 puntos, debe aparecer 4 veces en el cuadernillo, dando a los jueces la posibilidad de señalar en qué punto del proceso de ese reactivo se debe tomar en cuenta el punto de corte.

Como parte de la preparación del cuadernillo de la prueba que se empleará en el proceso, se debe considerar la inclusión de un único ítem por página del documento, además de la recomendación de incluir en esa misma página el proceso requerido para responder correctamente el reactivo.

En el caso de contar con un amplio banco de ítems para la prueba, no es necesario centrar la atención sólo en los reactivos que formen parte de la prueba que se aplicó en una cierta población, sino que se puede incluir cualquier otro que sea equivalente en cuanto a contenido, tipo y por supuesto, dificultad. De manera semejante, en el folletillo se pueden incluir más ítems que los que tendría una prueba que vaya a ser aplicada a la población meta, esto permite que en caso de que existan brechas importantes en la dificultad de los ítems, se pueden complementar con otros de una dificultad que sea menor que el ítem de mayor dificultad y mayor que el de dificultad menor, esto permite que los jueces coloquen el marcador con mayor precisión.

Los autores Cizek y Bunche (2007), señalan que si bien el uso de un listado de ítems del banco permite fijar los puntos de corte, una de la ventajas de realizar el proceso con un folleto de prueba real es que la interpretación de los datos y las conclusiones es más sencilla.

Durante el desarrollo del proceso que conlleva la aplicación del método bookmark, la pregunta que debe orientar a los jueces debe ser: ¿Es probable que el examinado mínimamente calificado responda este ítem de manera correcta? Ahora bien, para unificar el concepto de "probable", se sugiere considerar que el 67% de la población responda el ítem adecuadamente o su equivalente, es decir, 2 de cada 3 examinados.

Retomando el papel del modelo de Rasch para la configuración del cuadernillo de la prueba, se destaca que la ecuación del modelo considera la probabilidad de acertar un ítem $p(x=1)$, con una dificultad β_j y una habilidad del individuo θ_i y se expresa como:

$$p(x=1|\theta_i,\beta_j)=\frac{e^{\theta_i-\beta_j}}{1+e^{\theta_i-\beta_j}}$$

Ahora, si para el método Bookmark se estima la probabilidad de que 2 de cada 3 examinados acierten el ítem, esto implica que $p(x=1|\theta_i,\beta_j)=\frac{2}{3}$, este valor se puede sustituir en la ecuación y despejarla en términos de θ_{i}:

$$\begin{gathered}\frac{2}{3}=\frac{e^{\theta_i-\beta_j}}{1+e^{\theta_i-\beta_j}}\Longrightarrow\\ \frac{2}{3}\left(1+e^{\theta_i-\beta_j}\right)=e^{\theta_i-\beta_j}\Longrightarrow\\ \frac{2}{3}+\frac{2}{3}e^{\theta_i-\beta_j}=e^{\theta_i-\beta_j}\Longrightarrow\\ \frac{2}{3}=e^{\theta_i-\beta_j}-\frac{2}{3}e^{\theta_i-\beta_j}\Longrightarrow\\ \frac{2}{3}=\frac{1}{3}e^{\theta_i-\beta_j}\Longrightarrow\\ 2=e^{\theta_i-\beta_j}\Longrightarrow\\ \ln 2=\theta_i-\beta_j\Longrightarrow\\ \theta_i=\beta_j+\ln 2\end{gathered}$$

Nótese que en el penúltimo paso se aplicó el logaritmo natural a ambos lados de la ecuación, mientras que en el último paso, se despeja la variable θ_i.

La fórmula obtenida permite expresar la capacidad o habilidad del examinado (según el modelo de Rasch) en función de la dificultad que se consideró para cada uno de los ítems y con ello ordenar los ítems de menor a mayor dificultad partiendo de la consideración de dicho modelo.

Se debe considerar que algunos otros trabajos como el de Buckendahl et al. (2002), aplican el método bookmark con algunas variaciones, por ejemplo para estimar el índice de dificultad no recurren al ajuste indicado por el modelo de Rasch y además, la fijación del punto de corte considera el puntaje obtenido hasta aquel punto donde se coloca el marcador en el cuadernillo.

2.3.1.2. Indicaciones para el equipo de jueces

Una vez construido el cuadernillo que se empleará para la aplicación del método y ya organizados los ítems según su nivel de dificultad, se procede a indicar a los jueces que su tarea consiste en indicar o marcar el punto en el que la probabilidad de que el sujeto límite acierte los ítems sea menor a $\frac{2}{3}$, es de esperar que al inicio del cuadernillo se considere que los ítems serán acertados por el sujeto límite, pues estos son los más sencillos, sin embargo, conforme se avanza en el folleto, está probabilidad caerá y en ese punto es donde el juez debe colocar la marca, ya sea con un marcador de páginas, una nota adhesiva u otro similar. Se recomienda que la exploración del cuadernillo por parte de los jueces se realice en pequeños grupos, pues esto alienta la discusión respecto a qué diferencia el nivel de cada uno de los reactivos de la prueba.

2.3.1.3. El cálculo de las puntuaciones

En los textos que abordan algunas experiencias en el uso del método bookmark, se señalan varias opciones para configurar el resultado final de las puntuaciones asociadas al punto de corte bajo el método bookmark, una de ellas corresponde al cálculo del promedio en el cual se fijó el límite de la probabilidad de acierto de $\frac{2}{3}$, sin embargo, esto puede variar en función de las consideraciones que se hayan tenido para fijar la dificultad y el nivel de habilidad requerido para la ejecución de la prueba.

2.3.1.4. Limitaciones del método

Entre las limitaciones que se señalan para la implementación del método, se destaca en documentos como el de Feseker et al. (2021), la subjetividad y la falta de capacitación de las personas que participan en el proceso de establecer los puntos de corte con este proceso, aspecto al cual se le debe prestar atención. Además, se puede considerar una limitación el partir del análisis de un banco de ítems, por ello es importante asegurar que los datos obtenidos para fijar la dificultad de cada ítem sean claros y consistentes.

Por otro lado, se señala también como una dificultad de este método la existencia de brechas amplias en la dificultad de los ítems, lo cual hace que los jueces duden al momento de colocar el marcador en un determinado punto y esto puede inducir a cierta incertidumbre en el proceso. Es por ello que en Cizek y Bunche (2007) se recomienda agregar algunos ítems del banco que completen esas brechas en caso de que existan.

Finalmente, Cizek y Bunche (2007) señalan que otra limitación es el caso de las pruebas muy fáciles o muy difíciles, pues la ubicación del ítem marcador es relativa

debido a la percepción de la probabilidad de acierto, esto hace que el grupo de jueces ubiquen el marcador en una posición errónea en comparación con la dificultad percibida de la prueba en general.

2.3.2. El método Angoff

El método Angoff se identifica como un método de fijación para los puntos de corte en el cual los evaluadores centran su atención en las preguntas de la prueba y no en los examinados, es decir, el evaluador se centra en tomar decisiones sobre los ítems que dan cuenta de la habilidad mínima aceptable por parte de los sujetos (Jalalizadeh et al., 2019).

En este método se considera la revisión del test que se aplicó a la población, para que los jueces procedan a determinar la probabilidad de que una persona, con habilidad mínima, logre responder de manera correcta una pregunta (Wyse, 2020). Coincide con el método bookmark en que esta primera fase es individual y hay una segunda en colectivo, en la cual, con los resultados obtenidos se realiza una revisión conjunta con el panel de expertos y la suma de los puntajes asignados para todas las preguntas corresponde al puntaje mínimo para “aprobar” la prueba (Angoff, 1971).

En raras ocasiones este método es empleado tal y como se propuso originalmente en la segunda edición del texto *Medición Educativa* de Thorndike del año 1971, en realidad, son más conocidas sus variaciones y es ampliamente utilizado en procesos de obtención de licencias y certificaciones que consideran pruebas con un formato de opción múltiple. Este método destaca por el número de investigaciones que se han hecho sobre él y por el uso frecuente que se le ha dado.

La implementación del método Angoff requiere la elaboración de un cuadernillo que contenga los ítems de la prueba, junto con un espacio para que cada juez estime el porcentaje de personas mínimamente competentes que responderían correctamente a cada ítem, y justifique su estimación. Posteriormente, se analizan estos valores para identificar tendencias entre los jueces y, a partir de ellas, se determina el punto de corte. A continuación, se detallará el procedimiento de implementación de este método, así como sus limitaciones.

2.3.2.1. El proceso

Originalmente, Angoff indicó que para cada ítem de la prueba, el juez asignará un 1 si considera que el "sujeto límite" (un integrante de la muestra ya descrita) responderá de manera correcta al reactivo y un 0 si considera que lo fallará, posteriormente, se calcula la suma de las puntuaciones de los ítems de la prueba y ese será el valor bruto obtenido por la persona mínimamente aceptable según los objetivos de la prueba.

Teniendo en cuenta lo anterior, la práctica estándar para la aplicación del método considera una alternativa que describe el mismo Angoff, esto es solicitar a cada juez que calcule la probabilidad de que el sujeto límite conteste cada ítem de manera correcta (las llamadas calificaciones de Angoff), ante esto, los jueces considerarán un grupo de sujetos límite en lugar de un sólo individuo y estimarán la proporción de examinados que responderá correctamente el ítem, posteriormente, la suma de esas probabilidades será la puntuación mínima aceptable.

En términos prácticos, se procede en primer lugar, a construir un cuadernillo de prueba que incluya los reactivos de la prueba para la que se busca fijar el punto de corte, seguidamente, se define una muestra de la población que realizará la prueba y para ello,

se consideran aquellos examinados que apenas aprueban o que están justo por encima del límite hipotético entre la destreza aceptable o inaceptable.

Considerando estos dos elementos, cada participante emite un juicio sobre cada uno de los elementos del test que se incluyen en el cuadernillo de la prueba y ofrece una estimación de la proporción de la muestra de examinados que responderían correctamente.

Habitualmente, se pide que consideren una muestra de 100 individuos mínimamente competentes y que indiquen de ese número, cuántos responderán de manera correcta el ítem, usualmente se solicitan múltiplos de 10. Posteriormente, se realiza la sumatoria de cada una de esas probabilidades para obtener la nota mínima aceptable para el test.

Una característica distintiva del método Angoff es la realización de múltiples rondas de evaluación, con el objetivo de lograr la convergencia de las estimaciones de los jueces. No obstante, se recomienda limitar el número de rondas a tres. Esta restricción se fundamenta en la prevención del agotamiento físico y mental que podría experimentar el panel de jueces al repetir el proceso de juzgamiento.

El objetivo de contar con más de una iteración del método es que las personas participantes del equipo de jueces puedan reconsiderar las probabilidades que asignaron en la fase anterior, así como la nota preliminar que se obtuvo al final de cada ronda, para ello, se les permite dialogar sobre los juicios que emitieron y a la vez se les brinda información adicional entre cada una de las rondas que permita una estimación de las probabilidades más precisa, aunque también se señala que la convergencia se puede

deber a que las personas juezas tomen conciencia del impacto de sus juicios sobre las tasas de aprobación de la prueba.

Para cada ronda o iteración del método Angoff, es esperable que se presenten variaciones entre las notas que asigna cada juez en cada ronda, es por ello que para obtener la nota de aprobación recomendada por el método Angoff, se procede a promediar las notas asignadas por los evaluadores a cada uno de los ítems y a la prueba en general y con los datos que se obtengan en la ronda final se concluye el valor del punto de corte. Para mayor comprensión, considere el siguiente ejemplo: para una prueba hipotética que consta de 5 ítems, se establece un equipo de 3 jueces y se desarrollan dos iteraciones o rondas del método:

Tabla 1

Ejemplo de rondas del método Angoff

ID del juez	Iteración del método	Ítem 1	Ítem 2	Ítem 3	Ítem 4	Ítem 5	Media de cada juez para la prueba
J1	Ronda 1	80	80	60	50	90	72
	Ronda 2	80	80	60	60	90	74
J2	Ronda 1	70	60	60	90	70	70
	Ronda 2	70	70	70	80	70	72
J3	Ronda 1	70	60	60	70	70	66
	Ronda 2	80	70	70	70	80	74
Media para cada ítem	Ronda 1	73,33	66,67	60	70	76,67	69,33
	Ronda 2	76,67	73,33	66,67	70	80	73,33

Nota: adaptado de Cizek y Bunche (2007).

Los valores indicados en cada columna (ítem 1, ítem 2,…) corresponden al número de personas mínimamente competentes de una población de 100 que los jueces consideran que responderán de manera correcta esa pregunta, mientras que en la primera columna se indica cuál juez emitió ese criterio. Por ejemplo, en la ronda 1 el juez 1 indicó que 80 personas de las 100 acertarán el ítem 1, y en la primera y segunda ronda,

el juez 3 indicó que el ítem 4 lo responderán de manera correcta 70 de las 100 personas. Las dos últimas filas indican las medias para cada una de las rondas. Y la última columna señala la media para cada juez en cada ronda.

Considerando las recomendaciones sobre el uso del método que brinda la literatura analizada y que sugieren centrar la atención en la segunda ronda de implementación del método por la reducción de la variabilidad en los valores asignados, para este ejemplo se obtendría un valor de 73,33% lo que equivale a 3,67 ítems correctos de los 5 disponibles.

2.3.2.2. Consideraciones para la implementación del método Angoff

Autores como Cizek y Bunch (2007), así como el propio Angoff (1971), sugieren que una implementación adecuada del método Angoff requiere una preparación o entrenamiento exhaustivo de los jueces participantes. En este proceso, se debe enfatizar la definición clara y precisa del concepto de "persona mínimamente competente", ya que una definición inadecuada puede introducir sesgos significativos en el resto del proceso.

Por otra parte, esos mismos autores recomiendan la aplicación del método Angoff para ítems dicotómicos, lo cual limita la posibilidad de implementarlo en otros tipos de prueba que requieran producción o desarrollo de ideas por parte de las personas examindas.

Finalmente, otra consideración al implementar este método, es en el caso de obtener puntajes que no sean enteros, pues en este caso, Cizek y Bunche (2007) señalan la posibilidad de redondear ese valor bajo criterios objetivos, mismos que deben considerar el impacto de ese redondeo en la fijación del punto de corte.

2.3.2.3. Limitaciones del método

Respecto a este método, se señala que una de sus limitaciones es el considerar cada ítem de manera aislada, pues esto puede afectar la percepción de la dificultad y obtener valores distintos al considerarlo por separado o incluido en las otras pruebas.

Tal y como se señaló previamente, una ventaja es el tener que analizar los ítems por separado por parte de cada juez, esto hace que su visión sea más objetiva y al final del proceso, el método ofrece que el cálculo de promedios permita homogenizar los resultados que indicaron cada uno de los jueces para los ítems, logrando un valor más consistente.

En general, ambos métodos analizados para establecer los puntos de corte, poseen la desventaja de que se centran en la prueba y que se requiere juzgar los ítems desde el punto de vista de un equipo de jueces, por ello, durante el proceso pueden existir algunos sesgos que se limitan por medio de la integración de más de un método para fijar el punto de corte, la consideración de un equipo de jueces amplio y especializado, la consideración de una ronda de consenso para el método bookmark y la implementación de dos rondas en el caso del método Angoff.

En general, ambos métodos analizados para establecer puntos de corte presentan la desventaja de centrarse en la prueba y de requerir la evaluación de los ítems por parte de un panel de jueces. Esto puede introducir sesgos, los cuales en esta investigación se buscó limitarlos mediante la integración de múltiples métodos para fijar el punto de corte, la selección de un panel de jueces amplio y especializado, la inclusión de una ronda de consenso en el método bookmark y la implementación de más de una ronda en el método Angoff.

Si bien este trabajo se centra en el uso de los métodos bookmark y Angoff, a continuación se presenta un breve apartado para describir otras aproximaciones para la fijación de puntos de corte.

2.3.3. Otros métodos

Existen algunos otros métodos para fijar los puntos de corte que obedecen a variaciones de los métodos Angoff y Bookmark, sin embargo, a continuación se presentan otras distintas.

En primer lugar, se tiene el contraste de grupos (Ebel y Frisbie, 1991) y consiste en la comparación de los resultados que logra un grupo de control contra los que obtuvieron quienes realizaron la prueba. Se persigue el identificar en qué punto se diferencia el rendimiento de cada grupo, es decir, cuándo el de control se ve superado por el nivel de destreza o competencia del grupo evaluado.

Por otra parte se tiene el método de la comparación con criterio externo (Kane, 2006), bajo esta metodología, existe la asignación de un puntaje mínimo para la aprobación de una prueba que depende del juicio de un ente externo. Para fijar esos puntos de corte, se consideran aspectos de los objetivos de la prueba, de la comparación con otros grupos e incluso, aspectos de leyes o estándares educativos y son estos los que indican el valor que se debe fijar.

Adicionalmente, en este listado de métodos de puntos de corte, se considerará al método de la regla de corte modificada (Livingston y Lewis, 1995), en este método, la

atención se centra en la opinión de los jueces respecto a la probabilidad de que un estudiante “límite” los acierte y se complementa ese primer dato con algunos aspectos adicionales como la correlación entre ítems y la consistencia interna de la prueba.

Si bien se han mencionada algunos otros métodos para fijar puntos de corte, la lista es amplia y en algunos casos se encuentran modificaciones de los mismo métodos que agregan o reducen la información que se considera para estimar los valores, sin embargo, en coherencia con los objetivos de esta investigación, la atención se centrará en el Bookmark y el Angoff.

2.4. La Prueba de Habilidades Cuantitativas

En apartados previos se detalló información sobre varias pruebas y sus distintas clasificaciones, sin embargo, en este punto se centrará la atención en la Prueba de Habilidades Cuantitativas y algunas de sus características.

La PHC es una prueba que se aplica a las personas estudiantes que aspiran a ingresar a las carreras de la Universidad de Costa Rica que poseen varios cursos de matemática en sus planes de estudio. La historia de la prueba inicia en 2003, año en el que las autoridades indicaron su necesidad, sin embargo, hasta el 2015 es cuando se empieza a aplicar de manera regular (Rojas y Ordóñez, 2019) como requisito de ingreso a ciertas carreras.

La Prueba de Habilidades Cuantitativas (PHC), desarrollada por un equipo del Instituto de Investigaciones Psicológicas de la Universidad de Costa Rica, tiene como objetivo proporcionar un instrumento que evalúe si los aspirantes a carreras con una carga significativa de cursos de matemática poseen las habilidades necesarias para cursarlos con éxito (Rojas et al., 2019).

Respecto a las habilidades que se requieren para enfrentar algunos de los cursos de matemática, se considera el constructo de la PHC que se detalla a continuación.

2.4.1. Constructo de la PHC

Para la Prueba de Habilidades Cuantitativas, el constructo de interés es el razonamiento cuantitativo (RC) y cómo se adelantó en el apartado anterior, el interés de la misma es brindar un referente para elegir a las personas que quieran ingresar a las carreras que se caracterizan por tener un alto contenido matemático en su plan de estudios (Rojas y Ordóñez, 2019).

La definición del constructo en cuestión ha tenido varios acercamientos, sin embargo, en uno de los trabajos más recientes que se han recopilado para esta investigación, las personas autoras indican que se le puede comprender como la capacidad que tienen los individuos de comprender y utilizar argumentos cuantitativos en varios contextos (Ordóñez y Rojas, 2024).

Al respecto, esa concepción del razonamiento cuantitativo indica que no es únicamente el conocer sobre distintos contenidos de matemática, sino que se requiere el comprenderlos para poderlos aplicar en la solución de distintas situaciones. Por ello, no basta con saber mucho de matemática, sino que se requiere saber cómo aplicar esos contenidos a situaciones conocidas o nuevas.

Los autores Mora y Rojas (2023b) subrayan que el razonamiento cuantitativo considera los procesos de análisis que se vinculan con información cuantitativa, por ejemplo, su organización o las deducciones que se pueden realizar a partir de la información con la que se cuenta.

Por ello, autores como Ryan y Gass (2017) y Mayes (2019), sostienen que un buen nivel de RC se relaciona con que los especialistas de las áreas que requiere de las matemáticas en su ejercicio profesional, puedan utilizar la matemática para resolver un problema de manera exitosa, precisamente el objetivo planteado inicialmente para la PHC.

Conociendo cuál es el propósito y el contenido de la PHC, seguidamente se detallará cuál es la estructura que posee dicha prueba.

2.4.2. Estructura de la PHC

Tal y como se mencionó en el apartado de antecedentes, para configurar la PHC se tomó en cuenta cuáles eran los temas del área de matemática que se estudian en la educación secundaria en Costa Rica y a partir de esa revisión, las áreas consideradas fueron geometría, aritmética, álgebra y análisis de datos, cada una con 10 ítems que las representan 40 reactivos en total (Rojas y Ordóñez, 2019).

Al considerar que la PHC posee implicaciones en la selección de las personas que ingresarán a una carrera determinada, es necesario abordar los temas de validez y confiabilidad, dicho aspecto se presenta en el siguiente apartado.

2.4.3. Validez y confiabilidad del uso de la PHC

La validez, como principio fundamental en la evaluación, se define en este estudio como el grado en que la teoría y la evidencia empírica sustentan las inferencias derivadas de la aplicación de la Prueba de Habilidades Cuantitativas (PHC). Esta definición se basa en los estándares establecidos por la American Educational Research Association, American Psychological Association y National Council on Measurement in Education

(2018). Es importante reconocer que el concepto de validez ha evolucionado a lo largo del tiempo, con contribuciones significativas de autores como Messick (1989) y estudios recientes como los de Sánchez et al. (2020), Medina y Verdejo (2020) y Taipe (2021).

La confiabilidad, esencial para la validez de cualquier evaluación, se define en este estudio como la consistencia y precisión con la que la PHC mide las habilidades cuantitativas en una población específica bajo condiciones determinadas. Esta definición se apoya en investigaciones recientes, como las de Medina y Verdejo (2020) y Taipe (2021). En términos prácticos, la confiabilidad se manifiesta en la similitud de los resultados obtenidos al aplicar la PHC en diferentes momentos a la misma población.

La literatura coincide en que la confiabilidad de una prueba se evalúa mediante el coeficiente alfa de Cronbach, cuyos valores permiten clasificar la consistencia interna en distintos niveles. Un Alfa de Cronbach igual o superior a 0,9 se considera excelente, indicando una alta homogeneidad entre los ítems. Valores entre 0,8 y 0,9 sugieren una buena confiabilidad, mientras que valores entre 0,7 y 0,8 se consideran aceptables para muchas investigaciones. Sin embargo, es importante ser cauteloso con valores entre 0,6 y 0,7, ya que se consideran cuestionables y podrían indicar la necesidad de revisar o mejorar el instrumento. Valores entre 0,5 y 0,6 se consideran de pobre confiabilidad, y aquellos inferiores a 0,5 son inaceptables, lo que supone deficiencias en la consistencia interna de la prueba. Es importante recordar que la interpretación del valor del Alfa de Cronbach debe realizarse en el contexto específico de la investigación, considerando factores como el número de reactivos, las dimensiones del constructo y el propósito de la prueba.

Los conceptos de validez y confiabilidad en el campo de la investigación afectan los procesos de medición y evaluación, por ello son temas centrales para la construcción del conocimiento científico y para la toma de decisiones y el ajuste de las propuestas de intervención (Jornet et al., 2020). Además, para una prueba que tiene impacto en el ingreso a una institución, su proceso debería estar libre de sesgos y contar con jueces objetivos que analicen la información de manera integral, equitativa y ponderada, para así seleccionar a los “mejores” candidatos (M. Sánchez et al., 2020).

En cuanto a algunos de los parámetros para la validez, se señala que una de las posibles técnicas es la realización de un Análisis de Varianza (ANOVA) para detectar diferencias entre los resultados que obtiene la población en la prueba, pero también existen otras rutas para poder establecer dichas evidencias, por ejemplo el análisis factorial confirmatorio, mismo que evalúa si un modelo teórico establecido previamente y basado en una serie de factores, explica de manera adecuada las relaciones entre un conjunto de variables observadas (Martínez, 2021).

En cuanto a la fiabilidad de la prueba, Rojas y Ordoñez (2019) y Rojas, Mora y Ordoñez (2019) señalan los valores que ha obtenido la prueba en el Alpha de Cronbach para los años 2015, 2016 y 2018, mismos que respectivamente equivalen a 0,87, 0,86 y 0,85 respectivamente. Según los parámetros de esa prueba, estos valores indican que la confiabilidad del test es buena.

Por su parte, para asegurar la validez de las conclusiones realizadas a partir de la prueba, las personas autoras indican que se han basado en el análisis de la estructura interna de la prueba con el apoyo de un Análisis Factorial Confirmatorio, mismo que debería asegurar que los ítems carguen en un único factor para asegurar la

unidimensionalidad de la prueba. En cuanto a las cargas factoriales, los ítems han obtenido resultados superiores a 0,30, valor que se ha fijado como deseable para definir dichas cargas factoriales.

Por ello, se tiene un punto de partida claro para asegurar la confiabilidad y la validez de la prueba que se analizará en esta propuesta de investigación.

En síntesis, este estudio se fundamenta en la evaluación educativa comprendida como un proceso integral, sistémico y continuo, que va más allá de la clasificación de estudiantes y se enfoca en la mejora de los procesos formativos. La Prueba de Habilidades Cuantitativas (PHC), como prueba estandarizada y referida a criterio, se analiza bajo el prisma de la Teoría de Respuesta al Ítem (TRI) y el modelo de Rasch, lo que permite una medición objetiva de habilidades latentes para el constructo del razonamiento cuantitativo. La aplicación de los métodos Bookmark y Angoff para establecer puntos de corte se sustenta en la organización de ítems por dificultad y la evaluación de jueces expertos, buscando definir niveles de desempeño claros y objetivos. Si bien estos métodos, presentan limitaciones como la subjetividad y la dependencia del juicio de expertos, se complementan y fortalecen al considerar múltiples rondas y la integración de diferentes enfoques. La interpretación de los resultados de la PHC, basada en estos fundamentos teóricos y psicométricos, como la validez y confiabilidad demostradas a través del análisis factorial confirmatorio y el Alfa de Cronbach, permite identificar niveles de habilidad en razonamiento cuantitativo y tomar decisiones informadas sobre los procesos de admisión para la Universidad de Costa Rica.

En el capítulo siguiente, se aborda el proceso metodológico que se empleó para el desarrollo de esta investigación.

3. Marco metodológico

Este capítulo detalla la metodología empleada en la presente investigación, enfocándose en el proceso metodológico implementado. Se aborda el diseño de investigación utilizado, la población seleccionada, los elementos analizados, y las técnicas e instrumentos empleados para el establecimiento de los puntos de corte de la PHC

Considerando las especificidades del estudio y sus alcances, se optó por proponer una investigación de tipo cuantitativo bajo un diseño instrumental, mismo que consideró como población de estudio a algunas personas expertas de la Escuela de Matemática y del Instituto de Investigaciones Psicológicas, ambos de la Universidad de Costa Rica.

3.1. Tipo de investigación

El objetivo de esta investigación fue establecer los puntos de corte para la PHC, esto a partir de la implementación de dos de los métodos usuales para tal fin, como lo son los métodos bookmark y el Angoff.

Como se indicó al inicio de esta capítulo, el tipo de investigación que se propuso fue el cuantitativo, esto debido a que la investigación se centra en la fijación de valores que reflejen la dificultad que posee una serie de ítems que componen un banco para una prueba, esto para dilucidar el nivel mínimo de conocimientos que debe tener una persona estudiante al ingresar a una carrera que aplica la PHC. Es decir, se pretende fijar una nota de aprobación para una prueba desde la medición de una serie de parámetros indicados por los análisis realizados a los ítems que conforman la prueba y desde el punto de vista de algunas personas cercanas a un proceso de evaluación como lo es la PHC.

Los estudios cuantitativos permiten analizar datos para describir, predecir y explicar su ocurrencia y consecuencias (F. A. Sánchez, 2019). Esta investigación determina el nivel de complejidad de cada ítem de la prueba a partir de la percepción de dificultad de jueces expertos y del análisis de los ítems que la componen.

A su vez, una investigación cuantitativa ofrece la oportunidad de generalizar los resultados obtenidos. Esto resulta de interés en esta investigación debido a que ofrece la posibilidad de contar con una predicción del desempeño académico que tendrá una persona al inicio de su vida universitaria según el resultado que obtenga en la PHC.

Algunos autores como Galeano (2021) y Páramo et al. (2020), resaltan que las investigaciones que recurren a valores numéricos, es decir, los valores obtenidos a partir de medidas; se pueden considerar como un medio para el análisis de la información, misma que puede obtenerse desde distintas fuentes. En el marco de este documento, esto se puede ver reflejado cuando se aplican los métodos bookmark y Angoff, los cuales parten de la jerarquización de los ítems según su nivel de dificultad para establecer una nota mínima de aprobación en la PHC.

Es por ello que en coherencia con el estudio, esta investigación cuantitativa permitió obtener los valores para los puntos de corte respectivos para las personas que realizan la PHC y a su vez, facilitó una serie de espacios para que el equipo de jueces pudiera realimentar, validar y complementar el proceso que se siguió para el logro de los objetivos propuestos, esto debido a que se les consideró como los indicados para dotar de validez a los resultados obtenidos a partir de su experiencia en el tema, así como de su especialidad en el área de interés.

3.2. Diseño de la investigación

El diseño del estudio corresponde a una investigación instrumental, la cual según autores como Losada et al. (2022) y Contreras-Cazarez y Campa-Álvarez (2022) se puede definir como aquella que estudia el desarrollo de pruebas e instrumentos y que para ello considera tanto su diseño como su adaptación.

En el caso particular de esta investigación, se analizó el instrumento de la PHC con el objetivo de establecer sus puntos de corte. Para lograr este objetivo, se consideraron los ítems de la prueba, sus análisis de dificultad y el criterio de un equipo de jueces. Este equipo, fijó el punto de corte por medio de la implementación de los métodos Angoff y bookmark.

3.3. Población y muestra

En el contexto de esta investigación, la población a considerar se dividió en dos grupos. En primer lugar, la cohorte de personas estudiantes que presentó al PHC en el año 2021, de esa población se utilizarán los resultados obtenidos en las pruebas para obtener el índice de dificultad de los ítems que se requiere previo a la implementación de los métodos Angoff y bookmark. En segundo lugar, se consideró un equipo de 6 docentes de la Escuela de Matemática de la Universidad de Costa Rica que han impartido el curso de Cálculo I en los últimos 5 años. Por la formación y experiencia que este grupo de profesores tienen en el área, se constituyen en fuentes de información relevantes para el objetivo de este trabajo.

La selección de las personas participantes fue de manera intencional y se consideró como criterios de inclusión que laboraran para la Escuela de Matemática de la Universidad de Costa Rica, que fueran docentes de Matemática, que hubieran impartido

el curso de Cálculo I en los últimos 5 años y que además tuvieran disponibilidad para asistir a las distintas rondas de trabajo que suponía el desarrollo del proyecto. Como criterio de exclusión, se tomó en cuenta el faltar de manera injustificada a alguna de las sesiones de trabajo para la implementación de los métodos.

Se optó por un muestreo intencional, un método que implica la selección deliberada de unidades de análisis. Esta elección se fundamentó en la necesidad de incluir participantes que, por su conocimiento o experiencia, aportaran información relevante y coherente con el objeto de investigación (Mena, 2018).

Tomando como base las experiencias previas en la implementación de los métodos para fijar puntos de corte, al equipo de jueces se les dividirá de manera aleatoria en 2 equipos de 3 personas para poder intercambiar el orden en el que aplicarán los métodos cada equipo, es decir, el primer equipo trabajará primero con el método Bookmark y luego con el Angoff, mientras que el segundo grupo lo hará en orden inverso, esto tiene como objetivo aumentar la objetividad en los resultados y evitar caer en un sesgo por implementar un mismo orden en los métodos para todos los jueces.

Para designar a las personas que conformarán cada equipo de jueces, se optará por una elección aleatoria de cada integrante para los equipos. En esa línea, según Bolsover (2018) y Bracho (2022), la elección de informantes de manera aleatoria, consiste en asociar a cada elemento de la muestra con un conjunto númerico, mismo en el que cada representante tiene igual probabilidad de ser seleccionado. Lo anterior es significativo para conformar los equipos de jueces manteniendo un proceso objetivo para elegir qué persona integra cada equipo y así asegurar su homogeneidad.

La fijación del tamaño de la muestra se centró en la necesidad de realizar un manejo adecuado de la información y en ese sentido, con el tamaño de muestra seleccionado, cada persona aporta más información y se puede cotejar con mayor facilidad la información que brinde en cada momento, lo que facilita el análisis de los datos (Saiz et al., 2019).

3.4. Variables para el estudio

Dado el diseño de investigación instrumental propuesto, fue necesario definir los factores asociados al desarrollo del estudio, así como sus relaciones. Las variables consideradas fueron la dificultad de los ítems, los procesos de razonamiento implicados en cada ítem, el contenido de los ítems y los niveles de desempeño para la prueba. A continuación, se describe cada una de estas variables, así como el objetivo de tomarlas en cuenta.

3.4.1. Dificultad de los ítems

Para el establecimiento de los puntos de corte de una prueba, es necesario jerarquizar los ítems a partir de las respuestas que se hayan obtenido en una aplicación previa de la prueba (Feseker et al., 2021). Para ello, se utilizó la Teoría de Respuesta al Ítem y en específico a la variable independiente del índice de dificultad del modelo de Rasch, misma que se puede comprender como la variabilidad presente de que cada individuo responda correctamente un determinado ítem y desde el punto de vista práctico varía en una escala de -5 a 5, donde el 0 representa la dificultad media del ítem para la población (Jiménez y Montero, 2013). En otras palabras, es el nivel de habilidad necesario para que el examinado tenga una probabilidad del 50% de contestar correctamente al ítem y se suele representar con la letra *b*.

Para obtener este valor, se analizó la base de datos de la PHC actualizada al período del año 2021 de donde se obtuvo el índice de interés.

3.4.2. Procesos de razonamiento del ítem

Como ya se señaló, la PHC se encarga de medir el constructo del razonamiento cuantitativo. Por ello, al momento de establecer los puntos de corte de la prueba, es importante incluir en los cuadernillos del proceso los pasos que debe desarrollar la persona estudiante para resolver adecuadamente un ítem.

El procedimiento que se siga para resolver un ítem se vincula con la ejecución de un razonamiento, mismo que se puede comprender como un ejercicio de pensamiento que permite obtener conclusiones según una serie de premisas establecidas con anterioridad (Salgado y Salinas, 2012).

3.4.3. Contenido de los ítems

En la PHC se considera una serie de elementos que se deben conocer previamente para enfrentar las situaciones de razonamiento que se propone, esos temas que dan fondo a la prueba se conocen como los contenidos y se les puede definir como aquellos conocimientos que se requieren para generar destrezas específicas en las personas estudiantes (Garduno, 2009). Los contenidos de la PHC serán de interés en la construcción de los cuadernillos para la implementación de los métodos de punto de corte, pues los mismos sirvieron de guía a los jueces respecto a qué dificultades puede enfrentar cada individuo al responder cada ítem desde el punto de vista de los contenidos que requiere para responder cada ítem de manera correcta.

3.4.4. Niveles de desempeño

El objetivo principal de este trabajo es el establecer puntos de corte en función de las habilidades para el razonamiento cuantitativo que posee una persona, es decir, el nivel de desempeño que se considera que posee para enfrentar y resolver distintas situaciones en un contexto que requiera habilidades cuantitativas. Es por ello que resulta de interés referirse a la variable independiente de los niveles de desempeño, mismos que se relacionan pero no limitan al rendimiento académico de las personas estudiantes. Sin embargo, para los efectos de este estudio se consideraron como el resultado que obtiene una persona estudiante en la prueba y que son reflejo de su conocimiento y aptitudes para determinada temática (J. Contreras et al., 2019).

Es oportuno aclarar que aunque esta variable se considera en este apartado, la misma forma parte de los resultados de esta investigación y no necesariamente de los datos de partida para el desarrollo del proceso de establecimiento de los puntos de corte.

La función de los niveles de desempeño en educación es el describir “lo que saben y saben hacer los estudiantes en función del currículo, las metas y resultados de aprendizaje; están referidos al grado de procesamiento intelectual o a la demanda cognoscitiva involucrada en la resolución de problemas” (Cano et al., 2021, p. 113). Es decir, con esa variables se relacionan las capacidades de una persona con un estándar determinado a fin de poder clasificarle en cierto nivel respecto a aquella población con la que se le compara o la que pertenece.

En congruencia con lo anterior, Flores-Lueg et al. (2018) señalan que los niveles de desempeño de una persona y en el caso específico de la matemática, consideran también la resolución creativa de problemas varios, así como el encontrar soluciones a

las necesidades específicas de cualquier campo disciplinar, aspecto vital en la creación de laprueba de Habilidades Cuantitativas, pues como lo indica Ordóñez-Gutiérrez (2023) la prueba no se centra en el saber, sino, en la aplicación de los conocimientos para resolver distintas situaciones en un contexto de aplicaciones en distintas disciplinas.

Como paso previo a la implementación de los métodos para establecer puntos de corte, fue necesario fijar el número de niveles de desempeño deseados, por ello, se consultó a los jueces involucrados en el proceso si según su criterio bastaba con un nivel de desempeño o si se hace necesario contar con más de ellos. Era de esperar que los niveles de desempeño partieran de la división por la nota que actualmente se considera para recomendar el ingreso de las personas estudiantes a las carreras que aplican la PHC y que consideran en sus planes de estudio el curso Cálculo I (MA1001).

En el marco de esta investigación, se propone la creación de un punto de corte que permita clasificar la población que realiza la PHC (Prueba de Habilidades Cuantitativas) en dos grupos o niveles de desempeño para cada una de las áreas consideradas en la PHC:

- Insuficiente: subcategoría que comprende a las personas estudiantes que obtienen una nota menor que el punto de corte.
- Competente: población que alcanza una nota igual o superior al punto de corte fijado para la PHC.

El aporte de esta categoría a la investigación es que brindó información para establecer el punto de partida respecto a los niveles de desempeño que se esperaría obtener para la población. Esto se realizó a partir de los puntos de corte establecidos con la implementación de los métodos bookmark y Angoff. Una vez fijado el número de

niveles, será posible orientar a las personas juezas en cuanto al número esperado de ítems que indican los puntos de corte para el método Bookmark.

3.5. Instrumentos para la recolección de datos

Los instrumentos que se consideraron para el desarrollo de la investigación fueron el cuadernillo de la Prueba de Habilidades Cuantitativas utilizado en el 2021 y los cuadernillos de trabajo para el análisis de cada uno de los métodos, cada uno de ellos se describe seguidamente.

3.5.1. Prueba de Habilidades Cuantitativas del 2021

El primer instrumento para recolectar la información necesaria para la investigación fue uno de los cuadernillos de la Prueba de Habilidades Cuantitativas aplicada en el año 2021. Para acceder al cuadernillo de la prueba, y a los resultados del análisis de cada uno de sus reactivos, se solicitó el acceso a los mismos al equipo del Instituto de Investigaciones Psicológicas que se encarga del desarrollo de la prueba.

Como parte de este punto de la recolección de información, fue de interés conocer algunos aspectos técnicos de la prueba, por ejemplo el análisis psicométrico de sus ítems, esto con el propósito de conocer los ítems del banco y principalmente sus niveles de dificultad según el modelo de Rasch.

Para la organización de la información se elaboró una ficha o plantilla de análisis para cada uno de ellos, misma que contenía un identificador para cada ítem, así como un espacio para anotar los datos relevantes de cada reactivo a analizar, entre ellos, el área a la que pertenece (aritmética, geometría, álgebra o análisis de datos), su índice de dificultad y los procesos que implica resolver cada ejercicio de manera correcta.

Es oportuno aclarar que para la aplicación de los métodos es necesario conocer el índice de dificultad de los reactivos, por ello aunque el cuadernillo de la PHC se conforma por 40 ítems en total, para los objetivos de esta investigación se descartaron 4 de los mismos, uno por cada sección de la prueba. El motivo para no considerarlos obedece a que los mismos corresponden a reactivos experimentales que al momento de la aplicación de la prueba no formaban parte del banco de ítems de la misma.

Tomando en cuenta lo anterior, el análisis que se realizó corresponde a los 36 ítems restantes que ya formaban parte del banco de preguntas de la PHC y que contaban con datos previos sobre su índice de dificultad según la Teoría de Respuesta al Ítem.

3.5.2. Cuadernillo de trabajo de Análisis para cada método

Para la implementación de los métodos bookmark y Angoff, se diseñaron dos cuadernillos específicos con el fin de facilitar el proceso de evaluación por parte de los jueces expertos. El primero de los folletos integraba ambos métodos y el segundo incluía solamente el método Angoff, permitiendo así una aplicación flexible y adaptada a las necesidades de cada fase propuesta para el estudio.

La estructura de los cuadernillos se organizó en torno a las dimensiones fundamentales del razonamiento cuantitativo evaluadas por la PHC y representadas en esa edición de la prueba: validar, clasificar, relacionar y generalizar. Cada cuadernillo iniciaba con una descripción detallada de los procesos de razonamiento y contenidos evaluados, proporcionando un marco de referencia de fácil consulta para los jueces.

Los ítems, ordenados por nivel de dificultad según el método Rasch, se presentaron individualmente en páginas sucesivas, acompañados de posibles procedimientos de resolución paso a paso. Esta disposición buscaba asegurar que los

jueces pudieran evaluar cada ítem de manera detallada. Los valores de la dificultad de cada ítem se obtuvieron por medio de la consulta del análisis de los reactivos que fue proporcionado por el equipo encargado de la prueba.

En el cuadernillo que integraba el método bookmark, se incluyó un espacio al final de cada página para que los jueces registraran sus observaciones, indicaran si consideraban el ítem como el marcador de los niveles y para que justificaran sus decisiones en términos de los procesos de razonamiento implicados. Este diseño permitía recopilar todos los datos para el análisis de los puntos de corte.

Para el método Angoff, tanto en el cuadernillo combinado como en el exclusivo, se incluyó una pregunta estandarizada para cada ítem: "¿cuántas personas examinadas de un grupo de 100 con un nivel aceptable acertarán el ítem?". Esta pregunta se repitió en cada ronda de aplicación del método Angoff, asegurando la consistencia en la valoración de los jueces.

La construcción de estos cuadernillos se realizó con el objetivo de proporcionar a los jueces una herramienta clara y organizada, que facilitara la aplicación de los métodos y permitiera la recopilación de datos precisos y relevantes para el establecimiento de los puntos de corte de la PHC.

3.6. Procedimiento

Para el análisis de la información necesaria para la investigación, se propone el desarrollo de 5 etapas, mismas que se describen a continuación.

3.6.1. Construcción del perfil de la persona mínimamente competente para resolver la PHC

Como paso inicial en la investigación, se procedió a construir un perfil detallado de la persona estudiante mínimamente competente en la PHC. Este perfil tenía como objetivo definir las habilidades básicas que el examinado debe demostrar para superar la prueba, proporcionando así un marco de referencia para el equipo de jueces.

La creación de este perfil se basó en la revisión de la documentación existente en el proyecto de la PHC, lo que permitió identificar los procesos de razonamiento cuantitativo fundamentales que se esperan de un estudiante competente. Este documento inicial se sometió a la revisión de expertos que colaboran en la elaboración de la PHC, ellos aportaron su conocimiento y experiencia para refinar y validar el perfil propuesto.

La versión revisada del perfil del estudiante mínimamente competente se presentó posteriormente al equipo de jueces y sirvió como una guía esencial durante el proceso de evaluación, asegurando que todos los jueces compartieran una comprensión común de las habilidades esperadas y que sus juicios se basaran en criterios consistentes.

La implementación de este perfil favoreció que los jueces se centraran en determinar la probabilidad de éxito de un estudiante mínimamente competente en cada ítem, minimizando la influencia de interpretaciones individuales en la tarea que se les propuso.

3.6.2. Conformación del equipo de jueces

La conformación del equipo de jueces inició con la solicitud de referencias a los coordinadores de la cátedra del curso de Cálculo I en la Escuela de Matemática de la

Universidad de Costa Rica, así como a otros contactos clave de la Unidad Académica. Se buscaba identificar a candidatos con experiencia y conocimiento en el área de razonamiento cuantitativo, que cumplieran con los criterios de inclusión establecidos para esta investigación y con disponibilidad para vincularse en las tareas del proyecto.

A partir de las recomendaciones recibidas, se estableció contacto individual con cada candidato potencial. El objetivo de estas comunicaciones fue el verificar la disponibilidad de los candidatos para participar en el estudio y a la vez validar que cumplieran con los criterios de inclusión previamente definidos.

Este proceso de selección garantizó la conformación de un equipo de 6 jueces altamente calificados y con el compromiso de desarrollar las tareas propuestas en el marco de la investigación.

Posteriormente a la conformación del equipo, se procedió a citarlos para informarles sobre el proceso en general y realizar una sesión de entrenamiento, esta fase de detalla a continuación.

3.6.3. Entrenamiento del equipo de jueces

Al tomar en consideración la posible novedad de los métodos para fijar puntos de corte para algunos de los jueces, se diseñó una sesión de entrenamiento introductoria. El objetivo principal de esta sesión fue clarificar el rol de los jueces y proporcionarles las herramientas necesarias para cumplir con su tarea de manera efectiva.

La sesión de entrenamiento se centró en la aplicación de los métodos Bookmark y Angoff utilizando ítems del área de álgebra del folleto de práctica de la Prueba de Habilidades Cuantitativas (PHC), disponible para acceso público. Esta sección de la prueba, compuesta por 10 ítems, sirvió para pilotear el proceso de fijación de puntos de

corte y facilitó la identificación de posibles ajustes y detalles que requerirían atención especial durante las rondas definitivas del proceso.

Un aspecto fundamental del entrenamiento fue asegurar que los jueces priorizaran el proceso de razonamiento requerido por cada ítem sobre el contenido específico del mismo. También se hizo énfasis en la importancia de considerar el perfil del estudiante mínimamente competente definido previamente. Además, se clarificó el concepto de "sujeto límite" según cada método, para evitar interpretaciones erróneas en la implementación de los resultados.

Finalmente, se abordó el tema de la confidencialidad. Debido al carácter sensible de los recursos y el banco de ítems de la PHC, se requirió que todos los jueces firmaran un acuerdo de confidencialidad. Este documento, proporcionado por el Instituto de Investigaciones Psicológicas de la Universidad de Costa Rica, garantiza la protección de la información y establece los procedimientos a seguir en caso de incumplimiento.

3.6.4. Sesiones de punto de corte

Una vez que se concluyó con la sesión de entrenamiento con el equipo de jueces así como con la contrucción de los cuadernillos de trabajo para cada uno de los métodos de punto de corte, se procedió a desarrollar el trabajo definitivo con las personas del jurado. Para esta fase, se dividió el equipo de jueces en dos grupos de tres personas, mismas que participaron de manera diferenciada en la implementación de cada método, en el caso del grupo A, su primera ronda incluyó el método Angoff y el bookmark y en la segunda iteración, solamente el cuadernillo del método Angoff. El grupo B procedió en orden inverso, es decir, comenzaron con el cuadernillo del método Angoff y para la segunda sesión se enfocaron en el folleto que incluye ambos métodos.

Una vez desarrolladas las sesiones indicadas, se convocó al grupo de jueces para una sesión grupal, misma en la que se consensuó el ítem bookmark preliminar a partir de lo que se indicó en las dos sesiones previas del método y adicionalmente, se definió con el grupo de jueces, las características del nivel competente a partir de los contenidos y procesos de razonamiento comprendidos en la PHC.

3.6.5. Análisis de los datos

En esta sección se busca resumir el proceso de análisis de los datos que se consideró para lograr los objetivos propuestos para la investigación.

En primer lugar, considerando que la literatura señala ciertos elementos que inducen variabilidad y subjetividad en la asignación de los puntos de corte bajo los métodos considerados en este estudio y que el objetivo es obtener valores que fijen el punto de corte para la prueba, es importante proponer una etapa en la que se analicen los resultados obtenidos en cada fase implementada para definir un único valor para cada método como resultado de este estudio, para ello, se proponen cinco niveles de análisis que se detallan a continuación.

3.6.5.1. Cálculo de los puntos de corte con cada método

Al tener presente que esta investigación aborda la implementación de dos métodos para fijar los puntos de corte, es necesario diferenciar dos procedimientos para poder obtener un valor final luego del desarrollo de las fases que implica cada uno de los métodos empleados.

En primer lugar, en el caso del método Angoff y tomando como referente los insumos que brinda la teoría para la aplicación adecuada de dicho método, es necesario considerar los valores que cada uno de los miembros del equipo de jueces consignó en

cada cuadernillo empleado en las rondas desarrolladas. Según la teoría analizada, sería esperable que con cada iteración del método, se reduzca la variabilidad entre los valores que asignan los jueces y por ello los resultados se deberían orientar a un valor único o al menos, más cercano; de ahí la importancia de desarrollar al menos dos rondas de este método. Una vez que se halla realizado el trabajo con las personas que integran el equipo de jueces, se procederá a resumir los datos obtenidos de los cuadernillos por medio de una base de datos, en la que se incluirá el valor de la probabilidad de acierto que asigan los jueces en cada ronda para cada ítem y las justificaciones que brindan para dicho juicio. A partir de la información recolectada, se construirá una tabla resumen para buscar tendencias entre los datos obtenidos, por ejemplo se prestará atención a los valores extremos de la probabilidad de acierto de los ítems, así como a la tendencia en la nota de aprobación para cada una de las dimensiones que incorpora la prueba. Finalmente, para obtener la nota de corte según este método, se considerará la mediana de los valores indicados por los juecas para la probabilidad de acierto de cada uno de los ítems, luego, a partir de esos valores se calculará su promedio y esto dará como resultado dos puntos de corte, uno para cada ronda de aplicación del método. Es oportuno aclarar que según el sustento teórico del método, la atención se debe centrar en los valores obtenidos en la segunda ronda del mismo.

Para el caso del método bookmark, el cálculo de la nota de corte demanda dos etapas: la determinación del ítem bookmark para cada una de las dimensiones de la prueba, así como la obtención de la nota de corte a partir de estos ítems.

En la primera fase, las personas juezas deberán determinar el reactivo que consideran el ítem bookmark en cada proceso de la prueba, una vez que todo el equipo

de jueces haya emitido su criterio en los cuadernillos correspondientes, se procederá a resumir los datos del ítem que cada juez indicó como el marcador o el bookmark para cada una de las secciones, así como la justificación que brindaron para esa elección. Con los datos que se obtienen para cada dimensión por parte de cada juez, se calculará la mediana de las posiciones de los ítems que indicaron, considerando este el valor representativo para fijar el punto de corte según este método.

Es importante agregar que en el caso de este método es importante buscar el consenso entre el equipo de jueces, por ello, una vez calculado el valor preliminar de ese punto de corte, se validará con el equipo de jueces para intentar asegurar la fiabilidad del mismo y que represente la opinión del colectivo participante en el proceso.

En el desarrollo de la segunda etapa del cálculo del punto de corte según este método, se deberá definir un ítem bookmark global para la prueba, con este objetivo, se tomará el ítem con la dificultad Rasch mínima de los ítems bookmark que indicaron los jueces según las dimensiones de la prueba. Se tomará en cuenta ese ítem, pues según la definición del ítem bookmark, todos los reactivos que tengan una dificultad menor a este, deberían tener una probabilidad de acierto superior o igual a $\frac{2}{3}$ para la población de interés.

Considerando el reactivo que se haya identificado como el ítem bookmark global, se procederá a identificar la dificultad β de su ítem predecedor en dificultad y con ella, se estimará la habilidad requerida de un sujeto para tener una probabilidad de acierto de $\frac{2}{3}$ en este ítem.

Para el cálculo de dicha habilidad según el modelo de Rasch, se aplicará la siguiente fórmula:

$$p(x = 1|\theta_i, \beta_j) = \frac{2}{3} = \frac{e^{\theta_i - \beta_j}}{1 + e^{\theta_i - \beta_j}} \Longrightarrow \theta_i = \ln 2 + \beta_j$$

La expresión anterior indica que para que la persona que realiza la PHC tenga una probabilidad de acierto de $\frac{2}{3}$ en un ítem, la habilidad (θ) requerida se obtiene al realizar la suma de la dificultad del ítem (β), más $\ln(2) \approx 0{,}693$. El proceso detallado para obtener esta fórmula se realizó en el apartado 2.3.1.1.

Finalmente, la nota en escala de 0 a 100 será obtenida por medio de la tabla de conversión de habilidad Rasch al número de preguntas correctas en el test.

3.6.5.2. Contraste entre los resultados de los puntos de corte

Considerando que en esta investigación se optó por integrar dos métodos distintos para establecer un punto de corte para una prueba estandarizada, es de interés comparar los resultados que se obtienen con cada uno de ellos en búsqueda de similitudes y diferencias que enriquezcan el proceso desarrollado. Para ello, se analizarán los criterios emitidos por el equipo de jueces según cada dimensión de la prueba en búsqueda de elementos que se deban destacar.

3.6.5.3. Definición de los niveles de desempeño

Como parte de los objetivos de esta investigación, se busca dotar de sentido a cada uno de los niveles de desempeño que surja a partir de los puntos de corte de cada una de las áreas, para ello, se analizará los contenidos y procesos que conlleva los ítems que se supone más probable de acertar para cada uno de los niveles de desempeño.

3.6.5.4. Validez predictiva del punto de corte

Finalmente, se buscará corroborar si existe relación entre la aprobación y reprobación de la PHC en función de la nota indicada en este estudio y el éxito académico al cursar MA1001, para ello se buscará realizar una prueba de hipótesis, con ese objetivo, se analizará una muestra de los datos de los que dispone el equipo de la PHC de las personas estudiantes que realizaron la prueba para el proceso de admisión del 2021. Se depurarán los datos facilitados y en función de su naturaleza se seleccionará una prueba estadística apropiada que permita validar la hipótesis nula de que no existe influencia entre la aprobación de la PHC y la aprobación del curso de Cálculo I o en su defecto, la hipótesis alternativa de que sí existe influencia entre ambos resultados. Adicionalmente, con los datos obtenidos de dicha base, se calcularán los valores correspondientes a la especificidad y sensibilidad de la prueba a partir del punto de corte establecido.

3.6.5.5. Aspectos de calidad del proceso de fijación de puntos de corte

Considerando los objetivos de esta investigación, así como su naturaleza, es fundamental dedicar un apartado a la discusión de las evidencias de calidad sobre los resultados del proceso desarrollado. En este sentido, se centrará la atención en los elementos metodológicos y teóricos que aportan en la garantía de la confiabilidad de los hallazgos.

4. Resultados

Los resultados, para efectos de esta investigación, se describen de acuerdo con las etapas desarrolladas en las que se generó información para la obtención del punto de corte, esto debido a que cada una tiene participantes, procedimientos, instrumentos y análisis diferentes. A continuación, se detallan los resultados obtenidos.

4.1. Perfil de la persona mínimamente competente

De manera previa al trabajo con los jueces que colaboraron con el proceso de fijación de los puntos de corte, fue necesario definir el perfil de la persona que posee el nivel mínimo de habilidad para enfrentar y aprobar el curso de Cálculo I.

Con ese objetivo, se procedió a analizar la documentación teórica que respalda parte de las áreas de trabajo de la PHC, en concreto, se logró tener acceso a su Marco Teórico. En este documento, el personal que ha estado a cargo de la elaboración, desarrollo y aplicación de la PHC han incluido una descripción general de los perfiles de razonamiento de las personas aspirantes que han realizado la PHC, esto en función de las notas que obtienen en la prueba. La versión preliminar de ese perfil fue sometida a revisión por parte de colaboradores del proyecto de la PHC, quienes la realimentaron a fin de contar con una versión del perfil más robusta y pertinente.

A partir de esta información, se concretó el perfil de la persona mínimamente competente y se representó con el nombre ficticio de “Cris”, esto con el objetivo de que fuera más sencillo referirse a ese perfil de estudiante durante el trabajo con las personas juezas. En el primer anexo de este documento se ha incluido el perfil que se facilitó y analizó con el equipo de jueces.

La información recopilada, se resumen en tabla 3, en la que se enumeran las tareas que puede desarrollar una persona que coincida con ese perfil, esto organizado también según las dimensiones que se incluyen en la PHC.

Tabla 2

Habilidades deseables en el perfil de la persona mínimamente competente

Dimensión de la prueba	**Nivel mínimo deseable**
Ejemplificación	Elabora un caso apropiado que satisfaga una condición **sencilla**.
	Ejemplo: identificar un número natural que sea par y mayor que 15.
Validación	Determina el valor de verdad de proposiciones considerando información asociada de forma poco explícita. Plantear contraejemplos de afirmaciones universales.
	Ejemplo: determinar si es correcta una afirmación respecto a la variación de la medida de los lados de un polígono y su efecto en la medida del área.
Generalización	Determina un patrón entre objetos de una secuencia o de un conjunto de figuras y lo aplica a un término **inmediato** a los términos presentados.
	Ejemplo: identificar un patrón en una secuencia para establecer el último dígito de determinado elemento de la misma.
Relacionar	Representa relaciones implícitas en el texto. Reconoce propiedades no explícitas en objetos matemáticos.
	Ejemplo: sustituir valores en una fórmula indicada en la prueba para realizar un cálculo determinado.
Clasificar	Identifica alguna propiedad **sencilla** en el objeto que le permita diferenciarlo de otros objetos sin vincularlos entre sí. Analiza conceptos **muy básicos** asociados a la estructura del objeto que le permitan diferenciarlo de los demás.
	Ejemplo: reconocer múltiplos de un número específico a partir de la factorización con números primos del valor dado.

A partir de la información anterior, fue posible incluir una descripción del perfil deseable de esa persona en cada uno de los cuadernillos que emplearon las personas juezas, esto les ayudó a contar con mayor certeza al momento de definir cuántas personas responderían acertadamente cada uno de los ítems de la prueba y centrar su atención en la emisión de los juicios que se les solicitó.

4.2. Resultados de la implementación del método bookmark

En este apartado se presenta la descripción de los resultados obtenidos tras implementar el método bookmark. Si bien en el apartado metodológico se profundizó en los detalles, el proceso desarrolladó implicó la revisión del cuadernillo de la prueba por cada uno de los jueces y juezas, esto con el objetivo de contar con un ítem señalado como el marcador para cada dimensión presente en ese folleto de la PHC. Con los resultados preliminares se analizaron los valores obtenidos para cada una de las cuatro dimensiones consideradas en este cuadernillo: relacionar, validar, clasificar y generalizar. Posteriormente, se convocó a una sesión más al jurado para conciliar la posición del ítem marcador y así obtener un único valor a considerar como punto de corte. A continuación, se presenta el análisis de los datos obtenidos en esta fase.

4.2.1. Ítem bookmark para la dimensión relacionar

Para la primera dimensión que se evalúa en la PHC, que además, para el cuadernillo seleccionado en esta investigación es la que presenta la mayor cantidad de reactivos (19 en total), los ítems señalados individualmente por los jueces como el bookmark, son los indicados en la tabla 3.

Tabla 3

Ítem bookmark indicado por los jueces para la dimensión relacionar

Identificador del juez	Ubicación del ítem bookmark
J1	Ítem 14
J2	Ítem 12
J3	Ítem 4
J4	Ítem 10
J5	Ítem 12
J6	Ítem 7

Mediana de la ubicación del ítem	11
Total de ítems de la dimensión:	19

Es importante señalar que esta dimensión presenta valores variados, los que incluyen el caso del juez 3 que indica que el ítem 4 debe ser el bookmark, lo que supondría que son pocos los ítems que debería responder de manera correcta la persona que aprobaría la PHC. En el otro extremo, se ubica al juez 1, quien indica que el ítem bookmark debe ser el 14, lo cual dejaría a la persona postulante un margen de error de 6 ítems para ofrecer evidencias de que cuenta con el nivel mínimo deseable en lo que refiere a la dimensión relacionar.

Observando la generalidad de los datos, la mediana de la ubicación que indican las personas juezas para el ítem bookmark de esta primera dimensión es el ítem 11. Lo anterior implica que se hace necesario acertar al menos 10 reactivos de los 19 totales para garantizar que la persona que realiza la PHC posee el nivel mínimo para la dimensión relacionar.

Al momento de justificar la decisión por la que fijaron el ítem bookmark de esta dimensión en ese punto, la mayoría de los jueces indica que se debe a la complejidad de las relaciones que se deben establecer entre los elementos de cada reactivo, ya sea para realizar el procedimiento que conduce a la respuesta o para vincular un valor obtenido con la opciones de respuesta que se indican en la prueba.

4.2.2. Ítem bookmark para la dimensión validar

En lo que refiere a la dimensión validar, que es la segunda en cuanto al número de reactivos que aporta para la prueba (11 en total), los resultados indicados por los jueces se resumen en la tabla 4.

Tabla 4

Ítem bookmark indicado por los jueces para la dimensión validar

Identificador del juez	Ubicación del ítem bookmark
J1	Ítem 8
J2	Ítem 6
J3	Ítem 5
J4	Ítem 10
J5	Ítem 5
J6	Ítem 4
Mediana de la ubicación del ítem	5,5
Total de ítems de la dimensión:	11

Al igual que en la dimensión anterior, la ubicación del ítem bookmark varía entre los valores extremos que indican los jueces 6 y 4, estos corresponden al ítem 4 y el ítem 10, de manera respectiva. Es oportuno destacar que el juicio que ofrece el juez 4 implicaría que la persona estudiante debe responder de manera correcta casi la totalidad de esta sección de la prueba, pues en total posee 11 reactivos, esto contrasta con el criterio de los demás jueces, quienes indicaron una posición más conservadora respecto a la ubicación del ítem bookmark.

Buscando una posición equilibrada entre los criterios de los jueces, la mediana de la ubicación para el ítem bookmark para esta sección, corresponde al valor de 5,5. Este valor se redondeará a 6 para los efectos de este proceso. Lo anterior implica que la

persona que se postula para ingresar a las carreras que aplican la PHC deben acertar al menos 5 preguntas de las 11 que se plantean para esta sección.

Como elementos que justifican el criterio de los jueces para fijar el ítem en la posición que señalan, ellos y ellas hacen referencia a la necesidad de integrar relaciones no explícitas entre los elementos que se incorporan en los reactivos de la sección, además, hacen referencia a que hay un proceso de análisis más retador para resolver los ejercicios y que además, pueden haber algunos obstáculos procedimentales al realizar el ejercicio.

4.2.3. Ítem bookmark para la dimensión clasificar

La cantidad de ítems que se vinculan con la dimensión clasificar se reduce de manera significativa en comparación con las dos secciones que se han abordado previamente. En este caso, el total de reactivos corresponde a 4, esto hace que los jueces tengan menos opciones para fijar el ítem marcador según su criterio. El resumen de los datos obtenidos es presentado en la tabla 5.

Tabla 5

Ítem bookmark indicado por los jueces para la dimensión clasificar

Identificador del juez	Ubicación del ítem bookmark
J1	Ítem > 4
J2	Ítem > 4
J3	Ítem 3
J4	Ítem 2
J5	Ítem > 4
J6	Ítem > 4
Mediana de la ubicación del ítem	>4
Total de ítems de la dimensión:	>4

Considerando el criterio que ofrecen los jueces, es evidente que las dos terceras partes del equipo considera que para que la persona que realiza la prueba cuente con el nivel mínimo deseado de habilidad para la dimensión clasificar, es necesario que acierten todos los ítems de esta sección, es por ello que en la celda correspondiente al valor que indican los jueces J1, J2, J5 y J6, el ítem bookmark se considera que está fuera del rango de los ítems que aparecen en el cuadernillo empleado para el método, esto se representa como Ítem > 4 . En contraste, el juez 4 indica que bastaría con acertar uno de los ítems y el juez 3 indica que con dos bastará.

Al calcular el valor de la mediana para la posición del ítem marcador para esta sección, se obtiene un valor que superaría a los 4 ítems de la dimensión, por ello se concluye que para evidenciar el nivel mínimo deseable para esta dimensión, la persona deberá responder adecuadamente los 4 reactivos que aporta esta sección a la prueba.

Para fundamentar el ítem que seleccionaron como el bookmark en esta dimensión, las personas juezas señalan dificultades en elementos procedimentales para resolver el ejercicio, pero también hacen referencia a elementos conceptuales que podría interferir con el proceso de clasificación que se indica en el ejercicio. Esta es la primera sección en la que se indica por parte de los jueces, que los contenidos que abordan los reactivos inciden en su elección del ítem señalado como bookmark, a diferencia de las otras secciones, en las que su criterio se centraba en los procesos de razonamiento.

4.2.4. Ítem bookmark para la dimensión generalizar

En el caso de la cuarta y última dimensión del cuadernillo de la PHC que se analiza, la cantidad de ítems se reduce aún más, pues cuenta con solamente dos. Los valores indicados por los jueces se presentan en la tabla 6.

Tabla 6

Ítem bookmark indicado por los jueces para la dimensión generalizar

Identificador del juez	Ubicación del ítem bookmark
J1	Ítem > 2
J2	Ítem > 2
J3	Ítem 1
J4	Ítem 1
J5	Ítem > 2
J6	Ítem > 2
Mediana de la ubicación del ítem	>2
Total de ítems de la dimensión:	>2

En coherencia con la sección anterior, las dos terceras partes de los jueces coinciden en que el ítem Bookmark supera a la segunda pregunta de la dimensión, sin embargo, la tendencia de los jueces 3 y 4 coincide con la de la sección anterior e indican valores más conservadores para este apartado de la prueba.

Al buscar la mediana para ubicar el ítem bookmark, según los datos disponibles, se consideraría que la posición de ese reactivo supera la posición 2. Por ello, la persona que posee el nivel mínimo deseable para la dimensión generalizar, deberá acertar los dos reactivos de esta sección.

Para justificar el por qué han señalado esa ubicación para el ítem bookmark, las personas juezas indican que su elección es debida a la cantidad y complejidad de los elementos que se deben combinar para resolver el ejercicio de manera satisfactoria, no refieren solamente a los aspectos conceptuales, sino también a los procedimentales.

4.2.5. Consenso con el equipo de jueces

Una vez que se obtuvieron los valores preliminares para fijar el ítem bookmark de cada dimensión, se procedió a convocar nuevamente al equipo de jueces y juezas para

discutir y analizar en detalle los valores obtenidos por medio del método bookmark y las justificaciones para establecer ese valor en cada ítem seleccionado.

Para desarrollar este encuentro con las personas juezas, se presentó la lista de valores preliminares que asignó cada juez en la ronda que desarrollaron de manera individual, se indicó la justificación de los valores indicados por cada juez en cada caso y se procedió a desarrollar un espacio de discusión a fin de unificar el criterio de los jueces y obtener un único ítem marcador. Tras analizar los motivos expuestos por cada juez, se logró llegar a un único juicio que supone el consenso del equipo de docentes.

Como referencia, en la tabla 7 se presentan los datos que se analizaron al inicio de esa sesión.

Tabla 7

Resumen de los Ítems bookmark por dimensión

Identificador del juez	Bookmark relacionar	Bookmark validar	Bookmark clasificar	Bookmark generalizar
J1	14	8	>4	>2
J2	12	6	>4	>2
J3	4	5	3	1
J4	10	10	2	1
J5	12	5	>4	>2
J6	7	4	>4	>2
Mediana de la posición del ítem bookmark	11	5,5	>4	>2

Tras el análisis en conjunto y el espacio de discusión, el equipo de jueces logró consensuar la ubicación del ítem marcador para cada dimensión. Inicialmente, se acordó una posición basada en los datos de la tabla 7. Sin embargo, como ya se indicó, luego

de una discusión profunda y una valoración de los argumentos presentados, se realizó un ajuste final, determinando que la ubicación definitiva del ítem marcador para cada dimensión debía corresponder con lo expuesto en la tabla 8, que se presenta a continuación.

Tabla 8

Resultado del consenso para el Ítem bookmark

Dimensión	Relacionar	Validar	Clasificar	Generalizar
Ítem bookmark consensuado	11	6	>4	>2

Debe señalarse que respecto a los valores preliminares se analizaron en conjunto con los jueces, a partir del consenso, la única dimensión que tuvo variaciones luego de la discusión que se realizó, fue la de validar, debido a que el valor inicial sin redondeo alguno fue de 5,5 y a criterio de los jueces, debe corresponder a un 6, lo cual en la práctica no supone una gran variación, pero da mayor contundencia a las conclusiones obtenidas en esta fase del estudio.

Las demás dimensiones se mantienen sin variación, pues para relacionar se mantuvo la posición del ítem 11 y para clasificar y generalizar se concluyó que en efecto las personas que realizaban la PHC debían acertar la totalidad de los ítems de esas dos dimensiones, es decir, el ítem bookmark sobrepasa la posición 4 y la 2 de manera respectiva.

A continuación, se reumen el proceso implementado para obtener un único ítem para la prueba en general a partir de los que ofrecieron los jueces y juezas para cada dimensión.

4.2.6. Ítem bookmark general

En este apartado, se busca identificar el ítem de la prueba que establece el límite entre los estudiantes que alcanzan el nivel mínimo de habilidad requerido para tener éxito en el curso MA1001, según el perfil definido en esta investigación. Además, se pretende relacionar este nivel de habilidad con una puntuación de aprobación en la Prueba de Habilidades Cuantitativas (PHC), utilizando una escala de 0 a 100.

Para lograr este objetivo, se determinó la dificultad Rasch de cada ítem del cuadernillo de examen de la PHC utilizado en esta investigación. Este análisis se realizó mediante el software *R* (R Core Team, 2022), aplicando el paquete *Extended Rasch Modeling* (eRm) de Mair y Hatzinger (2007).

Es importante señalar que la escala de dificultad Rasch se mide en logits, con un rango de -3 a 3. Un valor de -3 indica un ítem de muy baja dificultad, 3 representa un ítem de alta dificultad, y 0 logits corresponden a un ítem de dificultad media.

Los resultados obtenidos para los ítems de la prueba se resumen en la tabla 9.

Tabla 9

Dificultad Rasch para los ítems de la prueba por dimensión

Dimensión	Ítem de la dimensión	Dificultad Rasch
Relacionar	1	-1,733
	2	-0,709
	3	-0,559
	4	-0,473
	5	-0,443
	6	-0,367
	7	-0,289
	8	-0,285
	9	-0,281
	10	-0,128
	11	**-0,037**
	12	0,282
	13	0,407
	14	0,439
	15	0,633
	16	0,903
	17	1,056
	18	1,218
	19	1,310
Validar	1	-1,140
	2	-1,063
	3	-0,406
	4	-0,132
	5	0,004
	6	**0,062**
	7	0,114
	8	0,485
	9	0,497
	10	0,747
	11	0,977
Clasificar	1	-1,209
	2	-0,300
	3	0,093
	4	**0,287**
Generalizar	1	-0,127
	2	**0,167**

De la información presentada en la tabla anterior, se destaca que los ítems 1 de la dimensión relacionar y 1 de la dimensión clasificar son los de menor dificultad, con valores de -1,733 y -1,209 logits respectivamente. En contraste, los ítems 15 y 17 de la dimensión relacionar presentan la mayor dificultad, con valores de 1,218 y 1,310 logits respectivamente.

Para la aplicación del método Bookmark, se resaltaron en negrita los ítems de interés para definir el punto de corte global, junto con sus respectivas dificultades. Estos ítems son: el ítem 11 de relacionar (-0,037 logits), el ítem 6 de validar (0,062 logits), y los ítems de mayor dificultad en clasificar y generalizar (0,287 y 0,167 logits respectivamente).

Como se indicó en la sección metodológica, el foco de atención se centra en el ítem bookmark de menor dificultad Rasch, que en este caso es el ítem 11 de la dimensión relacionar ($\beta = -0{,}037$).

Siguiendo los principios teóricos del método bookmark, la estimación del nivel de habilidad del estudiante mínimamente competente no se basa en el ítem bookmark en sí, sino en el ítem inmediatamente anterior. Para facilitar esta evaluación, la tabla 10 presenta los ítems de la prueba ordenados por dificultad Rasch de menor a mayor.

Tabla 10

Dificultad Rasch para los ítems de la prueba

Ítem de la dimensión	Dificultad Rasch
1	-1,733
2	-1,209
3	-1,140
4	-1,063
5	-0,709
6	-0,559
7	-0,473
8	-0,443
9	-0,406
10	-0,367
11	-0,300
12	-0,289
13	-0,285
14	-0,281
15	-0,132
16	-0,128
17	**-0,127**
18	**-0,037**
19	0,004
20	0,062
21	0,093
22	0,114
23	0,167
24	0,282
25	0,287
26	0,407
27	0,439
28	0,485
29	0,497
30	0,633
31	0,747
32	0,903
33	0,977
34	1,056
35	1,218
36	1,310

En la tabla 10, el ítem bookmark de menor dificultad se encuentra en la posición 18, considerando la totalidad de los ítems de la prueba, y no su organización por dimensiones. Por lo tanto, el ítem de interés es el anterior, el 17, con una dificultad Rasch de -0,127 logits. Ambos ítems se han resaltado en negrita en la tabla 10.

El valor de dificultad del ítem 17 se utiliza para calcular el nivel de habilidad requerido (θ_i) para que un estudiante mínimamente competente tenga una probabilidad de acierto del $\frac{2}{3}$. Al aplicar la fórmula del modelo de Rasch, se obtiene:

$$\theta = \ln 2 + -0{,}127 \approx 0{,}566$$

Este resultado indica que, según el método bookmark, un estudiante que cumpla con el perfil de aprobación de la PHC debe tener una probabilidad de acierto superior a los $\frac{2}{3}$ en el ítem 17, lo que equivale a una habilidad mínima de 0,566 logits según el modelo de Rasch.

Para complementar el análisis del método bookmark, se relaciona el nivel de habilidad del estudiante con el número de respuestas correctas en la prueba. Para ello, se utilizó nuevamente el paquete eRm en *RStudio*, calculando los parámetros de los estudiantes. Los resultados obtenidos se resumen en la tabla 11, que muestra la relación entre la cantidad de respuestas correctas y el nivel de habilidad.

Tabla 11

Correspondencia entre el número de preguntas correctas y el grado de habilidad de la persona examinada

Preguntas correctas	Nivel de habilidad
2	-3,044
3	-2,592
4	-2,258
5	-1,988
6	-1,758
7	-1,556
8	-1,373
9	-1,206
10	-1,049
11	-0,902
12	-0,762
13	-0,626
14	-0,496
15	-0,368
16	-0,243
17	-0,119
18	0,004
19	0,127
20	0,251
21	0,376
22	0,503
23	**0,634**
24	0,768
25	0,908
26	1,054
27	1,209
28	1,375
29	1,556
30	1,756
31	1,984
32	2,251
33	2,582
34	3,030
35	3,765
36	4,569

Al combinar la información de la tabla 11 con los valores obtenidos para el ítem 17, se observa que los estudiantes que alcanzaron al menos ese nivel de habilidad obtuvieron 23 respuestas correctas en la prueba. Según la tabla 11, este puntaje corresponde a un nivel de habilidad de 0,634 logits, valor resaltado en la tabla.

Para determinar la nota de aprobación correspondiente a 23 respuestas correctas en una prueba de 36 ítems, se aplicó la fórmula estándar de cálculo de calificaciones:

$$\frac{23}{36} \times 100 = 63{,}88 \approx 64$$

De este cálculo se concluye que, según el método bookmark, la nota de aprobación en la PHC debe ser de 64 en una escala de 0 a 100.

4.3. Resultados de la implementación del método Angoff

Esta sección se presentan los resultados del proceso de aplicación del método Angoff. Para ello, se solicitó a cada juez o jueza, en dos rondas distintas, que estimara el porcentaje de estudiantes (en un grupo de 100) que, con el nivel de habilidad mínimo requerido para aprobar los cursos de Cálculo I, responderían correctamente a cada ítem. Las estimaciones debían realizarse en múltiplos de 10.

Los criterios de los jueces fueron variados en algunos ítems y oscilaron desde la nota máxima (100) que suponía un ítem que todos los postulantes con el nivel requerido debería acertar, hasta el 30, dato que supondría que ese reactivo sólo lo responderán de manera correcta 30 de las 100 personas con el nivel de habilidad mínimo que se solicita según el perfil que sirvió de base para este análisis.

En la tabla 12 se presenta un resumen de los valores obtenidos tras la implementación de este método, los ítems se han organizado según la dimensión

respectiva y considerando un índice de dificultad ascendente. Adicionalmente, se incluyen dos columnas a la derecha que indican la mediana del porcentaje de acierto que indicaron todas las personas juezas en la primera y segunda ronda del método. De la misma manera, se agregó una columna en la que se indica el recorrido (valor máximo - valor mínimo) de los criterios que emitieron los jueces en cada ronda y se destaca con negrita el ítem bookmark indicado por el equipo de jueces.

Al finalizar cada dimensión, se resume el promedio de las medianas de los valores que asignaron los jueces para los ítems de esa área y finalmente, en la fila inferior de la tabla, se indica el promedio de acierto que indica cada juez para la totalidad de la prueba y las celdas en el extremo inferior derecho, indican la nota de punto de corte obtenido en cada ronda.

Si bien para los cálculos que se presentan en la tabla se consideraron todos los ítems, para el análisis y comparación de los resultados del método Angoff con el bookmark, se dio prioridad a las dimensiones de relacionar y validar, pues son las más representativas de la prueba por la cantidad de reactivos que aportan a la misma.

Tabla 12

Resumen de los resultados del método Angoff por cada ítem y juez

		Juez 1		Juez 2		Juez 3		Juez 4		Juez 5		Juez 6		Mediana por ítem		Recorrido	
Dimensión	Ronda /Ítem	R1	R2	R1	R2	R1	R2	R1	R2	R1	R2	R1	R2	R1	R2	R1	R2
	I1	100	100	80	90	90	90	80	80	90	80	80	80	85	85	20	20
	I2	90	100	80	80	60	70	60	80	90	80	70	70	75	80	30	30
	I3	90	100	80	80	70	70	70	60	80	70	70	70	75	70	20	40
	I4	90	90	80	80	90	70	80	90	80	60	50	60	80	75	40	30
	I5	80	90	70	70	100	60	80	80	70	90	60	60	75	75	40	30
	I6	80	80	70	70	90	90	50	60	60	50	60	60	65	65	40	40
	I7	70	80	70	70	90	90	80	70	70	60	50	50	70	70	40	40
	I8	70	70	70	70	100	80	90	80	80	80	50	60	75	75	50	20
	I9	70	70	70	70	70	60	70	70	70	60	50	60	70	65	20	10
Relacionar	I10	70	60	70	60	90	50	40	40	50	50	50	50	60	50	50	20
	I11	**70**	**70**	**60**	**60**	**50**	**60**	**70**	**60**	**80**	**70**	**50**	**60**	**65**	**60**	**30**	**10**
	I12	60	60	60	60	80	70	50	70	60	60	60	60	60	60	30	10
	I13	60	60	60	50	70	50	70	70	70	80	50	60	65	60	20	30
	I14	50	50	50	50	60	60	60	60	50	60	40	50	50	55	20	10
	I15	70	50	40	40	90	60	70	60	60	50	60	70	65	55	50	30
	I16	70	50	40	40	70	40	80	80	70	60	60	70	70	55	40	40
	I17	60	50	40	40	80	40	80	80	40	30	60	70	60	45	40	50
	I18	40	40	40	30	60	30	50	40	60	50	40	50	45	40	20	20
	I19	50	30	30	30	70	60	50	50	80	60	40	50	50	50	50	30
Media de las medianas para el área relacionar														66,32	62,63		
	I20	100	100	80	80	100	90	90	90	70	90	70	60	85	90	30	40
	I21	90	90	80	80	90	90	90	80	70	80	60	60	85	80	30	30
	I22	80	90	80	80	70	70	70	70	60	70	60	60	70	70	20	30
	I23	80	80	80	80	80	80	90	90	80	70	60	50	80	80	30	40
	I24	70	80	70	70	80	50	70	70	50	50	60	60	70	65	30	30
Validar	**I25**	**70**	**70**	**60**	**60**	**60**	**40**	**80**	**80**	**70**	**60**	**70**	**60**	**70**	**60**	**20**	**40**
	I26	60	70	50	60	50	80	70	70	60	60	70	60	60	65	20	20
	I27	50	60	50	50	80	60	90	80	50	50	60	70	55	60	40	30
	I28	60	50	40	50	70	70	80	80	40	40	50	60	55	55	40	40
	I29	40	50	40	50	70	90	70	70	50	50	50	60	50	55	30	40
	I30	30	40	40	50	100	70	90	90	60	60	60	60	60	60	70	50
Media de las medianas para el área validar														67,27	67,27		
	I31	100	100	90	80	90	70	90	90	50	80	70	70	90	80	50	30
Clasificar	I32	90	80	80	70	80	60	80	60	70	80	50	60	80	65	40	20
	I33	70	70	80	60	70	50	80	80	80	90	60	70	75	70	20	40

	I34	60	50	60	50	90	80	80	80	80	80	50	50	70	65	40	30
Generalizar	I35	70	80	80	80	50	50	80	60	60	60	50	50	65	60	30	30
	I36	50	50	70	70	40	50	80	80	60	40	40	40	55	50	40	40
Media de las medianas para las otras áreas														72,50	65,00		
Media para la prueba		69,72	69,72	63,61	62,78	76,39	65,28	73,89	72,22	65,83	64,17	56,67	60,00	67,64	64,44		
Promedio de los recorridos por ronda																34,2	30,3

Con el propósito de detallar la información que se presenta en la tabla de resumen, es importante analizar cada una de la áreas, esto se realiza en los siguientes apartados.

4.3.1. Análisis Angoff para la dimensión relacionar

En la dimensión de relacionar, los ítems 1 y 19 fueron identificados como los de menor y mayor dificultad respectivamente, lo cual es consistente con la estructura de la prueba. Se observó una mayor variabilidad en las estimaciones de los jueces para los ítems 17, 6, 7 y 16, con rangos de 40 a 50 puntos porcentuales entre las estimaciones individuales. Sin embargo, las estimaciones de cada juez mostraron una notable consistencia entre las dos rondas.

En general, se detectó una ligera disminución en las estimaciones promedio de los jueces entre la primera y la segunda ronda, pasando de aproximadamente 66% a 62%.

Los jueces, en sus cuadernillos de trabajo, señalaron un punto de inflexión en la dificultad de los ítems entre los números 8 y 12. Los ítems anteriores a este punto se consideraron viables de resolver con la aplicación de cálculos directos y sencillos, mientras que los posteriores requerían la integración de conceptos y conocimientos previos más complejos, como ecuaciones, fórmulas notables y procesos algebraicos.

Según el consenso del equipo de jueces, se considera necesario obtener una puntuación mínima de 63 en la escala de 0 a 100 para asegurar un nivel de competencia aceptable en esta área de la prueba.

4.3.2. Análisis Angoff para la dimensión validar

En la dimensión validar, el equipo de jueces y juezas coincidieron en que el ítem inicial era el más sencillo, evidenciado por una mediana de 90% de probabilidad de acierto asignada en la segunda ronda.

A diferencia de la dimensión anterior, los ítems 28 y 29 fueron considerados los más difíciles. Los jueces justificaron esta percepción indicando que estos ítems requerían el uso de múltiples representaciones de datos y la integración de diversos contenidos matemáticos, como los porcentajes y las probabilidades presentados en formatos de números decimales y representaciones fraccionarias.

Es importante destacar que la dificultad percibida en estos ítems se atribuyó principalmente a los conocimientos previos necesarios para interpretar la información, más que al concepto evaluado en sí.

En general, se observó una variación significativa en las estimaciones de probabilidad de acierto entre los ítems 25 y 27. Esta variabilidad se explicó por la necesidad de convertir datos entre diferentes representaciones, lo cual fue identificado como un factor de dificultad clave en esta dimensión.

El porcentaje de aprobación estimado por los jueces para estudiantes mínimamente competentes fue consistente en ambas rondas, situándose en aproximadamente 68% en la escala de 0 a 100. Sin embargo, se observó una alta

variabilidad en las estimaciones individuales para el ítem final, con rangos de 40% a 90% en la segunda ronda y de 30% a 100% en la primera.

4.3.3. Análisis Angoff para las dimensiones clasificar y generalizar

Debido al número reducido de ítems en las dimensiones clasificar y generalizar en la confección del folleto de esta versión de la prueba, se decidió fusionarlas para el análisis. El limitado número de reactivos de estas dimensiones limitó la capacidad de los jueces para emitir juicios precisos sobre los niveles de aprobación. Por ejemplo, en la dimensión clasificar, todos los ítems mostraron una mediana de probabilidad de acierto superior al 65%, mientras que en generalizar, solo dos ítems (35 y 36) obtuvieron estimaciones de 60% y 50% respectivamente.

Al considerar ambas dimensiones como una sola, se identificó que el ítem 31 fue percibido como el más sencillo y el 36 como el más difícil. Los ítems de generalizar mostraron porcentajes de probabilidad de acierto generalmente más bajos que los de clasificar. Sin embargo, a diferencia de las otras dimensiones, la información cualitativa proporcionada por los jueces sobre los conocimientos y habilidades que influyeron en sus estimaciones fue limitada.

Durante el proceso de aplicación de este método, los jueces señalaron que, a pesar del número reducido de ítems en generalizar, estos requerían habilidades complejas y la integración de diversos conocimientos para identificar patrones y generalizar fenómenos. Se consideró que esta dimensión era una de las más desafiantes,

debido a la necesidad de reconocer situaciones específicas y sus condiciones de generalización.

Finalmente, se observó que los ítems 33 y 36 presentaron la mayor variabilidad en las estimaciones de probabilidad de acierto, con un rango de 40 puntos en la segunda ronda del método Angoff.

4.3.4. Resultados generales para el método Angoff

Para poder llegar a una conclusión respecto a la implementación del método Angoff en esta investigación, es importante considerar que la atención se centra en la segunda ronda del método, pues tal y como se pudo visualizar en la tabla que resume las rondas del método, en 27 de los 36 ítems de la prueba hubo una reducción del rango de los valores que asignaron los jueces entre la primera y la segunda ronda, esto implica una menor variabilidad en dichos datos, esto se verifica con una reducción de 34,2 a 30,3 en el promedio de los recorridos de los datos entre la primera y la segunda ronda, por ello se considera la ronda fundamental para fijar el punto de corte.

A partir de los resultados de la segunda ronda, es posible indicar que en relación con los reactivos de la prueba, el ítem que fue percibido con la mayor probabilidad de acierto por parte del jurado, fue el ítem 20, esto con una mediana de 90 y presentando en algunos casos un valor de 100% de probabilidad de acierto por parte de al menos dos jueces distintos entre la primera y la segunda ronda.

En contraste, el ítem percibido con menor porcentaje de acierto fue el ítem 18, el cual mostró un valor menor extremo de 30 según lo indicado por dos jueces en sus segundas rondas.

Centrando la mirada en la revisión general de los resultados para la prueba, también es de relevancia concentrar la atención en la justificación que brindaron los jueces para establecer los valores de probabilidad de acierto para cada uno de los ítems. De manera general, el equipo de jueces coincide en que los ítems que son considerados de un nivel básico, es decir, los que poseen mayor probabilidad de acierto, son aquellos que incorporan un menor número de contenidos y procesos, o que aquellos resultan básicos para las personas estudiantes, esto ya se comentó para las dos dimensiones más representativas de la prueba.

Para los ítems de nivel alto, es decir, los de la dimensión generalizar, que fueron aquellos que presentaron un bajo porcentaje de acierto para los examinados mínimamente competentes, el equipo de jueces indica que la percepción de su dificultad radica en la incorporación de conceptos más complejos o abstractos para el nivel de las personas postulantes, aunado a ello, indican la posibilidad de que esos ítems requieran para su correcto abordaje procesos complicados o que requieren el establecimiento de relaciones no evidentes entre elementos de distintos contextos o más de dos elementos, además de requerir que se pueda reconocer un patrón en el ítem que se puede generalizar a otros elementos que cumplan ciertas condiciones.

Para los demás reactivos que quedan en niveles intermedios, las decisiones se fundamentaron en la aplicación de algún concepto o algoritmo que puede ser de conocimiento y aplicación para un grupo de estudiantes de nivel intermedio.

Desde una óptica general, un elemento destacable de los elementos de justificación que emplearon las personas juezas es que consideraron como un aspecto que puede incidir en la percepción de la dificultad de los ítems a la omisión de

representaciones gráficas que ilustren las situaciones expuestas en los reactivos que así lo podrían incluir.

Según indican los jueces, lo anterior se vincula con la capacidad de abstracción que pueden tener las personas estudiantes para convertir la descripción de una situación en una imagen en una, dos o tres dimensiones. Aunado a lo anterior, algunos de los jueces comentan que esta capacidad de razonamiento abstracto y de poder realizar representaciones gráficas de distintas situaciones, son elementos indispensables para enfrentar un curso de Cálculo I.

Finalmente y como cierre de este apartado, las notas de corte que se indican como producto de la información brindada por los jueces, corresponde a un 67,64 en el caso de la primera ronda y un 64,44 para la segunda iteración. Esta diferencia supone una variabilidad importante en la aplicación del método Angoff, sin embargo, como ya se ha recalcado, la teoría indica que los resultados a considerar serán los de la segunda ronda, pues la misma asume un mayor conocimiento de los jueces del método, del cuadernillo empleado y de la misma prueba, por ello, debería concluir un resultado más confiable. En conclusión, por medio de la aplicación de este método para fijar el punto de corte para la PHC, es posible llegar a un valor de 64, una vez que se apliquen las reglas de redondeo para el valor original.

4.4. Contraste entre los resultados de los puntos de corte

Considerando el proceso desarrollado en este punto de la investigación, gracias a la aplicación de los métodos bookmark y Angoff para el establecimiento de puntos de corte, ha sido posible concluir dicho valor que resultó ser consistente entre ambas metodologías, es decir una nota de 64.

Sin embargo, se debe hacer una diferenciación entre la nota obtenida en este estudio y el resultado práctico en términos de la PHC. Esta observación responde a que en el marco de la prueba y tal y como se indicó en el proceso de construcción de los cuadernillos que emplearon los jueces, hay 4 ítems de la prueba que se descartaron, pues se incluyen como pilotaje en las pruebas de cada año y para el resultado final, estos puntos se cuentan a favor de la persona que realizó la prueba.

En otras palabras, la nota de 64 que indica esta investigación, debe aumentar considerando los 4 puntos de los reactivos de prueba de esa aplicación de la PHC.

Para ello, considere el proceso seguido para determinar el punto de corte según el método bookmark. En él se concluyó que eran necesarios 23 ítems para aprobar la PHC y con ese valor se pudo concluir que la nota correspondiente era un 64 pero tomando como referencia un puntaje total de la prueba de 36 ítems. Ahora, se procederá a calcular la nota para efectos prácticos y por ello el puntaje obtenido (*Po*) pasará de 23 a 27 y el puntaje total de la prueba ya no será de 36, sino de 40.

Ahora, se considerará la siguiente fórmula para calcular la nota obtenida en una prueba en función de los puntos obtenidos (Po) y los puntos totales de la misma (Pt).

$$N = \frac{Po \cdot 100}{Pt}$$

Sustituyendo los datos del escenario que se está considerando, obtenemos:

$$N = \frac{27 \cdot 100}{40} = 67{,}5$$

Es decir, a partir de los resultados de esta investigación que incorporó los métodos de fijación de puntos de corte Bookmark y Angoff, se puede asegurar que ese valor límite corresponde a un 67,5 para los efectos prácticos de la PHC.

Ahora bien, el proceso desarrollado para cada uno ha sido distinto y por ello, en este apartado se busca detallar los hallazgos obtenidos al contrastar los resultados y juicios que se obtuvieron luego de la participación de los jueces en el proceso.

Con este fin, la sección abordará esos elementos según cada una de las dimensiones de la prueba que se analizó.

4.4.1. Los resultados de la dimensión relacionar

Para el caso de la dimensión relacionar, el ítem bookmark se ubicó en la pregunta número 11 de esta dimensión, entonces, analizando la teoría que respalda la metodología que siguió esta investigación, no es de extrañar que este ítem también se encuentre cerca del punto en el que para el método Angoff se diferencian aquellos reactivos con mayor porcentaje de acierto para la población con el perfil deseable de aquellos que no lo satisfacen. Esto se puede corroborar en la Tabla 13, donde en las cercanías del ítem 11 se pasa de un porcentaje de respuestas correctas de 65 a 60 o menos.

Ahora, en cuanto a los aspectos que consideró el equipo de jueces para establecer la diferenciación en los niveles, ellos y ellas señalan que los ítems previos al que señalaron como bookmark y los que poseen valores de acierto más altos, son aquellos que resultan más sencillos por el tipo de contenidos que abordan, además, indican que las relaciones que se deben establecer entre los elementos que incorpora el ítem, son pocos, son sencillos y son conocimientos de nivel básico para quien resuelve la PHC.

La diferenciación que hacen con los ítems que señalan como límites en los métodos, responde a un procedimiento más amplio o más detallado. También indican mayor complejidad en los algoritmos que requieren para resolver los ejercicios. En específico, uno de los jueces señala que hay elementos menos tangibles que son requeridos para resolver el ítem que eligió como bookmark.

De ahí en adelante, es decir, para los ítems que superan la complejidad de los que fueron considerados como la frontera, lo que se señala como los elementos considerados fueron el número de relaciones a establecer entre más elementos o la variación entre la representaciones para los conceptos matemáticos involucrados en el ejercicio o también

procesos de análisis y abstracción más complejos que los que estaban antes de las preguntas consideradas como el límite.

4.4.2. Los resultados de la dimensión validar

De manera similar a la dimensión anterior, en la Tabla 13 se puede confirmar que en las cercanías al ítem 25 de la prueba en general (ítem 6 de la dimensión), el procentaje de acierto pasa de un 65 a un 60 o menos, lo cual una vez más indica que hay similitud entre las posiciones generales indicadas para los ítems que marcan un límite para fijar los puntos de corte.

En lo que refiere a los elementos que se consideraron para diferenciar los niveles en ambos métodos, los jueces indican que antes del ítem bookmark y donde hay mayor porcentaje de acierto de los ítems, los reactivos se pueden resolver considerando casos particulares de una situación o que en su defecto consideran análisis básicos de alguna propiedad o escenario para poder resolverlos.

En el otro extremo, es decir, los ítems más dificultosos o que superan el ítem bookmark, son considerados como aquellos que para resolverlos requieren la aplicación de operaciones más complejas y que además, se pueden dificultar por el uso de distintas representaciones para los números, es decir, se pueden presentar valores en forma fraccionaria, porcentual o decimal y esto exige mayor habilidad por parte de las personas estudiantes. Para los ítems que se ubican cercanos a la frontera entre los niveles de desempeño, las personas juezas los encuentran como aquellos que equilibran un nivel de análisis básico, pero apropiado para perfil que se construyó y a la vez, para su solución requieren de un análisis básico y cercano al nivel de quienes se postulan para ingresar a las carreras que aplican la PHC.

4.4.3. Los resultados de las dimensiones clasificar y generalizar

Tal y como se ha indicado en el resto de los resultados de esta investigación, las dimensiones de clasificar y generalizar mostraron un comportamiento atípico. Esto según el criterio de los jueces, debido a la naturaleza y cantidad de los reactivos. Es por ello que en el caso de la primera dimensión, es decir, de clasificar, no se alcanza a delimitar un ítem bookmark, pues el equipo de jueces indica que dicho reactivo debería estar más allá de los 4 que se incluyen en la prueba. Este resultado es consistente con el método Angoff, pues los valores de porcentaje de acierto para los ítems de la dimensión muestran valores altos de aprobación, superiores a 65 de mediana. Es decir, bajo ninguno de los dos métodos es posible establecer un límite entre quienes tienen el nivel competente y aquellas personas que no lo tienen.

Por su parte, la dimensión generalizar muestra valores bajos en cuanto a su aprobación, esto según el método Angoff, pero tampoco se evidencia un ítem bookmark de manera concreta. Una vez más, el criterio de los jueces es que ese ítem bookmark era complejo de asignar debido al número limitado de reactivos disponibles en la sección.

A la luz de estos resultados, es posible establecer una diferenciación entre las habilidades de las personas que aprueban la PHC y quienes no, por ello, en el apartado siguiente se procederá a establecer las caraterísticas del grupo de personas que aprueban la PHC.

4.5. Definición de los niveles de desempeño

Como corolario del proceso que se desarrolló para fijar los puntos de corte la Prueba de Habilidades Cuantitativas, es posible diferenciar dos niveles de desempeño para las personas que realizan la prueba en lo que refiere al razonamiento matemático.

Estos niveles se componen por las personas que obtienen una nota igual o superior a 64, es decir, los aprobados en la PHC y quienes no lo hacen.

Según lo que se ha definido en esta investigación, ese grupo de personas que obtienen una nota igual o superior al punto de corte establecido, se identificará como el nivel competente, mientras que aquellas personas excluidas de este nivel serán aquellas de la categoría insuficiente. Los párrafos siguientes se centrarán en la caracterización de las habilidades de las personas integrantes de cada nivel de desempeño para el razonamiento matemático.

La descripción del perfil del nivel competente se basa en la información indicada por las personas juezas al momento de establecer los puntos de corte según los dos métodos empleados en esta investigación. Se dará por sentado que aquellas personas que se excluyen de esta descripción, serán las que obtienen una nota inferior al punto de corte establecido y con esto se infiere que sus habilidades no se encuentran al nivel que el equipo de jueces consideró necesario para cumplir con el perfil deseable del estudiantado que enfrentará el curso de Cálculo I.

Según se desprende del perfil que se construyó para el desarrollo de las sesiones de trabajo con las personas juezas y la consecuente definición de los puntos de corte, es posible considerar las 5 dimensiones que considera la PHC para definir los niveles de desempeño en el contexto de esta investigación, cada una de ellas se aborda a continuación.

4.5.1. Dimensión relacionar

El relacionar conceptos matemáticos fue la habilidad más evaluada en este cuadernillo de la prueba. Según los jueces, quienes aprueben la PHC demostrarán la

capacidad de identificar relaciones implícitas en enunciados matemáticos y reconocer propiedades no explícitas de diversos objetos matemáticos. Los jueces consideran que las relaciones y conceptos evaluados en esta dimensión son generalmente básicos y de fácil reconocimiento por su uso previo en la educación secundaria.

Por otro lado, los jueces señalan que quienes superen la nota de corte podrán establecer conexiones más complejas entre un mayor número de conceptos matemáticos, en contraste con aquellos que se encuentren cerca del puntaje mínimo, quienes se limitarán a establecer relaciones más simples o evidentes entre dos conceptos básicos. Además, las personas estudiantes podrán hacer uso de representaciones gráficas simples, como un cuadrado inscrito en un círculo o un cubo compuesto por cubos más pequeños, esto para poder visualizar problemas y ayudarse para resolverlos. Algunos jueces consideran que la prueba y error también es una estrategia que podrían aplicar para resolver algunos problemas las personas que aprueben la PHC, así como el construir casos específicos para analizar una situación específica.

También en esta dimensión, será posible que las personas apliquen una fórmula, conocida previamente o no, para resolver un problema. Esto implica que pueda establecer relaciones entre las variables que se le indican en un problema y las que requiere para aplicar la fórmula que se la ha indicado.

En síntesis, quienes aprueben la PHC tendrán la capacidad de establecer relaciones claras y sencillas entre dos conceptos matemáticos básicos, aplicar fórmulas para resolver problemas específicos así como de utilizar representaciones gráficas simples para representar las situaciones que se le plantean como problemas.

4.5.2. Dimensión validar

En cuanto al número de reactivos presentes en el cuadernillo de la PHC que se analizó, el segundo lugar lo ocupa la dimensión validar.

En este sentido, la persona estudiante que obtenga una nota superior al punto de corte, es a criterio de los jueces y juezas, capaz de determinar el valor de verdad de distintas proposiciones, esto considerando información vinculada a los conceptos analizados pero que no se presentan explícitamente. Además, serán capaces de constuir contraejemplos para ciertas afirmaciones que se les planteen.

Además, las y los jueces destacan que son capaces de realizar distintas operaciones o identificar las características generales de un conjunto numérico y sus elementos para poder determinar si cierto resultado es aplicable para todos los integrantes de ese conjunto y en caso de no serlo, pueden determinar las razones por las que no se puede indicar que es un resultado aplicable para todo el conjunto o modificarlo para poder generalizarlo.

Se indica por parte de los jueces que en esta dimensión se puede dar por sentado que las personas que superen el punto de corte, pueden realizar de manera correcta operaciones aritméticas básicas al nivel de la educación secundaria para dar respuesta a las preguntas que se plantean, además de poder establecer relaciones de orden con el mismo objetivo.

Otro de los aspectos fundamentales que se señala por parte del equipo de jueces y que pudiera ser valioso para el perfil de quien enfrenta el curso de MA1001, es que son capaces de plantear distintas situaciones hipotéticas, siempre y cuando sea en un contexto cercano a su conocimiento o realidad.

Entre las limitaciones de esta dimensión que se indican por parte del equipo de jueces para quienes apenas superen el punto de corte, destaca el que tendrán dificultad para validar resultados o afirmaciones que requieran el cambio de distintas notaciones, por ejemplo el unificar la representación de valores entre números racionales expresados en notación decimal, fraccionaria o como porcentajes.

En resumen, quien apruebe la PHC podrá determinar el valor de verdad de una afirmación que incorpore elementos ya conocidos, aunque la información necesaria no se presente de manera explícita y podrá determinar las razones por las que una propiedad no se puede generalizar para todos los elementos de un conjunto.

4.5.3. Dimensión clasificar

La dimensión clasificar es la tercera en orden descendente según la cantidad de ítems que aportó a la prueba y además, posee la particularidad de que el equipo de jueces, en general, indicó que la persona que aprobara la PHC debería poder responder de manera correcta todos los ítems de esta dimensión.

Ahora bien, a criterio de las y los jueces, al aprobar la PHC, la persona estudiante tendrá la capacidad de identificar rasgos de un objeto que le permita diferenciarlo de otros, además, podrá identificar aspectos de la estructura del objeto que resulten ser particulares para destacarlo de los demás.

Adicionalmente, el equipo de jueces resalta que quienes hayan aprobado la PHC serán capaces de resolver situaciones problemáticas que requieran aplicar la abstracción, entendida como el poder analizar una situación específica que no se puede representar gráficamente o que no responde a algún elemento concreto de la realidad cercana a la persona estudiante.

Considerando la ubicación de los ítems bookmark que señalaron los jueces, es de interés que quien supere la PHC también se caracterizará por tener una mayor capacidad de análisis, pues en reiteradas ocasiones los jueces indicaron que, para esta dimensión, uno de los elementos que diferenciaba a la persona mínimamente competente de los demás, era su capacidad de análisis, dando por sentado que hay una diferenciación importante en esta habilidad entre quienes aprueban la PHC y aquellas personas que no lo hacen.

Se concluye que quienes aprueban la PHC según el punto de corte indicado y en el contexto de la dimensión clasificar, serán capaces de detectar particularidades de un objeto para diferenciarlo de los demás, así como de poder resolver problemas que requieran un cierto grado de abstracción y a la vez, se les reconoce un mayor nivel de análisis para enfrentar distintas situaciones problemáticas en el contexto del razonamiento matemático.

4.5.4. Dimensión generalizar

La dimensión generalizar es la que aportó menos elementos al cuadernillo de la prueba seleccionado para este análisis, sin embargo, su aporte en la definición del perfil de la persona estudiante no es irrelevante.

Las personas cuya nota supere el punto de corte de la PHC, serán capaces de determinar un patrón o una secuencia entre un subconjunto determinado de elementos de un conjunto. Sin embargo, según lo indicado por el equipo de jueces y juezas, el factor que les ayudó a determinar el punto de inflexión para esta dimensión, fue el poder aplicar el patrón detectado a otros elementos del conjunto o de otro.

En otras palabras, a criterio del jurado, el reconocer patrones y secuencias se da por sentado para un amplio rango de las y los postulantes que realizan la PHC, pero el poder reconocer ese patrón o secuencia, expresarlo de algún modo y poder replicarlo para otros elementos es lo que marca la pauta para definir quién aprueba la PHC y quién no.

Por otro lado, se identifica como una caraterística de quienes sobrepasan el punto de corte de la PHC, el poder generalizar el patrón o la secuencia para elementos lejanos de los que se consideraron inicialmente para analizar el patrón. Por ejemplo, resulta sencillo generalizar y aplicar una propiedad para los primeros 5 elementos de un conjunto, pero si se pide que determine el resultado para el elemento número 467, será mucho más complejo, aunque no se requiera hacer el cálculo de manera consecutiva para cada uno de los elementos, sino identificar un patrón como una secuencia en los últimos dígitos del resultado obtenido o algo similar.

Además, se señala como otra dificultad en esta dimensión el poder diferenciar resultados particulares en una secuencia de aquellos que son generalizables, es decir, puede que la persona estudiante analice 5 casos que definen un patrón y considera que ese es el resultado en general, pero el caso siguiente muestra un comportamiento distinto que pasó desapercibido a los ojos de quien realiza la prueba.

En síntesis, considerando la dimensión generalizar, quien apruebe la PHC podrá reconocer patrones y definir secuencias para aplicarlas a otros elementos de un conjunto, siempre y cuando esos patrones sean claros y se definan considerando pocos elementos del conjunto originalmente aportado en el ejercicio.

4.5.5. Dimensión ejemplificar

La dimensión ejemplificar no se consideró como parte de los reactivos de la prueba que se consideró para esta investigación, sin embargo, a partir del trabajo que se realizó con el equipo de jueces, es posible establecer que la persona que aprueba la PHC incorpora en su perfil la capacidad de elaborar un caso apropiado en el contexto matemático que cumpla una condición sencilla.

Es importante que esta dimensión aunque es exclusiva de las demás, incorpora elementos cercanos a la de validar, misma que supone la construcción de contraejemplos para determinadas condiciones, así como a la de clasificar que implica detectar particularidades para poder diferenciar u organizar elementos de distintos conjuntos. Lo anterior no la hace ni más ni menos importante que las otras, pero al considerar su sinergia, sí las dota de mayor relevancia por las implicaciones que tiene en el desarrollo formativo de quien enfrenta un curso como Cálculo I.

Realizando una revisión general de las dimensiones que componen la PHC, es posible caracterizar a la persona que obtiene una nota superior al punto de corte establecido en la PHC, como aquella que es capaz de establecer relaciones entre distintos elementos, así como de aplicar fórmulas variadas para resolver diferentes situaciones problemáticas. A su vez, puede realizar operaciones básicas de manera correcta, construir contraejemplos para ciertas afirmaciones que sean cercanas a sus conocimientos previos, además conoce los conjuntos numéricos $\mathbb{N}$, $\mathbb{Z}$, $\mathbb{Q}$, $\mathbb{I}$ y $\mathbb{R}$, así como sus propiedades. También es capaz de plantear hipótesis para distintas situaciones. Posee también un nivel de abstracción superior al resto de los examinados, lo cual resulta

relevante para un curso que estudia el cálculo de límites, el proceso de derivación e integración y que a la vez les diferencia de quienes no aprueban la PHC.

Considerando los elementos expuestos en este apartado, será de interés para este estudio el considerar si existe alguna relación entre la calificación obtenida en la PHC y la aprobación del curso de Cálculo I en la población que realizó la prueba seleccionada para esta investigación, este proceso se aborda en la sección siguiente.

4.6. La validez predictiva del punto de corte

Si bien ya en otras publicaciones como por ejemplo en Rojas (2014) o en Rojas y Ordoñez (2019) se abordado en detalle el proceso de construcción de la PHC, sus alcances y su correlación con algunos elementos de la vida académica de las personas estudiantes que la realizan, en este apartado interesa centrarse en las implicaciones que tiene el aprobarla o no en términos de la nota de corte que se ha fijado.

Para analizar los efectos del resultado en la PHC con el éxito académico de las personas estudiantes, en específico con los resultados que obtienen en el curso de Cálculo I, se procedió a solicitar a los encargados de la prueba una base de datos en la que se incluyera el resultado que obtuvo cada persona que realizó la PHC para el proceso de admisión del año 2021, así como el resultado que obtuvo en el curso MA1001.

Una vez que se obtuvo la base de datos, se procedió a depurarla, para ello, se procedió a descartar los casos en los que presentaron la PHC pero que no matricularon Cálculo I, además se descartó a quienes realizaron retiro de matrícula del curso y a quienes lo equipararon en cualquier momento por otro curso. Finalmente, la base de datos que se utilizó estuvo conformada por 129 casos que contaban con la información requerida para el análisis propuesto.

Una vez que se contó con la base de datos refinada y que contuviera solamente los casos en los que había registro de la nota obtenida en PHC y el resultado de la primera vez que cursaron MA1001, se procedió a cargar los datos en el software *RStudio* para realizar el análisis que se detalla en los apartados siguientes.

4.6.1. La correlación de los datos

Ya en otras publicaciones que realizan procesos de análisis sobre la PHC se ha indicado que en efecto existe una correlación entre los resultados de la prueba y el éxito en la vida académica de las personas estudiantes, específicamente al cursar asignaturas vinculadas con la matemática. Sin embargo, resulta de interés conocer si en esta muestra se mantiene esa tendencia y además, en qué grado se presenta esa correlación.

Para cumplir con la tarea propuesta, se procedió a realizar un análisis de la correlación de las variables Nota obtenida en la PHC y Nota obtenida en MA1001, esto utilizando el método de correlación de Spearman. El resultado obtenido en dicho proceso fue de un 0,54 y esto implica que en efecto existe una correlación entre los datos considerados y que la misma es moderadamente significativa. En otras palabras, entre mayor sea la nota obtenida en la PHC, mayor es la nota de aprobación en el curso de Cálculo I.

Se concluye entonces que los resultados sobre la correlación de los datos de las publicaciones previamente citadas, se mantienen también para la muestra abordada en este estudio.

4.6.2. La prueba de la hipótesis

Para este punto del proceso de análisis de los datos, es de interés comparar la proporción de personas estudiantes que aprueban el curso de Cálculo I entre aquellos

que aprobaron la PHC con la nota de corte establecida y aquellos que no aprobaron. En caso de que la proporción de aprobados fuera significativamente mayor en el primer grupo, entonces, se podría llegar a la conclusión de que la PHC es un buen predictor del éxito en el curso MA1001.

Considerando que este análisis se centra en una comparación de proporciones, se recurrirá a la prueba de proporciones incluida en RStudio. Los datos que requiere la función corresponden a los aprobados y reprobados en la PHC y en Cálculo I y se han organizado según lo descrito en la siguiente tabla:

Tabla 13

Resumen de los datos de aprobación en la PHC y Cálculo I

		Prueba de Habilidades Cuantitativas				
		Aprobados	Reprobados	Total		
Cálculo I	Aprobados	50 (VP)	29 (FN)	79	Sensibilidad:	63,29%
	Reprobados	13 (FP)	37 (VN)	50	Especificidad:	74,00%
	Total	63	66	129		

Para realizar la prueba, es necesario establecer dos hipótesis que serán las que se validen con el proceso de realización de la misma. En este caso, la hipótesis nula y alternativa a considerar, serán las siguientes:

- Hipótesis nula: No hay diferencia entre la proporción de aprobados en Cálculo I entre aquellos que aprobaron y no aprobaron la PHC según la nota establecida en este estudio.

- Hipótesis alternativa: Existe una asociación entre quienes aprueban la PHC y quienes aprueban Cálculo I.

Adicionalmente, el nivel de significancia que se consideró para la prueba fue de 0,05, valor que se asume como el habitual para este tipo de procesos.

Una vez ejecuta la prueba para los datos considerados, se obtuvo como resultado un *p-value* de 0,00007901, el cual es evidentemente inferior al nivel de significancia deseado de 0,05. Por ello, es posible afirmar con un nivel de significancia del 5%, que la aprobación de la PHC según la nota definida en este estudio se asocia con la aprobación del curso de Cálculo I.

Retomando los datos expuestos en la Tabla 13, es posible constatar la presencia de las siguientes categorías de examinados:

- Verdaderos Positivos **(VP)**: quienes aprobaron la PHC y también aprobaron Cálculo I. En este caso son 50 personas.
- Verdaderos Negativos **(VN)**: las 37 personas que reprobaron tanto la PHC como Cálculo I.
- Falsos Positivos **(FP)**: aquellos 13 que aprobaron la PHC y reprobaron Cálculo I.
- Falsos Negativos **(FN)**: las 29 personas que reprobaron la PHC y aprobaron Cálculo I.

Estos valores son importantes para poder abordar dos aspectos fundamentales al establecer puntos de corte en pruebas estandarizadas, como lo son la sensibilidad y la especificidad.

La primera de ellas refiere a la proporción de personas que realmente poseen el nivel de habilidad para aprobar Cálculo I y que en efecto son identificadas al aprobar la PHC. En otras palabras, qué tan bien identifica la prueba los verdaderos positivos. Su valor obtiene con el siguiente cálculo: $\frac{VP}{VP+FN} = \frac{50}{79} = 63{,}29\%$.

Por otra parte, la especificidad indica la proporción de las personas que no poseen las habilidades para aprobar Cálculo I y que son identificadas por la PHC (quienes la reprueban), esto refiere al reconocimiento de los verdaderos negativos. Para calcularla, basta con realizar la siguiente operación: $\frac{VN}{FP+VN} = \frac{37}{50} = 74{,}00\%$.

Ambos valores tienen un rango que va desde el valor mínimo de 0% hasta un máximo de 100%, es por ello que al tener en cuenta los valores que se han indicado, se pueden interpretar como moderadamente buenos.

Ahora, considerando las conclusiones a las que se ha llegado hasta este punto, es importante brindar un espacio a la revisión de los elementos que hacen que esta investigación cuente con insumos para asegurar la validez en las interpretaciones que se hagan de los resultados obtenidos.

4.7. Aspectos de calidad del proceso de fijación de puntos de corte

Para este apartado se ha considerado el organizarlo en dos secciones, la primera de ellas refiere a la fundamentación teórica en la que se basó el proceso investigativo, mientras que la segunda, se centra en los aspectos metodológicos que aportan a la calidad del proceso desarrollado.

4.7.1. Fundamentación teórica

Al inicio de este documento se abordó en detalle el posicionamineto teórico que sentó las bases para el desarrollo de este proceso. Entre los elementos que fueron considerados, se incluyó la selección y el análisis de los métodos para fijar los puntos de corte.

Esta selección consideró los antecedentes en el uso de cada método, para poder replicar el uso de los mismos en el escenario de esta investigación, a su vez, se tomó nota de la investigaciones previas sobre las limitaciones que se tuvieron y se hizo lo posible por atenderlar o limitarlas en el caso de esta investigación. Por ello, se puede hacer referencia a la validez inherente de la aplicación de los dos métodos para la fijación de los puntos de corte.

Por otro lado, la teoría indicaba que era apropiado realizar este proceso con al menos dos métodos para fijar el punto de corte y por ello, en este caso se optó por integrar el método bookmark y el Angoff para esta tarea. Es oportuno destacar que la diferencia entre ambos métodos no se limita a algo superficial y tampoco son una variación entre ellos, sino que los mismos se diferencian en su fundamentación, en su aplicación y hasta en la manera en la que conciben la prueba, es decir, en el Angoff se considera cada ítem como una unidad independiente y en el bookmark se considera la prueba como un todo.

Otro aspecto que fundamenta la calidad de las conclusiones de esta investigación, es la consideración de distintos insumos para construir el perfil de la persona mínimamente competente que debía aprobar la PHC. Este perfil consideró los insumos teóricos de los que dispone el proyecto de la PHC para su construcción y a la vez, fue validado por miembros del equipo de la prueba, esto no sólo habla de la validez del

proceso, sino de la contextualización de los resultados del estudio para las necesidades de la prueba.

4.7.2. Elementos metodológicos

En lo que refiere a los elementos metodológicos, uno de los aspectos vitales fue la conformación del equipo de jueces. El perfil del jurado fue específico y requería que además de una formación básica como docentes de matemática, también tuvieran experiencia impartiendo los cursos de Cálculo de la Universidad de Costa Rica, esto habla de su especialidad en el área. Además, su compromiso fue vital, pues asumieron la tarea con seriedad y esto aportó a la conclusión exitosa del proceso desarrollado.

Por otro lado, si bien se fijaron algunos elementos para poder excluir a algunos posibles jueces y poder seleccionar a otros, la conformación final del equipo fue aleatoria y bastante variada. Gracias al proceso de selección del equipo de jueces se logró contar con representantes de ambos sexos, además de tener la posibilidad de incluir docentes de distintas sedes de la Universidad de Costa Rica que con su aporte enriquecieron el proceso que se desarrolló. Aunado a ello, con la variedad de docentes con los que se contó, hubo diferenciación en sus años de servicio, es decir, algunos integrantes tenían bastantes años de desarrollar los cursos en la Universidad, mientras que otros eran docentes más jóvenes.

En el caso del método bookmark, a partir de los resultados que fueron obtenidos de manera preliminar, fue posible desarrollar una ronda adicional con miembros del equipo de jueces, esto con el objetivo de lograr un consenso en los ítems que fueron seleccionados para ese método. Este proceso resultó exitoso y el lograr unificar los criterios del equipo de jueces que participó en el proceso, a la vez que facilitó el análisis,

también aseguró una mayor representatividad de la opinión del jurado en cuanto a los ítems seleccionados como los marcadores.

4.7.2.1. Evidencias de validez y confiabilidad de los usos del punto de corte

Para el caso del método, Angoff, se contó con dos rondas para poder establecer los valores deseados y tal y como se mostró en apartados anteriores, la segunda ronda fue la considerada, pues tanto la teoría que se indagó previamente como su aplicación en este caso específico, mostraron que era donde había menor variabilidad en el juicio emitido por los y las integrantes del equipo de jueces.

Otro aspecto del proceso que incide en su validez, responde al contenido, esto en tanto el equipo de jueces que participó en el proceso, consideró que los ítems que emplearon para establecer el punto de corte se concentraban en el constructo evaluado, cuando no era así, lo reflejaban penalizando el ítem por medio de la asignación de menores tasas de repuesta para la población meta.

Por otra parte, es necesario tener también en consideración que gracias a las pruebas estadísticas realizadas, se cuenta con elementos que suman a las evidencias de validez basada con otras variables. Ejemplo de ello es que existe una relación entre el éxito en ambos escenarios académicos, es decir, entre la aprobación de la PHC y del curso Cálculo I, esto muestra que en efecto las conclusiones de este proceso de fijación de los puntos de corte tiene un efecto esperable en la predicción de los resultados del curso.

Finalmente, uno de los resultados más satisfactorios de este proceso, fue la cercanía de los valores que se logró obtener en cada uno de los métodos la cual puede considerarse como evidencia de validez basada en las intepretaciones de los puntajes,

ya que los métodos realizados proporcionan don intepretaciones del puntaje obtenido que pueden justificar los usos pretendidos en esta investigación, pues el resultado de la misma “se extiende más allá de la interpretación o el uso de puntajes previsto por el desarrollador de la prueba” (American Educational Research Association, American Psychological Association y National Council on Measurement in Education, 2018, p. 20).

Si bien al final se realizó un redondeo, ya desde el resultado original ambos valores variaban en menos de una unidad. Este elemento que podría verse como una coincidencia, más bien dota de más sentido el valor que se obtuvo luego del proceso metodológico desarrollado.

5. Discusión de los resultados

Una vez que se han presentado y analizado los datos obtenidos como parte del proceso de esta investigación, se ha reservado este apartado para discutir algunos elementos particulares que se han evidenciado durante la planificación del proceso, la implementación de la metodología o al analizar los datos resultantes. Al cierre de este apartado, la persona lectora podrá encontrar algunas conclusiones y recomendaciones que surgen del desarrollo de este proyecto.

5.1. Importancia de contar con un punto de corte definido científicamente

Uno de los elementos que justificó el iniciar con esta investigación, fue el interés por poner en práctica una metodología clara y reconocida que permitiera complementar el análisis de los resultados que se realiza una vez aplicada la PHC en cada cohorte. Sin embargo, este interés no surgió de la nada, sino que se justificaba con la posibilidad de aprovechar los métodos elegidos para fijar un punto de corte para una prueba considerando el criterio de expertos, con la aplicación de un método claro y hasta repetible en otros escenarios, es decir, sería definido científicamente.

En la indagación previa que se realizó para el desarrollo de este proyecto, se encontró que en el país hay otras pruebas que se consideran estandarizadas y de altas consecuencias, que poseen una nota límite para la aprobación de las mismas que no necesariamente responden a un criterio científico. Entre estas se puede mencionar a las pruebas prácticas y teóricas que se aplican como parte de los procesos para la obtención de las licencias de conducir del Consejo de Educación Vial (artículos 217 y 221 de la ley de tránsito, así como decreto 138 del MOPT), también el filtro que aplica el Servicio Civil

para la selección de las personas mejor capacitadas para desempeñar un puesto público, los puntos de corte del sistema educativo (65 o 70), las notas mínimas que se solicitan en las pruebas que se realizan para ingresar e colegios profesionales, entre otros más.

La posibilidad de desarrollar procesos más transparentes y con una metodología con fundamento científico, podría ayudar en el logro de los objetivos que se han propuesto para cada una de las pruebas que se han ejemplificado, sin que los valores empleados respondan únicamente a un criterio de tradición o de una disposición administrativa que no considera los efectos que puede tener en la integralidad del proceso para el que se construyeron las pruebas.

Adicionalmente, es importante mencionar la validez basada en las consecuencias. Al respecto, esta investigación contribuye a establecer criterios más claros y sólidamente fundamentados para la utilización de los resultados de la PHC. Al definir un punto de corte validado estadísticamente y basado en el perfil de habilidad requerido para el éxito académico (como se evidencia en su relación con el desempeño en Cálculo I), se busca reducir los efectos negativos que podrían derivar de la aplicación de criterios arbitrarios. De este modo, el estudio realizado promueve una toma de decisiones más justa y efectiva en los procesos de selección y orientación académica, lo que a su vez puede impactar positivamente en la trayectoria y el rendimiento académico del estudiantado universitario.

5.2. Las interpretaciones del punto de corte establecido

Uno de los elementos primordiales en los objetivos de este estudio, fue el poder ofrecer una interpretación de lo que significa lograr aprobar la PHC según la nota fijada como resultado de la aplicación de los métodos de punto de corte, pues el proceso

desarrollado no sólo busca asignar un valor, sino que intenta dotar de sentido al número que se encontró.

Tomando en cuenta el proceso para el establecimiento de puntos de corte según la metodología Angoff (Angoff, 1971) y considerando que la nota de corte de 67,5 que se obtuvo en este estudio, este valor implica que para aquellas personas que poseen el perfil deseado, es esperable que, en promedio, tengan un total de 67,5% de aciertos en la prueba. Claramente, aquellas personas que no cumplen con el perfil deseado, obtendrán un promedio de aciertos inferior al que se ha mencionado previamente.

Aunado a lo anterior, y desde la metodología bookmark (Cizek y Bunche, 2007), aquellos que cumplan con el perfil deseado poseen un nivel de habilidad que les asegura una alta probabilidad de responder correctamente el 67,5% de los ítems de la prueba.

Estas interpretaciones justifican el que se pueda decir que las personas estudiantes que superan la PHC poseen las habilidades mínimas requeridas para poder afrontar el curso de Cálculo I con éxito, esto quiere decir que a criterio del equipo de jueces, pueden desarrollar los procesos para establecer relaciones, validar, ejemplificar, generalizar y clasificar con el nivel mínimo que se requiere para enfrentar los contenidos incluidos en el curso de Cálculo I.

Lo anterior significa que serán capaces de establecer relaciones sencillas, representar situaciones gráficamente, realizar ejercicios por medio de la prueba y el error y el aplicar fórmulas. Además, poseen un nivel aceptable para desarrollar operaciones básicas, pueden construir contrajemplos para determinada propiedad, conocen los conjuntos numéricos estudiados en la Educación Secundaria y también pueden reconocer patrones en distintos contextos.

Estas interpretaciones no son antojadizas, sino que se apoyan en un proceso de indagación teórica detallado y acompañado de una minuciosa implementación de los procesos que en conjunto con la selección de un adecuado equipo de jueces, permite asegurar una buena calidad en el proceso de fijación de los puntos de corte.

Al respecto, es importante recordar que el proceso seguido cuenta con un respaldo teórico que refiere a la indagación sobre los antecedentes de implementación de los métodos bookmark y Angoff, sus aciertos, desaciertos y limitaciones, esto con el fin de tomarlos en cuenta para robustecer el proceso seguido en esta investigación.

Adicionalmente, el equipo de jueces que colaboró en el estudio cuenta con un perfil académico y profesional que respalda los juicios que emitieron de manera individual y las interpretaciones realizadas para los resultados colectivos. En complemento, los valores obtenidos tras la aplicación de las pruebas estadísticas del apartado anterior, brindan mayor relevancia a las interpretaciones realizadas en este proceso.

Adicionalmente, es importante recordar que la enseñanza de la matemática posee varios fines que justifican su inclusión en los distintos niveles del proceso formativo (Ministerio de Educación Pública, 2012), uno de ellos refiere propiamente a su aplicación práctica en las distintas disciplinas, pero también se incorpora en distintos currículos por las habilidades que permite desarrollar y que van más allá de la mera aplicación práctica de los conceptos matemáticos y sus algoritmos. Esto implica para las carreras universitarias que aplican la PHC, que las personas estudiantes que superan ese examen, no solo tienen mayor probabilidad de aprobar el curso de Cálculo I en el primer intento, sino que además, cuentan con las habilidades que ya se han mencionado

previamente y que se pueden explotar en los otros cursos que componen la estructura curricular de esas carreras.

Por ejemplo, una persona que supere la PHC no solo debería tener éxito al cursar MA1001, sino que también las habilidades demostradas le deberían ser útiles en otros cursos de ciencias básicas como Química o Física, así que los resultados de la prueba no sólo se limitan al área matemática, sino que inciden en el posible éxito en otras áreas que requieran la aplicación del razonamiento cuantitativo.

5.3. Las innovaciones metodológicas

Para el desarrollo de esta investigación, se realizaron algunas adaptaciones en los procesos empleados, los cuales a criterio del investigador, lejos de perjudicar el proceso, ayudan a enriquecerlo o a facilitar el rol que desempeñaron las y los jueces.

Estas modificaciones surgen a raíz de lo que se expone en aplicaciones previas de los métodos, por ejemplo los casos expuestos por Wyse (2017), Kampa et al. (2019a) y Wyse (2020).

En detalle, en el caso del método Angoff, los ítems se presentaron al equipo de jueces organizados de menor a mayor índice de dificultad, esto contribuyó a que los jueces tuvieran una idea más clara de que el porcentaje de la población que respondía correctamente debía disminuir, además, en el cuadernillo que incorporaba los dos métodos, fueron conscientes de que el ítem bookmark debía localizarse cerca (o coincidir) con el primer ítem en el que el porcentaje de la población de interés que lo acertaría fuera menor al 66%. En la aplicación tradicional del método, los ítems no responden a este orden.

Por otra parte, el método bookmark considera la prueba como un todo y se debe seleccionar un único ítem por parte del equipo de jueces, sin embargo, en el contexto de esta investigación y siendo conscientes de que se valoraban distintas habilidades, se optó por trabajarlas de manera separada y en congruencia con el perfil que se construyó para la persona con el nivel mínimamente deseable, esto ayudó al equipo de jueces para saber qué era lo que se buscaba en ese individuo que debía aprobar el examen.

Ahora bien, al ser objetivos, el tomar esta decisión tiene sus ventajas y desventajas. A favor se cuenta con que la división en secciones asegura que cada habilidad estará concentrada en un único apartado y que no será posible que en otras secciones se le pida resolver ítems que superan el nivel que posee en determinada habilidad. En contra, puede suceder que algunas secciones requieran demostrar su nivel de habilidad solamente con ítems considerados sencillos, lo que puede influir en la nota de corte obtenida.

En resumen, pese a la desventaja señalada, la división de la prueba por habilidades resolvió una necesidad del proceso de fijación del punto de corte de la PHC y fue el que los examinados debían demostrar su nivel para todas las habilidades y sin embargo, no se tiene la posibilidad de establecer varios puntos de corte para definir la entrada a las distintas carreras que aplican la PHC.

5.4. Posibilidades a futuro y limitaciones

El primer aspecto a considerar para esta sección, responde a algunas consideraciones metodológicas que se podrían valorar en otros escenarios en los que se desee implementar un proceso de establecimiento de puntos de corte.

Específicamente, para esta investigación se decidió que al momento de tomar el punto de corte indicado por el método bookmark, el ítem bookmark general de la prueba correspondería a aquel que tuviera la dificultad Rasch más baja entre los seleccionados de las distintas dimensiones del cuadernillo de la PHC. Sin embargo, también se pudo haber optado por elegir el ítem de mayor dificultad entre los ítems bookmark de la prueba, e incluso, se podría centrar la atención en las dos dimensiones de mayor presencia en el cuadernillo de la PHC analizado.

Sin embargo, al reproducir el proceso para los otros posibles reactivos, la nota de corte no sufrió una variación significativa, esto pues los niveles de habilidad requeridos eran relativamente cercanos.

Otra alternativa que se podría considerar en un nuevo proceso de fijación de los puntos de corte, es el aplicar los métodos sin considerar las dimensiones de la prueba, es decir, tomarla como una unidad y para los procesos de análisis, solamente considerar la dificultad de los ítems para organizarlos. Este aspecto podría facilitar la elección del ítem bookmark, pues sería solamente uno para todo el cuadernillo, pero en este caso no fue viable pues resultaba de importantancia conocer las consideraciones del equipo de jueces para diferenciar los niveles de habilidad en cada dimensión y poder establecer conclusiones para la habilidad del razonamiento matemático.

Por otra parte, otra consideración sería la posible variación en el número de reactivos de cada dimensión en otros cuadernillos de la prueba, pues en este caso, las últimas dos dimensiones analizadas (clasificar y generalizar) dificultaron algunos elementos del proceso por tener pocos representantes en el cuadernillo analizado y en

el caso de la dimensión ejemplificar, no hubo siquiera reactivos presentes de esa dimensión.

Retomando lo expuesto en el párrafo anterior, resulta de interés profundizar en lo acontecido con las dimensiones con menor representación en el cuadernillo de la PHC que se analizó, específicamente al momento de implementar el método bookmark. Las dimensiones clasificar y generalizar aportaron únicamente 4 y 2 reactivos de los 36 del cuadernillo. Sin embargo, según lo que se ha mencionado por parte de algunos integrantes del equipo de jueces, esto les dificultó el establecer el ítem bookmark, en coherencia con esto, en el caso del método Angoff, no existe una tendencia clara entre los reactivos de ambas dimensiones que permita evidenciar cuál marca la diferencia entre los niveles de competencia deseados para las personas que realizan la prueba. Es por ello, que en el desarrollo metodológico de esta investigación, se optó por centrar la atención en los resultados obtenidos de las primeras dos dimensiones que eran las más representativas en el cuadernillo de la prueba.

El equipo de jueces señaló que si esas categorías tuvieran más representantes, eso les hubiera facilitado su labor al fijar el ítem bookmark respectivo. En ambos casos, ese ítem marcador se encontraría fuera de los límites de los ítems presentes en el cuadernillo analizado.

Otro elemento que puede estar sujeto a discusión fue la elección de fijar un único límite para establecer los niveles de desempeño de las personas estudiantes, las cuales en este caso fueron las categorías competente e insuficiente, que se diferenciaban por quienes aprobaban la PHC en los términos de esta investigación y quienes no lo hacían.

Considerando la variedad de carreras que consideran la PHC como uno de sus requisitos complementarios para el ingreso a carrera (según los requisitos de ingreso actualizados en la resolución de la Vicerrectoría de Docencia VD-12961-2024), sería de interés valorar la posibilidad de complementar el proceso que se ha llevado a cabo con el establecimiento de puntos de corte diferenciados para las necesidades de las distintas carreras, esto tomando en cuenta las exigencias en cuanto a razonamiento matemático que consideran pertinente para sus carreras. Ahora bien, esta variación supondría que al momento de trabajar con el equipo de jueces, se debe señalar esta pauta y ya no bastaría con su entrenamiento considerando un único perfil de la persona con el nivel de mínimamente competente, sino que se deberían considerar varias opciones.

Claramente, esto haría más extenso el laborioso trabajo del equipo de jueces, pero existe la posibilidad de considerar más integrantes para este equipo y diferenciarlos por áreas y así evitar que se confundan al analizar los distintos requerimientos según cada tipo de perfil.

Adicionalmente, la calificación de 67,5 que se concluyó en esta investigación, resulta válida para el curso de Cálculo I (MA1001), pero este proceso se pueda adaptar para las necesidades de distintos cursos.

También es importante que desde el primer momento en que haya contacto con los jueces, se indique claramente cuál es la tarea por ejecutar, cuál es su rol y qué se pretende hacer con la información que brindarán. Además, es vital explicar detalladamente el perfil de la persona que aplicará la prueba y que se desea que cumpla con los requisitos para aprobar dicha prueba, así como también explicar en términos generales cómo se aplica cada uno de los métodos seleccionados para el proceso.

Durante las sesiones de trabajo con el equipo de jueces, fue posible observar cómo en reiteradas ocasiones consultaban el perfil de la persona que debían considerar, además de realizar consultas al investigador sobre detalles específicos de la tarea que se les asignó. Por ello, se destaca la importancia de que se puede ofrecer acompañamiento en dicho proceso.

5.5. Algunas conclusiones del proceso

Como las conclusiones más relevantes del proceso desarrollado, se consideran elementos que son resultados de la implementación de los métodos, pero también hay algunos que responden propiamente al planteamiento metodológico elegido.

En primer lugar, destaca la construcción del perfil de la persona competente en términos de los requerimientos de la PHC, sin bien esta tarea no se había considerado de manera explícita inicialmente, resultó fundamental para el desarrollo de todo el proceso de fijación de los puntos de corte.

Este perfil delimitó las competencias o habilidades que debía tener una persona estudiante que aprueba la PHC y esto implica elementos de orden declarativo y procedimental, con el detalle adicional de que esa persona competente debe poder aplicar eso conocimientos y procesos a distintas situaciones en marcos referenciales distintos.

Otra conclusión de este proceso fue el establecer la nota de punto de corte, que aunque podría haber sido más de una, el proceso desarrollado permitió obtener un resultado consistente en la aplicación de ambos métodos. El valor obtenido facilita la tarea de diferenciar quién tendrá mayor posibilidad de aprobar el curso de MA1001 al ingreso a la universidad en función de la nota que haya obtenido en la PHC.

Como resultado de la nota de punto de corte que se logró fijar, también se puede concluir que existen al menos dos niveles de habilidad que se pueden diferenciar al considerar esa nota como límite entre categorías. Por una parte el nivel competente que es aquella persona con las capacidades básicas de análisis y de conocimientos matemáticos que podrá enfrentar con éxito el curso MA1001 y por otra, aquella que tendrá mayor dificultad para hacerlo.

Es relevante que aunque en el proceso de fundamentación y construcción de la PHC descrito por Rojas y Ordoñez (2019), es claro que los contenidos del área matemática estudiados en secundaria que se incluyen en la prueba son de conocimiento muy básico y que se limitan a lo estudiado en noveno año con el objetivo de no generar sesgos por las diferencias que pudieran haber en la educación diversificada, el equipo de jueces indica que sí hay influencia de esos conocimientos en los niveles de dificultad de los reactivos. Ejemplo de esto es el que consideraron que el operar con porcentajes, o con número racionales en distintas representaciones dificulta significativamente el responder correctamente ese reactivo.

En lo que refiere a la validez predictiva del punto de corte, con el apoyo de pruebas estadísticas se logró comprobar que existe una conexión entre la aprobación o no de la PHC con la nota obtenida como punto de corte y la aprobación o reprobación del curso de Cálculo I y además, esta confirmación de la hipótesis resultó estadísticamente significativa. Es decir, una persona que realizó la PHC y obtuvo una nota superior a 67,5, tendrá una alta probabilidad de enfrentar ese curso con las herramientas requeridas para aprobarlo.

En cuanto a los elementos que se consideraron para construir y desarrollar esta investigación, como lo fue la elección de los métodos para establecer los puntos de corte, la integración de dos de ellos para realizar esta tarea, así como la elección de un equipo de jueces selecto y apropiado a las necesidades del proceso, los mismos dan cuenta de la validez de los resultados de esta investigación y de las interpretaciones que se desprenden de los mismos.

En síntesis, el proceso llevado a cabo permitió la construcción de un perfil de la persona competente, el establecimiento de un punto de corte predictivo para el éxito en Cálculo I y la validación de la metodología empleada. No obstante, la reflexión sobre las limitaciones del estudio y las oportunidades de mejora metodológica abren paso a una serie de recomendaciones para futuras investigaciones y aplicaciones prácticas.

5.6. Recomendaciones para investigaciones futuras

Considerando la aplicabilidad potencial de la metodología de esta investigación en diversas áreas, se recomienda explorar la adaptación y validación de este proceso en otros contextos educativos o para otras pruebas estandarizadas. Para ello, se debería replicar el estudio en diferentes poblaciones estudiantiles o en relación con la predicción del éxito en diferentes disciplinas, realizando los ajustes necesarios al perfil de competencias y a los métodos de establecimiento del punto de corte.

Es oportuno aclarar que una de las principales limitaciones de este estudio fue el considerar que los resultados fueran deterministas, es decir, para un proceso a futuro o para otros estudios similares se podría valorar la opción de considerar un margen de error apropiado para los valores de punto de corte que se obtengan, sin embargo, ello requiere

una análisis detallado para valorar qué tan amplio debe ser ese margen y las implicaciones que esto tendría en términos de la aplicación del punto de corte obtenido.

Por otra parte, si bien la integración de los dos métodos seleccionados para esta investigación proporcionó resultados consistentes, se sugiere explorar e incorporar metodologías adicionales para el establecimiento de puntos de corte en futuras investigaciones, esto podría enriquecer la validez y robustez del punto de corte final. Por ejemplo, se podría incluir métodos basados en el juicio de expertos con mayor estructuración o análisis basados en el impacto en las tasas de éxito en el curso objetivo de la prueba, mismo que no necesariamente debería ser Cálculo I.

Estas consideraciones que se han expuesto resultan de interés para aquellos evaluadores o investigadores que tengan algún grado de interés en desarrollar procesos similares para pruebas estandarizadas. Adicionalmente, las consideraciones metodológicas se basan en la revisión de varias investigaciones previas y de la experiencia específica en el desarrollo de este proyecto. Claro que es una actividad sujeta a la mejora, pero las recomendaciones y conclusiones son aplicables a otros procesos afines en diversas áreas.

6. Referencias

Al-Musawi, N. (2016). Using Item Response Models to Develop a Criterion—Referenced Test to Measure the Students' Achievement in Educational Evaluation. *Journal of Educational & Psychological Studies*, *10*(4), 727–736. Education Source. https://doi.org/10.24200/jeps.vol10iss4pp727-736

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

Angoff, W. H. (1971). Scales, norms, and equivalent scores. En R. Thorndike (Ed.), *Educational measurement* (pp. 508–600). DC:American Council on Education.

Arribas, J. M. (2017). La evaluación de los aprendizajes. Problemas y soluciones. *Profesorado. Revista de Currículum y Formación de Profesorado*, *21*(4), 381–404.

Baldwin, P., Margolis, M. J., Clauser, B. E., Mee, J., y Winward, M. (2020). The Choice of Response Probability in Bookmark Standard Setting: An Experimental Study. *Educational Measurement: Issues & Practice*, *39*(1), 37–44. Education Source. https://doi.org/10.1111/emip.12230

Benítez, J. E., y López, M. D. (2018). La aplicación de la evaluación de los aprendizajes. Un estudio en la Universidad Bolivariana de Venezuela. *Revista Arbitrada Interdisciplinaria Koinonía*, *3*(5), 67–83.

Bolsover, G. (2018). Slacktivist USA and Authoritarian China? Comparing Two Political Public Spheres With a Random Sample of Social Media Users. *Policy and internet*, *10*(4), 454–482. https://doi.org/10.1002/poi3.186

Bracho, R. (2022). Estimación estadística basada en un diseño muestral multietápico (estudio de MYPES). *Tecnociencia*, *24*(1), 5–21.

Buckendahl, C., Smith, R., Impara, J., y Plake, B. (2002). A comparison of Angoff and Bookmark standard setting methods. *Journal of Educational Measurement*, *39*, 253–264.

Calua, J. (2016). *Potencia predictiva de variables académicas en el rendimiento académico de estudiantes universitarios del primer ciclo-2015-1. Caso de la universidad privada del Norte-Cajamarca* [Tesis de Doctorado, Universidad Nacional de Cajamarca]. http://190.116.36.86/bitstream/handle/UNC/1348/T016_26691613_D.pdf?sequence=1&isAllowed=y

Cano, Y., Obaco, E. E., Delgado, L., Herrera, G. N., y Romero, J. M. (2021). Trabajo por niveles de desempeño cognitivo en el contexto ecuatoriano: ¿Alternativa o necesidad? *Tendencias pedagógicas*, *38*(38), 112–123. https://doi.org/10.15366/tp2021.38.010

Carballo, M., y Guelmes, E. L. (2016). Algunas consideraciones acerca de las variables en las investigaciones que se desarrollan en educación. *Revista Universidad y Sociedad*, *8*(1), 140–150.

Carvalho, M. da P. (2018). A avaliação da aprendizagem em uma perspectiva teórica e prática. *ARANDU UTIC*, *5*(1), 247–262.

Cizek, G., y Bunch, M. (2007). *Standard Setting: A guide to establishing and evaluating performance standards on tests*. SAGE Publications.

Clauser, B. E., Baldwin, P., Margolis, M. J., Mee, J., y Winward, M. (2017). An Experimental Study of the Internal Consistency of Judgments Made in Bookmark Standard Setting. *Journal of Educational Measurement*, *54*(4), 481–497. Education Source. https://doi.org/10.1111/jedm.12157

Contreras, C., y Campa, R. (2022). Diseño instrumental y validación de un cuestionario para la competencia informacional en estudiantes universitarios. *Sinéctica, Revista Electrónica de Educación*, *59*. https://doi.org/10.31391/S2007-7033(2022)0059-015

Contreras, J., Ramírezparis, X., y Hernández, V. (2019). Factores que influyen en el desempeño escolar de los estudiantes de Básica Primaria de una institución educativa del área metropolitana de Cúcuta. *Perspectivas*, *4*(1), 6–13.

Cristancho, J. G. (2019). La pregunta por los fundamentos epistemológicos de la investigación en la educación -aportes para una discusión epistemológica-.

Cuestiones De Filosofía, *5*(25), 37–61. https://doi.org/10.19053/01235095.v5.n25.2019.10681

Cuesta, H., Aguiar, M. V., y Marchena, M. R. (2015). Desarrollo de los razonamientos matemático y verbal a través de las TIC: descripción de una experiencia educativa. *Pixel-Bit. Revista de Medios y Educación*, *46*, 39–50. Redalyc. http://dx.doi.org/10.12795/pixelbit.2015.i46.03

Demarchi, G. D. (2020). La evaluación desde las pruebas estandarizadas en la educación en Latinoamérica. *En-Contexto*, *8*(13), 107–133.

Durán, F. (2019). Pruebas estandarizadas para el acceso a la Educación Superior en Chile: Performatividad y subjetividad de los estudiantes. *Calidad en la Educación*, *50*, 180–215. Education Source. https://doi.org/10.31619/caledu.n50.723

Ebel, R., y Frisbie, D. (1991). *Essentials of Educational Measurement* (5ta ed.). Prentice-Hall.

Escorcia, D., Moreno, M., Campo, K., y Palacio, J. (2014). Enseñanza y evaluación de la escritura en la universidad: Análisis de prácticas declaradas de docentes franceses y colombianos. *Zona próxima*, *20*, 92–107. https://doi.org/10.14482/zp.20.6007

Feseker, T., Gnambs, T., y Artelt, C. (2021). Setting a standard for low reading proficiency: A comparison of the bookmark procedure and constrained mixture Rasch model. *PLoS One*, *16*(11), e0257871–e0257871. https://doi.org/10.1371/journal.pone.0257871

Flores, C., Mena, C., Arteaga, P., Navarrete, L., y Gajardo, A. (2018). Nivel de desempeño autopercibido por futuras educadoras de párvulos sobre el uso de TIC. *Panorama*, *12*(22), 18–30. https://doi.org/10.15765/pnrm.v12i22.1070

Galeano, M. (2021). *Investigación cualitativa: Preguntas inagotables* (1a ed.). Universidad de Antioquia. https://doi.org/10.2307/j.ctv1pdrq25

Gallardo, M., Sierra, J. E., y Domínguez, A. (2015). El Portafolios de los Estudiantes como Estrategia Alternativa a las Pruebas Estandarizadas para la Evaluación de Competencias. (Spanish). *Qualitative Research in Education*, *4*(1), 71–101. Education Source. https://doi.org/10.4471/qre.2015.57

García, F. J., Pozuelos, F. J., y Alvarez, C. (2019). La Evaluacion de Aprendizajes del Alumnado por parte del Profesorado Universitario Novel. *Form. Univ*, *12*(2), 3–16. https://doi.org/10.4067/S0718-50062019000200003

García, L. (2020). *Algunas tipologías de evaluación*. Contextos universitarios mediados. https://aretio.hypotheses.org/4148

Garduno, R. (2009). Contenido educativo en el aprendizaje virtual. *Investig. bibl*, *23*(47), 15–44.

Giménez, J. A., Morales, F. J., y Parra, D. (2021). La utilización de instrumentos de evaluación en Educación Primaria: Análisis de caso en centros educativos de la provincia de Valencia (España). *Educatio siglo XXI: revista de la Facultad de Educación*, *39*(2), 193–212. https://doi.org/10.6018/educatio.483481

Gonzalez, A. O., Vasconez, B. E., Moso, G. M., y Sanguña, E. S. (2017). Conocimiento pedagógico en evaluación educativa de los docentes y la construcción de exámenes del idioma inglés. *Dominio de las Ciencias*, *3*(3), 374–389.

González, D. K., Cárdenas, G. A., y González, G. A. (2020). Estudio contrastivo de los modelos evaluativos de aula de instituciones públicas y privadas de la ciudad de Tuluá y su relación con índices de calidad educativa. *Boletín Redipe*, *9*(10), 195–211.

Guevara, R. (2017). La calidad, las competencias y las pruebas estandarizadas: Una mirada desde los organismos internacionales. *Educación y Ciudad*, *33*, 159–170. Education Source. https://doi.org/10.36737/01230425.v0.n33.2017.1658

Gutiérrez, J. G., y Acuña, L. A. (2020). Standardized evaluation of learning at UABC: Psycometric analysis innovation. *Apertura: revista de innovación educative*, *12*(1). https://doi.org/10.32870/Ap.v12n1.1698

Hernández, M., Ramírez, É., y Gamboa, S. (2018). La implementación de una evaluación estandarizada en una institución de educación superior. *Innovación Educativa*, *18*(76), 149–170.

Hernández, V. J. (2017). *Sistema de evaluación inteligente para medir habilidades de razonamiento matemático* [Tesis para optar al grado de Magíster en Ciencias de la Computación]. Benemérita Universidad Autónoma de Puebla.

Herrera Arroyo. (2016). *Folleto de práctica: Prueba de Habilidades Cuantitativas*. https://www.kerwa.ucr.ac.cr/bitstream/handle/10669/77397/Folleto%20práctica.pdf?sequence=1&isAllowed=y

Illesca, R. S., y Alfaro, J. E. (2017). Aptitud física y habilidades cognitivas. *Revista andaluza de medicina del deporte*, *10*(1), 9–13. https://doi.org/10.1016/j.ramd.2016.04.004

Instituto Colombiano para la Evaluación de la Educación. (2020). *Establecimiento de estándares de desempeño: Descripción de niveles y puntos de corte* (J. C. Gómez-Barrera, Ed.).

Jalalizadeh, M., Delavar, A., Farokhi, N., y Askari, M. (2019). Comparison of ANGOF-based IRT method and Bookmark method for standard Setting of MSRT language test. *Research in Teaching*, *7*(4), 69–49. https://doi.org/10.34785/J012.2019.698

Jiménez, K., y Montero, E. (2013). Aplicación dle modelo de Rasch, en el análisis psicométrico de una prueba de diagnóstico en matemática. *Revista Digital: Matemática, Educación e Internet*, *13*(1), 1–24.

Jordán, Á., Morán, L., y Camacho, G. (2018). La evaluación de los aprendizajes y su influencia en la calidad del proceso de enseñanza aprendizaje en el contexto universitario. *Opuntia Brava*, *9*(1), 215–224.

Jornet, J. M., Perales, M. J., y González, J. (2020). The concept of validity of teaching evaluation processes. *Revista española de pedagogía*, *78*(276), 233–252.

Kampa, N., Wagner, H., y Köller, O. (2019a). The standard setting process: Validating interpretations of stakeholders. *Large-scale Assessments in Education*, *7*(1), 3. https://doi.org/10.1186/s40536-019-0071-8

Kampa, N., Wagner, H., y Köller, O. (2019b). The standard setting process: Validating interpretations of stakeholders. *Large-scale assessments in education*, *7*(1), 1–25. https://doi.org/10.1186/s40536-019-0071-8

Kane, M. (2006). Validating the Interpretations and Uses of Test Scores. *Journal of Educational Measurement*, *43*(1), 1–38.

Karsenty, R. (2014). Mathematical Ability. En S. Lerman (Ed.), *Encyclopedia of Mathematics Education* (pp. 372–375). Springer Netherlands. https://doi.org/10.1007/978-94-007-4978-8_94

Lestari, W. y Jailani. (2018). Enhancing an Ability Mathematical Reasoning through Metacognitive Strategies. *Journal of Physics: Conference Series*, *1097*, 012117. https://doi.org/10.1088/1742-6596/1097/1/012117

Lewis, D., y Cook, R. (2020). Embedded Standard Setting: Aligning Standard-Setting Methodology with Contemporary Assessment Design Principles. *Educational measurement, issues and practice*, *39*(1), 8–21. https://doi.org/10.1111/emip.12318

Livingston, S., y Lewis, C. (1995). Estimating the consistency and accuracy of classifications based on test scores. *Journal of Educational Measurement*, *32*, 179–197.

Losada, A., Zambrano, C., y Marmo, J. (2022). Clasificación de métodos de investigación en Psicología. *Revista Psicología UNEMI*, *6*(11), 13–31. https://doi.org/10.29076/issn.2602-8379vol6iss11.2022pp13-31p

Lyons, I. M., y Ansari, D. (2015). Foundations of Children's Numerical and Mathematical Skills. En J. B. Benson (Ed.), *Advances in Child Development and Behavior* (Primera, Vol. 48). Academic Press.

Mahias, P., y Polloni, M. del P. (2019). *Desarrollo de instrumentos de evaluación; pruebas* (M. R. García, Ed.). Centro de Medición MIDE UC.

Mair, P., y Hatzinger, R. (2007). Extended Rasch Modeling: The eRm Package for the Application of IRT Models in R. *Journal of Statistical Software*, *20*(9), 1–20. https://doi.org/10.18637/jss.v020.i09

Martínez, M. (2021). Análisis factorial confirmatorio: Un modelo de gestión del conocimiento en la universidad pública. *Revista Iberoamericana para la Investigación y el Desarrollo Educativo*, *12*(23). https://doi.org/10.23913/ride.v12i23.1103

Martínez, M., Soberanes, A., y Sánchez, J. M. (2017). Análisis correlacional de competencias matemáticas de pruebas estandarizadas y pre-requisitos

matemáticos en estudiantes de nuevo ingreso a Ingeniería en Computación. *RIDE. Revista Iberoamericana para la Investigación y el Desarrollo Educativo*, *8*(15), 946–974. https://doi.org/doi.org/10.23913/ride.v8i15.328

Mayes, R. (2019). Quantitative Reasoning and Its Rôle in Interdisciplinarity. En B. Doig, J. Williams, D. Swanson, R. Borromeo Ferri, y P. Drake (Eds.), *Interdisciplinary Mathematics Education: The State of the Art and Beyond* (pp. 113–133). Springer International Publishing. https://doi.org/10.1007/978-3-030-11066-6_8

Medina, M. del R., y Verdejo, A. L. (2020). Validez y confiabilidad en la evaluación del aprendizaje mediante las metodologías activas. *ALTERIDAD. Revista de Educación*, *15*(2), 270–284.

Mena, L. (2018). La muestra cualitativa en la práctica: Una propuesta. *Revista Eixo*, *7*(3).

Minervino, C. A. da S. M., y Dias, É. B. (2017). Teste de habilidades preditoras da leitura: Normas de habilidade para crianças. *Avaliação psicologica*, *16*(4), 415–425.

Ministerio de Educación Pública. (2012). *Programas de estudio. Matemática*. MEP.

Mitzel, H., Lewis, D., Patz, R., y Green, D. (2001). The Bookmark procedure: Psychological perspectives. En *Setting performance standards: Concepts, methods, and perspectives* (pp. 249–281). Mahwah, NJ: Lawrence Erlbaum.

Mora, M., y Rojas, L. (2023a). Procesos de análisis para resolver problemas de razonamiento cuantitativo: Un estudio de respuestas de estudiantes. *Revista de Estudios e Investigación en Psicología y Educación*, *10*(1), 38–60. https://doi.org/10.17979/reipe.2023.10.1.9389

Mora, M., y Rojas, L. (2023b). Procesos de análisis para resolver problemas de razonamineto cuantitativo: Un estudio de respuestas de estudiantes. *Revista de Estudios e Investigación en Psicología y Educación*, *10*(1), 38–60. https://doi.org/10.17979/reipe.2023.10.1.9389

Ordóñez, G. (2023). *Habilidades de Razonamiento Cuantitativo Requerido por Estudiantes de Química en la Universidad de Costa Rica* [Tesis de Doctorado]. Universidad de Costa Rica.

Ordóñez, G., y Rojas, G. (2024). Estudio observacional del razonamiento cuantitativo en Química. *Revista de Investigación y Evaluación Educativa*, *11*(2), 7–24. https://doi.org/10.47554/revie.vol11.num2.2024.pp7-24

Orlando, M. (2014). *Razonamiento, solución de problemas matemáticos y rendimiento académico* [Tesis doctoral, Universidad de San Andrés]. https://repositorio.udesa.edu.ar/jspui/bitstream/10908/10908/1/%5BP%5D%5BW%5D%20T.%20D.%20Edu.%20Orlando%2C%20Mario.pdf

Ozakbas, S., Yigit, P., Akyuz, Z., Sagici, O., Abasiyanik, Z., Ozdogar, A. T., Kahraman, T., Bozan, H. R., y Hosgel, I. (2021). Validity and reliability of “Cognitive Reserve Index Questionnaire” for the Turkish Population. *Mult Scler Relat Disord*, *50*, 102817–102817. https://doi.org/10.1016/j.msard.2021.102817

Páramo, D., Campo, S., y Maestre, L. (2020). *Métodos de investigación cualitativa: Fundamentos y aplicaciones* (Primera edición.). Editorial Unimagdalena.

Parcerisa, L., Villalobos, C., Santa Cruz, E., y Saura, G. (2022). Movimientos Anti-Estandarización en Educación: Un Análisis Comparado de Chile, España y México. *Revista Izquierdas*, *51*, 1–22.

Park, J., Yim, M. K., Kim, N. J., Ahn, D. S., y Kim, Y.-M. (2020). Similarity of the cut score in test sets with different item amounts using the modified Angoff, modified Ebel, and Hofstee standard-setting methods for the Korean Medical Licensing Examination. *J Educ Eval Health Prof*, *17*(0), 28–0. https://doi.org/10.3352/jeehp.2020.17.28

Pérez, D. D., Cantera, L., y Pereyra, G. A. (2020). La evaluación formativa: Aportes para pensar la universidad latinoamericana desde un caso uruguayo. *Educación física y deporte*, *39*(1). https://doi.org/10.17533/udea.efyd.v39n1a08

Pérez, M., y Valmaseda Balanzategui, M. (2019). Evaluación de la LSE. Cuatro pruebas en desarrollo: Inventario MacArthur-Bates (CDI), prueba de vocabulario, RST-test de habilidades receptivas y PT-test de habilidades narrativas. *REVLES: Revista de estudios de lenguas de signos*, *1*, 209–237.

Popham, J. (1983). *Evaluación basada en criterios*. Magisterio Español.

Popham, W. J. (2014). Criterion-Referenced Measurement: Half a Century Wasted? *Educational Leadership*, *71*(6), 62–68. Education Source.

R Core Team. (2022). *R: A language and environment for statistical computing* [Software]. R Foundation for Statistical Computing, Vienna. https://www.R-project.org/

Rahmi, N., y Surya, E. (2017). An analysis of students' mathematical reasoning ability in VIII grade of Sabilina Tembung Junior High School. *IJARIIE*, *3*(2), 3527–3533.

Rasch, G. (1960). *Studies in mathematical psychology: I. Probabilistic models for some intelligence and attainment tests*. Nielsen & Lydiche.

Reguant, M., y Martínez-Olmo, F. (2014). *Operacionalización de conceptos/variables*. Dipòsit Digital de la UB.

Rodríguez, P. (2017). Creación, Desarrollo y Resultados de la Aplicación de Pruebas de Evaluación basadas en Estándares para Diagnosticar Competencias en Matemática y Lectura al ingreso a la Universidad. *Revista Iberoamericana de Evaluación Educativa*, *10*(1), 89–107. https://doi.org/10.15366/riee2017.10.1.005

Rojas, L. (2013). Predicción de la dificultad de la prueba de Habilidades Cuantitativas de la Universidad de Costa Rica. *Revista Digital: Matemática, Educación e Internet*, *13*(1), 1–14. https://doi.org/10.18845/rdmei.v13i1.1627

Rojas, L., Mora, M., y Ordóñez, G. (2019). Asociación del razonamiento cuantitativo con el rendimiento académico en cursos introductorios de matemática de carreras STEM. *Revista Digital: Matemática, Educación e Internet*, *19*(1). https://doi.org/10.18845/rdmei.v19i1.3851

Rojas, L., y Ordóñez, G. (2019). Proceso de construcción de pruebas educativas: El caso de la Prueba de Habilidades Cuantitativas. *Revista Evaluar*, *19*(2), 15–29. https://doi.org/10.35670/1667-4545.v19.n2.25080

Rojas-Torres, L. (2014). Predicción de la reprobación de cursos de matemática básicos en las carreras de Física, Meteorología, Matemática, Ciencias Actuariales y Farmacia. *Revista Electrónica EDUCARE*, *18*(3), 3–15. http://dx.doi.org/10.15359/ree.18-3.1

Ruano, A. L., Vizuete, A., Moreno, J. C., y Quispe, W. (2018). Habilitación profesional: Caso Ecuador. *Revista Ecuatoriana de Medicina Eugenio Espejo*, *7*(9), 15–23.

Ryan, A.-M., y Gass, S. (2017). Quantitative Reasoning: Exploring Troublesome Thresholds. *Discussions on University Science Teaching: Proceedings of the Western Conference on Science Education*, *1*(1).

Saiz, M. C., Escolar, M. del C., y Rodríguez, J. (2019). *Investigación cualitativa: Aplicación de métodos mixtos y de técnicas de minería de datos*. Burgos: Universidad de Burgos, Servicio de Publicaciones e Imagen Institucional.

Salgado, M., y Salinas, M. (2012). El razonamiento inductivo como generador de la construcción del número en 5 años. En *Investigaciones en Pensamiento Numérico y Algebraico e Historia de la Matemática y Educación Matemática* (pp. 119–125). Departamento de Didáctica de la Matemática de la Universitat de València y SEIEM.

Rojas, L., Mora, M., & Ordóñez, G. (2019). Asociación del razonamiento cuantitativo con el rendimiento académico en cursos introductorios de matemática de carreras STEM. *Revista Digital: Matemática, Educación e Internet*, *19*(1). https://doi.org/10.18845/rdmei.v19i1.3851

Sánchez, F. A. (2019). Fundamentos Epistémicos de la Investigación Cualitativa y Cuantitativa: Consensos y Disensos. *Rev. Digit. Invest. Docencia Univ*, *13*(1), 101–122. https://doi.org/10.19083/ridu.2019.644

Sánchez, M., García, M., Martínez, A., y Buzo, E. (2020). El Examen de Ingreso a la Universidad Nacional Autónoma de México: Evidencias de Validez de una Prueba de Alto Impacto y Gran Escala. *Revista iberoamericana de evaluación educativa*, *13*(2), 107–128. https://doi.org/10.15366/riee2020.13.2.006

Sánchez, M., y Martínez, A. (Eds.). (2022). *Evaluación y aprendizaje en educación universitaria: Estrategias e instrumentos* (Primera). UNAM, Coordinación de UNiversidad Abierta, Innovación Educativa y Educación a Distancia.

Saxton, D., Grefenstette, E., Hill, F., y Kohli, P. (2019). Analysing mathematical reasoning abilitiesof neural models. *Conference paper at ICLR 2019*, 1–17.

Shin, S.-Y., y Lidster, R. (2017). Evaluating different standard-setting methods in an ESL placement testing context. *Language testing*, *34*(3), 357–381. https://doi.org/10.1177/0265532216646605

Šifrar, M., y Trenc, A. (2014). Relating reading comprehension in the Spanish as a foreign language national exam to the CEFR: some aspects of evaluation. *Linguistica*, *54*(1), 309–323. https://doi.org/10.4312/linguistica.54.1.309-323

Sitotaw, B., y Tadele, K. (2018). Perception and trends in assessment of students' learning in Physics courses. *Latin-American journal of physics education*, *12*(1).

Smith, V. (2014). *Compendio de Instrumentos de Medición*. Instituto de Investigaciones Psicológicas.

Sondergeld, T. A., Stone, G. E., y Kruse, L. M. (2020). Objective Standard Setting in Educational Assessment and Decision Making. *Educational policy (Los Altos, Calif.)*, *34*(5), 735–759. https://doi.org/10.1177/0895904818802115

Sukirwan, Darhim, y Herman, T. (2018). Analysis of students' mathematical reasoning. *Journal of Physics: Conference Series*, *948*, 012036. https://doi.org/10.1088/1742-6596/948/1/012036

Taipe, E. (2021). La validez y la confiabilidad de la prueba de comprensión de textos aplicada a estudiantes de cuarto grado de primaria. *Diagnóstico Educativo*, *4*(4), 71–77.

Tall, D. (2014). Making Sense of Mathematical Reasoning and Proof. En M. N. Fried y T. Dreyfus (Eds.), *Mathematics & Mathematics Education: Searching for Common Ground* (pp. 223–235). Springer Netherlands. https://doi.org/10.1007/978-94-007-7473-5_13

Torres, A., y Contreras, J. (2022). Perspectivas de directivos educacionales sobre el uso de pruebas estandarizadas.: El caso de una evaluación de bajas consecuencias. *Revista mexicana de investigación educativa*, *27*(93), 511–536.

Tristán, A., y Pedraza, N. Y. (2017). La Objetividad en las Pruebas Estandarizadas. *Revista Iberoamericana de Evaluación Educativa*, *10*(1), 11–31. https://doi.org/10.15366/riee2017.10.1.001

Villalobos, M. (2018). *Puntos de corte en pruebas referidas a criterios: Análisis comparativo de estrategias para la evaluación de competencias digitales* [Info:eu-repo/semantics/masterThesis]. http://uvadoc.uva.es/handle/10324/32906

Villalobos, M., Marbán, J. M., y Anguita, R. (2021). Mapeo científico como técnica de investigación: Puntos de corte en pruebas de evaluación educativa referidas a criterios como campo de conocimiento. *Science Mapping as a research approach: cut-off points in criteria-referenced educational evaluation tests as field of knowledge*, *10*(33), 1–18. https://doi.org/10.30827/Digibug.70947

Víquez, L., Mora, M., Ordóñez, G., y Rojas, L. (2021). *Introducción a la Prueba de Habilidades Cuantitativas* (1ª). Editorial UCR.

Wyse, A. (2017). Five Methods for Estimating Angoff Cut Scores with IRT. *Educational Measurement: Issues and Practice*, *36*(4), 16–27.

Wyse, A. (2020). Comparing Cut Scores from the Angoff Method and Two Variations of the Hofstee and Beuk Methods. *Applied Measurement in Education*, *33*(2), 159–173. Education Source.

7. Anexos

7.1. Perfil de la persona mínimamente competente

Caracterización de la persona estudiante que aprueba la Prueba de Habilidades Cuantitativas

Es importante recordar que la PHC es una prueba que evalúa el razonamiento cuantitativo. En el marco de la prueba, el mismo se puede comprender como el razonamiento que se basa en contenidos matemáticos, esto abarca la elaboración de conclusiones basadas en el contenido matemático y el determinar cómo utilizar la matemática para la resolución de un problema.

Las áreas de contenido que se incluyen en la prueba, provienen del plan de estudios de la educación primaria y secundaria (hasta noveno año) y corresponden a las siguientes:

- *Aritmética:* conjuntos numéricos y teoría de números naturales
- *Geometría:* geometría plana, geometría analítica y cuerpos sólidos
- *Álgebra:* operaciones fundamentales con expresiones algebraicas, simplificación y factorización de expresiones algebraicas y ecuaciones e inecuaciones
- *Análisis de datos (estadística):* descripción de datos, medidas de posición y probabilidad

Considerando el contexto de la PHC, la persona que la "apruebe" debería poseer las siguientes habilidades según cada dimensión:

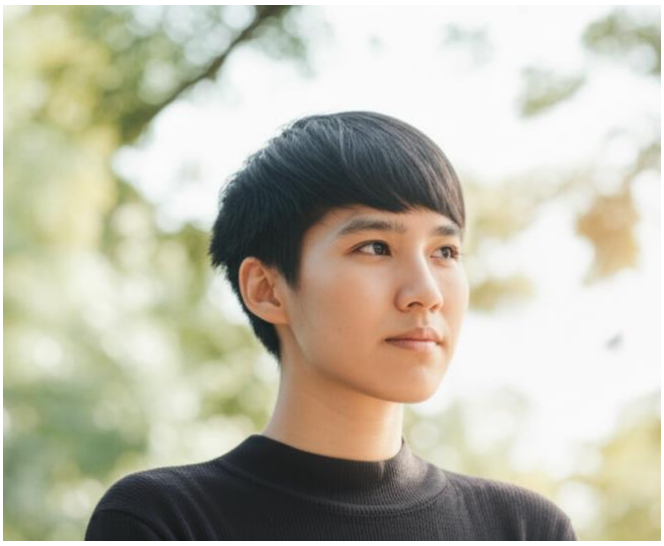
Dimensión de la prueba	Nivel mínimo deseable

Ejemplificación	Elabora un caso apropiado que satisfaga una condición **sencilla**.
	Ejemplo: identificar un número natural que sea par y mayor que 15.
Validación	Determina el valor de verdad de proposiciones considerando información asociada de forma poco explícita. Plantear contraejemplos de afirmaciones universales.
	Ejemplo: determinar si es correcta una afirmación respecto a la variación de la medida de los lados de un polígono y su efecto en la medida del área.
Generalización	Determina un patrón entre los elementos de una secuencia y lo aplica a un término **inmediato** a los términos presentados.
	Ejemplo: identificar un patrón en una secuencia para establecer el último dígito de determinado elemento de la misma.
Relacionar	Representa relaciones implícitas en el texto. Reconoce propiedades no explícitas en objetos matemáticos.
	Ejemplo: sustituir valores en una fórmula indicada en la prueba para realizar un cálculo determinado.
Clasificar	Identifica alguna propiedad **sencilla** en el objeto que le permita diferenciarlo de otros objetos sin vincularlos entre sí. Analiza conceptos **muy básicos** asociados a la estructura del objeto que le permitan diferenciarlo de los demás.
	Ejemplo: reconocer múltiplos de un número específico a partir de la factorización con números primos del valor dado.

Ahora, nuestra siguiente tarea es el acercarnos a definir el perfil de la persona que aprueba la PHC.

Rasgos generales del perfil de quien aprueba la PHC

A continuación, se busca establecer el perfil general de la persona que logre aprobar la PHC, esto con el propósito de unificar las habilidades deseables de ese grupo de la población y así homogenizar los criterios al momento de emitir los juicios respecto a los ítems al aplicar los métodos de punto de corte bookmark y Angoff. Considerando lo anterior, se les presenta a Cris:



Cris es una persona estudiante egresada de la Educación Secundaria del país. Además, se postulará para ingresar a alguna de las siguientes carreras que ofrece la Universidad de Costa Rica: Farmacia, Química, Física, Meteorología, Matemática, Ciencias Actuariales o Estadística. Cris ha analizado los planes de estudio de cada una de esas carreras y sabe que en el primer año de sus estudios se podrá encontrar alguno de los siguientes cursos en el área de la Matemática: MA1210, MA1001, MA0150, MA1021. Adicionalmente, sus orientadores le han informado sobre la PHC y sabe que para afrontar con éxito los cursos enumerados anteriormente, requiere contar al menos con las siguientes habilidades asociadas al razonamiento cuantitativo y que son evaluadas en esa prueba:

- Construir un ejemplo para satisfacer una condición compleja.
- Determinar el valor de verdad de proposiciones que abarcan varios casos y que requieren análisis diferenciados.
- Construir contraejemplos para afirmaciones universales, que abarcan casos verdaderos y falsos.
- Construir una regla de generalización y aplicarla a un término lejano a los términos presentados.
- Integrar múltiples relaciones entre elementos.
- Reconocer relaciones no indicadas que pueden utilizarse en un problema dado.
- Identificar alguna propiedad compleja en los objetos y que permita diferenciarlos de otros, esto sin vincularlos entre sí.
- Analizar conceptos básicos asociados a la estructura del objeto que permitan diferenciarlos de los demás.

¿Cuáles ítems de la PHC serán los indispensables para asegurar que Cris posee las habilidades requeridas?

7.2. Cuadernillo de práctica para jueces

Antes de iniciar con este ejercicio que forma parte del proceso planificado para el establecimiento de puntos de corte para la Prueba de Habilidades Cuantitativas (PHC), quiero agradecerle su participación y señalar nuevamente la importancia que su conocimiento disciplinar aunado a la experiencia en las aulas universitarias posee para el alcance del objetivo de esta investigación. A la vez, aclaramos que todos los juicios emitidos en este proceso se manipulan de manera anónima y en conjunto con los de los demás integrantes del equipo de jueces y únicamente para los alcances del Trabajo Final de Graduación titulado ESTABLECIMIENTO DE PUNTOS DE CORTE PARA PRUEBAS ESTANDARIZADAS REFERIDAS A CRITERIOS, EL CASO DE LA PRUEBA DE HABILIDADES CUANTITATIVAS.
Para el desarrollo de este ejercicio se ha seleccionado solamente una de las cuatro partes que integra la PHC, específicamente el apartado de aritmética, esto con el propósito de que su abordaje sea rápido, pero que a la vez permita conocer y practicar el procedimiento que se llevará a cabo con la versión real de los ítems que componen el banco de reactivos de la PHC.

Consideraciones generales

En este folleto encontrará una selección de 10 de los reactivos que conforman el libro de práctica de la PHC, mismo al que se puede tener acceso por medio del sitio web de la prueba. Estos ítems se han organizado de menor a mayor dificultad y para ello se ha considerado las habilidades de razonamiento requeridas y los contenidos involucrados en cada uno. Además, se ha resaltado con negrita la opción más adecuada para cada ejercicio. Para cada ítem se incluye el desarrollo de **una posible solución** para cada uno de los ejercicios y los conocimientos previos que se consideran necesarios para resolver cada planteamiento. Tanto las alternativas de solución como los conocimientos previos asociados se basan en el texto *Introducción a la Prueba de Habilidades Cuantitativas* de Víquez, L.; Mora, M.; Ordoñez, G. y Rojas, L. (2021).
Para cada ítem se ha dedicado una página individual y en cada una de ellas se ofrece un espacio para que indique cuántas personas examinadas de un grupo de 100 con un nivel básico (es decir, capaces de realizar los procesos mínimos que se han considerado en el perfil deseado de la persona que aprueba la PHC), considera que acertarán el ítem, esto según la ruta que ofrece el método Angoff. El valor que asigne deberá considerar múltiplos de 10, esto para facilitar la sistematización de los resultados obtenidos.
Por su parte, para la aplicación del método Bookmark, se requiere identificar un ítem de la prueba a partir del cual se requiera un “un nivel medio de razonamiento cuantitativo”. Para ello, considere que un ítem de nivel básico es aquel que 2 de cada 3 estudiantes con un nivel básico lo responden adecuadamente, mientras que un ítem de nivel medio es aquel que solo puede resolverlo menos del 66.6% de esa población. Se ha destinado un espacio para que indique cuál de los ítems es el primero que consideraría como de nivel medio y el porqué de esa decisión. Para justificar su elección, considere únicamente elementos asociados al perfil de razonamiento de PHC.
Por último, puede disponer del espacio en blanco de cada página para hacer las anotaciones que considere pertinentes para cada uno de los reactivos que se incluyen en este documento.

Temática de la sección de la prueba: Aritmética

Ítem #1 (opción correcta: a)
Dimensión asociada al ítem: Relacionar. [Posibles procesos para el nivel básico de *Relacionar*: Representa relaciones implícitas en el texto, reconoce propiedades no explícitas en objetos matemáticos.]

1. Al dividir 14 505 por un número natural *n*, el residuo es 25. Con base en lo anterior, ¿cuál de los siguientes números naturales podría ser el valor de *n*?

a) 7240
b) 7241
c) 7242
d) 7245

Procedimiento para la solución
1. Efectuar de manera correcta la división del número indicado por cada una de las opciones de respuesta. 2. Identificar cuál de las alternativas produce un residuo de 25. **Conocimientos previos asociados:** • Números naturales • Divisibilidad • Múltiplos

¿Cuántas personas examinadas de un grupo de 100 con un nivel aceptable, acertarán el ítem? [_____].	¿Es este el primer ítem que clasificaría como de nivel medio? [_____].
Justifique por qué postula ese valor:	En caso afirmativo, indique los procedimientos de razonamiento adicionales que diferencian a este ítem de los otros de esta categoría:

Ítem #2 (opción correcta: d)
Dimensión asociada al ítem: Clasificar. [Posibles procesos para el nivel básico de *Clasificar*: Identifica alguna propiedad sencilla en el objeto que le permita diferenciarlo de otros objetos sin vincularlos entre sí, analiza conceptos muy básicos asociados a la estructura del objeto que le permitan diferenciarlo de los demás.]

2. Considere las siguientes cantidades:

I. $0{,}2\%\ de\ 100$.

II. $95\%\ de\ \frac{1}{5}$.

III. $\frac{2^2}{20}$.

Con base en lo anterior, ¿cuál de las siguientes afirmaciones es, **con certeza**, verdadera?

a) La cantidad I es igual que la cantidad II.
b) La cantidad II es igual que la cantidad III.
c) La cantidad I es menor que la cantidad II y III.
d) La cantidad II es menor que la cantidad I y III.

Procedimiento para la solución
1. Calcular el valor numérico de cada una de las tres cantidades (respectivamente: 0.2, 0.19 y 0.2). 2. Validar cada una de las 4 opciones de respuesta y concluir cuál de ellas es verdadera. **Conocimientos previos asociados:** • Números racionales en distintas representaciones y sus relaciones de orden

¿Cuántas personas examinadas de un grupo de 100 con un nivel aceptable, acertarán el ítem? [_____].	¿Es este el primer ítem que clasificaría como de nivel medio? [_____].
Justifique por qué postula ese valor:	En caso afirmativo, indique los procedimientos de razonamiento adicionales que diferencian a este ítem de los otros de esta categoría:

Ítem #3 (opción correcta: b)
Dimensión asociada al ítem: Clasificar. [Posibles procesos para el nivel básico de *Clasificar*: Identifica alguna propiedad sencilla en el objeto que le permita diferenciarlo de otros objetos sin vincularlos entre sí, analiza conceptos muy básicos asociados a la estructura del objeto que le permitan diferenciarlo de los demás.]

3. Si m es un número entero que satisface la desigualdad $-2 < m + 5 < 2$, entonces, ¿cuál es la cantidad de posibles valores de m?

a) 2
b) 3
c) 4
d) 5

Procedimiento para la solución
1. Separar la desigualdad en dos más sencillas. 2. Despejar la incógnita "m"en cada una de las desigualdades. 3. Establecer el intervalo en el que puede estar el valor de la incógnita. 4. Definir la cantidad de valores que se encuentran en ese intervalo. **Conocimientos previos asociados:** • Números enteros • Inecuaciones de primer grado • Valor numérico de expresiones algebraicas

¿Cuántas personas examinadas de un grupo de 100 con un nivel aceptable, acertarán el ítem? [_____].	¿Es este el primer ítem que clasificaría como de nivel medio? [_____].
Justifique por qué postula ese valor:	En caso afirmativo, indique los procedimientos de razonamiento adicionales que diferencian a este ítem de los otros de esta categoría:

Ítem #4 (opción correcta: b)
Dimensión asociada al ítem: Clasificar. [Posibles procesos para el nivel básico de *Clasificar*: Identifica alguna propiedad sencilla en el objeto que le permita diferenciarlo de otros objetos sin vincularlos entre sí, analiza conceptos muy básicos asociados a la estructura del objeto que le permitan diferenciarlo de los demás.]

4. ¿Cuál de los siguientes números es un divisor de $78^2 + 2 \cdot 78 \cdot 14 + 14^2$?

a) 15
b) 46
c) 64
d) 78

Procedimiento para la solución
1. Rescribir la expresión como un producto notable [$(78 + 14)^2 = (2^2 \cdot 23)^2$]. 2. Identificar que un posible divisor del número dado debe ser un producto del 2 y el 23 pero con exponentes menores a 4 y 2 respectivamente. 3. Verificar cuál de las opciones se clasifica como un divisor de la expresión $(2^2 \cdot 23)^2$. **Conocimientos previos asociados:** • Números naturales • Múltiplos • Descomposición prima • Divisibilidad • Conocimiento de la primera fórmula notable[1]

¿Cuántas personas examinadas de un grupo de 100 con un nivel aceptable, acertarán el ítem? [_____].	¿Es este el primer ítem que clasificaría como de nivel medio? [_____].
Justifique por qué postula ese valor:	En caso afirmativo, indique los procedimientos de razonamiento adicionales que diferencian a este ítem de los otros de esta categoría:

[1] No es indispensable, pero facilita el procedimiento.

Ítem #5 (opción correcta: c)
Dimensión asociada al ítem: Clasificar. [Posibles procesos para el nivel básico de *Clasificar*: Identifica alguna propiedad sencilla en el objeto que le permita diferenciarlo de otros objetos sin vincularlos entre sí, analiza conceptos muy básicos asociados a la estructura del objeto que le permitan diferenciarlo de los demás.]

5. Si *n* es un número natural impar mayor que 1, entonces, ¿por cuál de los siguientes valores es divisible, **con certeza**, la expresión $4^n + 6^n$?

a) 6
b) 7
c) 8
d) 11

Procedimiento para la solución
1. Construir casos particulares para la expresión, considerando que *n* debe ser un impar mayor o igual que 3. 2. Se clasifican los valores obtenidos dependiendo de si son divisibles o no por cada una de las opciones, esto se puede lograr buscando contraejemplos en cada caso. 3. Se determina cuál opción no se logra descartar con el análisis de los casos. **Conocimientos previos asociados:** • Números naturales (en particular las potencias y sus leyes) • Divisibilidad • Múltiplos • Descomposición prima • Método del factor común para factorizar[2]

¿Cuántas personas examinadas de un grupo de 100 con un nivel aceptable, acertarán el ítem? [_____].	¿Es este el primer ítem que clasificaría como de nivel medio? [_____].
Justifique por qué postula ese valor:	En caso afirmativo, indique los procedimientos de razonamiento adicionales que diferencian a este ítem de los otros de esta categoría:

[2] Es deseable el conocimiento del método del factor para realizar la factorización, pero en su defecto, basta con el conocimiento de la propiedad distributiva de la multiplicación respecto a la suma.

Ítem #6 (opción correcta: d)
Dimensión asociada al ítem: Validar. [Posibles procesos para el nivel básico de *Validar*: Determina el valor de verdad de proposiciones considerando información asociada de forma poco explícita, plantea contraejemplos de afirmaciones universales.]

6. Si n es un número natural, tal que
   $2^{11} \cdot 7^3 \cdot 5^1 = 2^{10} \cdot n \cdot 7^3$,

Entonces, ¿cuál de las siguientes relaciones se cumple **con certeza**?

a) $n < 5$
b) $n > 10$
c) $n^3 < 64$
d) $\boldsymbol{n^2 < 128}$

Procedimiento para la solución
1. Considerar la factorización de cada extremo de la igualdad. 2. Inferir que en el extremo derecho de la igualdad falta el producto de 2 y 5. 3. Considerando la conclusión del paso anterior, se comprueba cada opción para definir la correcta. **Conocimientos previos asociados:** • Números naturales, en particular el concepto de factor, las leyes de potencias y las relaciones de orden • Descomposición prima • Ecuaciones de primer grado[3]

¿Cuántas personas examinadas de un grupo de 100 con un nivel aceptable, acertarán el ítem? [_____].	¿Es este el primer ítem que clasificaría como de nivel medio? [_____].
Justifique por qué postula ese valor:	En caso afirmativo, indique los procedimientos de razonamiento adicionales que diferencian a este ítem de los otros de esta categoría:

[3] Su necesidad depende de la estrategia de solución elegida por la persona examinada.

Ítem #7 (opción correcta: d)
Dimensión asociada al ítem: Generalizar. [Posibles procesos para el nivel básico de *Generalizar*: Construye una regla de generalización y la aplica a un término inmediato a los términos presentados.]

7. Considere la siguiente secuencia numérica:

$$u_2 = \left(\frac{2+1}{2}\right)$$
$$u_3 = \left(\frac{2+1}{2}\right)\left(\frac{3+1}{3}\right)$$
$$u_4 = \left(\frac{2+1}{2}\right)\left(\frac{3+1}{3}\right)\left(\frac{4+1}{4}\right)$$
$$\vdots$$
$$u_n = \left(\frac{2+1}{2}\right)\left(\frac{3+1}{3}\right)\left(\frac{4+1}{4}\right)\cdots\left(\frac{n+1}{n}\right)$$

Con base en la secuencia anterior, ¿cuál es el valor de u_{100}?

a) 1^{100}
b) 2^{100}
c) $\frac{100}{2}$
d) $\mathbf{\frac{101}{2}}$

Procedimiento para la solución
1. Calcular los primeros términos de la secuencia y realizar las operaciones obtenidas en los paréntesis de cada caso y obtener como resultado un valor racional expresado en notación fraccionaria. 2. Determinar el patrón que se genera en el paso anterior. 3. Generalizar la secuencia para definir qué sucederá en el caso de u_{100}. **Conocimientos previos asociados:** • Números racionales • Simplificación

¿Cuántas personas examinadas de un grupo de 100 con un nivel aceptable, acertarán el ítem? [_____].	¿Es este el primer ítem que clasificaría como de nivel medio? [_____].
Justifique por qué postula ese valor:	En caso afirmativo, indique los procedimientos de razonamiento adicionales que diferencian a este ítem de los otros de esta categoría:

Ítem #8 (opción correcta: b)
Dimensión asociada al ítem: Clasificar. [Posibles procesos para el nivel básico de *Clasificar*: Identifica alguna propiedad sencilla en el objeto que le permita diferenciarlo de otros objetos sin vincularlos entre sí, analiza conceptos muy básicos asociados a la estructura del objeto que le permitan diferenciarlo de los demás.]

8. Si x y y son números naturales pares consecutivos, entonces, cuál de las siguientes características corresponde, **con certeza**, al valor numérico de la expresión $\frac{x+y}{2}$?

a) Es un número par.
b) Es un número impar.
c) Es múltiplo de cuatro.
d) Es un número primo.

Procedimiento para la solución
1. Considerar algunos casos que cumplan las indicaciones ofrecidas (por ejemplo: 2 y 4, 4 y 6, 6 y 8…). 2. Calcular los valores correspondientes para cada caso según lo indica la expresión del ejercicio. 3. Determinar la característica común de los valores que se obtienen en cada caso. **Conocimientos previos asociados:** • Números naturales • Divisibilidad • Múltiplos • Números compuestos • Números primos

¿Cuántas personas examinadas de un grupo de 100 con un nivel aceptable, acertarán el ítem? [_____].	¿Es este el primer ítem que clasificaría como de nivel medio? [_____].
Justifique por qué postula ese valor:	En caso afirmativo, indique los procedimientos de razonamiento adicionales que diferencian a este ítem de los otros de esta categoría:

Ítem #9 (opción correcta: c)
Dimensión asociada al ítem: Relacionar. [Posibles procesos para el nivel básico de *Relacionar*: Representa relaciones implícitas en el texto, reconoce propiedades no explícitas en objetos matemáticos.]

9. ¿Cuál es el valor de la suma de los dígitos del número $(200)^6 + (700)^2$?

a) 8
b) 9
c) 23
d) 113

Procedimiento para la solución
1. Descomponer en factores primos cada término e identificar que $(200)^6$ se puede rescribir como $64 \cdot 1000000000000$ y que $(700)^2$ puede ser $49 \cdot 10000$. 2. Identificar cuáles dígitos de esos productos aportarán a la suma solicitada en el ejercicio. 3. Realizar la suma requerida para proceder a elegir la opción correcta. **Conocimientos previos asociados:** • Números naturales • Descomposición prima • Factorización por medio del método del factor común

¿Cuántas personas examinadas de un grupo de 100 con un nivel aceptable, acertarán el ítem? [_____].	¿Es este el primer ítem que clasificaría como de nivel medio? [_____].
Justifique por qué postula ese valor:	En caso afirmativo, indique los procedimientos de razonamiento adicionales que diferencian a este ítem de los otros de esta categoría:

Ítem #10 (opción correcta: b)
Dimensión asociada al ítem: Clasificar. [Posibles procesos para el nivel básico de *Clasificar*: Identifica alguna propiedad sencilla en el objeto que le permita diferenciarlo de otros objetos sin vincularlos entre sí, analiza conceptos muy básicos asociados a la estructura del objeto que le permitan diferenciarlo de los demás.]

10. Si p y m son números enteros positivos, tales que $p \div 2$ es entero y $m \div 3$ es par, entonces, ¿cuál de las siguientes opciones es, **con certeza**, verdadera?

a) $p \cdot m$ es múltiplo de 9.
b) $p \cdot m$ es múltiplo de 12.
c) $\frac{3p}{m}$ es entero.
d) $\frac{2m}{3p}$ es par.

Procedimiento para la solución
1. Según la información indicada, se reconoce que p es un número par y que m es múltiplo de 6. 2. Aplicar las reglas de divisibilidad junto con el cálculo de múltiplos de los valores dados para determinar la opción correcta, por ejemplo, el producto de los valores no podría ser múltiplo de 9, pero sí de 12. Por otra parte, $\frac{3p}{m}$ no es un valor entero y $\frac{2m}{3p}$ tampoco es par. **Conocimientos previos asociados:** • Números enteros • Divisibilidad • Múltiplos

¿Cuántas personas examinadas de un grupo de 100 con un nivel aceptable, acertarán el ítem? [_____].	¿Es este el primer ítem que clasificaría como de nivel medio? [_____].
Justifique por qué postula ese valor:	En caso afirmativo, indique los procedimientos de razonamiento adicionales que diferencian a este ítem de los otros de esta categoría: